ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 5** **Москва 2017**

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 004.414.2/.3 : 004.85.021

Д.В. Виноградов

Анализ результатов применения ВКФ-системы: успехи и открытая проблема

Описывается программная реализация ВКФ-метода интеллектуального анализа данных, успешно примененного к двум массивам из репозитария данных для тестирования алгоритмов машинного обучения. Упрощенный вариант ВКФ-системы для поиска случайного подмножества продемонстрировал впечатляющую скорость ее работы. Будет доказан теоретический результат, объясняющий наблюдаемый феномен. Общий случай пока не поддается такому анализу, однако на практике скорость работы и в общем случае очень высока.

Ключевые слова: *сходство, спаривающая цепь Маркова, ВКФ-кандидат, концентрация около среднего*

ВВЕДЕНИЕ

Основываясь на результатах группы отечественных исследователей под руководством профессора В.К. Финна [1, 2], автором работы [3] был предложен новый вероятностно-комбинаторный подход. Так как некоторые ингредиенты заимствованы из теории формальных понятий, автор назвал его вероятностно-комбинаторный формальный метод, сокращенно ВКФ-метод. Для апробации ВКФ-метода была создана программная система, названная ВКФ-системой, которую с успехом применили к двум массивам из репозитория данных для тестирования алгоритмов машинного обучения.

В ходе проведения экспериментов была обнаружена высокая скорость работы ключевого алгоритма

системы – спаривающей цепи Маркова. Хотя общий случай остается открытым, удалось достичь прогресса в понимании этого феномена в частном случае Булеана всех подмножеств признаков.

В настоящей работе мы опишем результаты применения ВКФ-системы к реальным данным массивов SPECT Hearts и Mushrooms из репозитория данных для тестирования алгоритмов машинного обучения. Основным теоретическим результатом будет оценка среднего времени спаривания и теорема о сильной концентрации времени спаривания около его среднего. Наконец, сформулируем открытую проблему нахождения среднего времени спаривания цепи Маркова для произвольной обучающей выборки.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Вероятностные когнитивные процедуры, основанные на операции сходства, были запрограммированы автором статьи в единой программной системе, получившей название ВКФ-система:

- программа реализована как консольное приложение. Она была создана в среде Code::Blocks (version 13.12) с использованием библиотеки boost (version1_56_0). Компилятор C++ - GNUC++ toolset (version 4.9.1);.
- примеры (обучающие, контр- и представленные для предсказания целевого свойства) представляются объектами класса *boost::dynamic_bitset*. Они сохраняются в контейнерах типа *std::vector* и *std::list* стандартной библиотеки C++;
- программа использует классы *boost::random* для датчиков случайных чисел. Это нужно для спаривающей цепи Маркова, реализуемой алгоритмом 3;
- для реализации многопоточности используются классы *boost::thread*.;
- программа платформенно независима: она собиралась и запускалась под Windows и под Linux.

Прокомментируем некоторые достоинства ВКФ-системы по сравнению с классическим ДСМ-подходом.

- Так как каждая ВКФ-гипотеза порождается независимым запуском цепи Маркова, то ВКФ-программа использует несколько потоков для вычисления индуктивного обобщения. Для ДСМ-системы подобное распараллеливание индукции невозможно.
- ВКФ-система вычисляет процедуру абдуктивного уточнения и принятия ВКФ-гипотез тоже в несколько потоков. В ДСМ-системе распараллеливание шага абдукции возможно, но пока не реализовано.
- Предсказание свойств по аналогии осуществляется в один поток, так как вычислительная сложность этого шага мала в сравнении с шагом индукции.
- На центрально процессорном устройстве с четырьмя потоками (i5-3210M) максимальная нагрузка процессора при вычислении в 4 потока достигает 90%. Для существующей параллельной версии ДСМ-системы она не превосходит 50%.

Простейшая ВКФ-система применялась к двум массивам из репозитория данных для проверки алгоритмов машинного обучения.

Первым массивом был SPECT Hearts (данные компьютерной томографии сердца).

- Обучающая выборка содержит 40 (+)- и 40 (–)-примеров.
- Тестовая выборка содержит 172 (+)- и 15 (–)-примеров.
- Каждый пример описывался 22 бинарными атрибутами.
- ВКФ-система добавила отрицания исходных признаков, чтобы отсутствие атрибута могло быть частью причины проявления свойства. Поэтому обучающая выборка - это матрица размера 40×44 .
- Точность предсказания простейшей ВКФ-системы достигла 85,56 % (151 из 172 (+)-примеров и 9 из 15 (–)-примеров).
- Авторы массива SPECT достигли 84,0% точности своей программой CLIP3, которая реализует обучение покрытию средствами целочисленного программирования.

Второй массив Mushrooms – данные из определителя грибов Северной Америки, оцифрованные в файл agaricus-lepiota.data.

- Исходные данные включают описания 8124 грибов, разделенные на две категории (съедобные и ядовитые). Мы случайным образом разделили их на обучающую и тестовую выборки.
- Обучающая выборка содержит 4032 объекта.
- Тестовая выборка содержит 2120 (+)- (съедобные грибы) и 1972 (–)-примеров (ядовитые грибы).
- Каждый пример описывался 22 признаками, представляющими различные характеристики грибов (цвет, форма шляпки, места произрастания, частота встречаемости и т.п.). Эти признаки – номинальные, принимающие одно из нескольких значений.
- ВКФ-система закодировала эти признаки битовыми строками длины 128 бит.
- Точность предсказания простейшей ВКФ-системы достигла 100% для 100 ВКФ-гипотез и их абдуктивного уточнения.

СИЛЬНАЯ КОНЦЕНТРАЦИЯ ВРЕМЕНИ РАБОТЫ ОКОЛО СРЕДНЕГО

Во время проведения экспериментов с ВКФ-системой был обнаружен феномен очень быстрого нахождения очередного ВКФ-кандидата. Хотя мы не смогли получить оценку в общем виде, для случая Булеана имеются результаты о среднем времени склеивания и сильной концентрации времени склеивания около своего среднего.

До конца этого параграфа мы ограничимся случаем Булеана:

Пусть $O=\{o_1,o_2,...,o_n\}$ будет множеством объектов, каждый из которых описывается признаками из списка $F=\{f_1,f_2,...,f_n\}$, и $o_i If_j \Leftrightarrow i\neq j$.

O\F	f_1	f_2	$\cdots$	f_n
o_1	0	1	$\cdots$	1
o_2	1	0	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
o_n	1	1	$\cdots$	0

Ясно, что $o_{j_1} \cap o_{j_2} \cap ... \cap o_{j_k} = F \setminus \{f_{j_1}, f_{j_2}, ..., f_{j_k}\}$, так как добавление в сходство примера o_k с номером k удаляет из фрагмента признак f_k с тем же самым номером k.

Очевидно, что таким образом может быть получено любое подмножество n-элементного множества F.

Для дальнейшего нам будет полезен явный вид операций «замыкай-по-одному» в решетке-Булеане

$$o_i I f_j \Leftrightarrow i \neq j,$$

для множества объектов $O = \{o_1, o_2, ..., o_n\}$, каждый из которых описывается признаками из списка

$$F = \{f_1, f_2, ..., f_n\}.$$

Ясно, что в этом случае

$$CbODown(\langle A, B\rangle, o_k) = \begin{cases} \langle A \cup \{o_k\}, B \setminus \{f_k\}\rangle, & \text{если } o_k \notin A \\ \langle A, B\rangle & \text{иначе,} \end{cases}$$

так как добавление в сходство примера o_k с номером k удаляет из фрагмента признак f_k с тем же самым номером k.

Аналогично,

$$CbOUp(\langle A, B\rangle, f_k) = \begin{cases} \langle A \setminus \{f_k\}, B \cup \{f_k\}\rangle, & \text{если } f_k \notin B \\ \langle A, B\rangle & \text{иначе,} \end{cases}$$

так как добавление в сходство признака f_k с номером k удаляет из родителей сходства объект o_k с тем же самым номером k.

Определение 1. ***Расстояние*** $\rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle)$ между кандидатами $\langle A_1, B_1\rangle$ и $\langle A_2, B_2\rangle$ определяется как число позиций, в которых отличаются битовые строки B_1 и B_2. Другими словами, расстояние равно минимальному количеству ребер между соответствующими вершинами гиперкуба.

Это расстояние на гиперкубе называется *метрикой Хэмминга*.

Оказывается, что после применения операций «замыкай-по-одному» в случае Булеана, это расстояние не увеличивается. Более точное утверждение сформулировано в следующей лемме.

Лемма 1. *Пусть* $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$. *Тогда*

$$\rho(CbOUp(\langle A_1, B_1\rangle, f), CbOUp(\langle A_2, B_2\rangle, f)) = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle), \text{ если } f \in B_1 \text{ и } f \in B_2.$$

$$\rho(CbOUp(\langle A_1, B_1\rangle, f), CbOUp(\langle A_2, B_2\rangle, f)) = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle) - 1, \text{ если } f \notin B_1 \text{ и } f \in B_2.$$

$$\rho(CbOUp(\langle A_1, B_1\rangle, f), CbOUp(\langle A_2, B_2\rangle, f)) = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle), \text{ если } f \notin B_1 \text{ и } f \notin B_2.$$

$$\rho(CbODown(\langle A_1, B_1\rangle, o), CbODown(\langle A_2, B_2\rangle, o)) = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle), \text{ если } o \in A_1 \text{ и } o \in A_2.$$

$$\rho(CbODown(\langle A_1, B_1\rangle, o), CbODown(\langle A_2, B_2\rangle, o)) = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle) - 1, \text{ если } o \in A_1 \text{ и } o \notin A_2.$$

$$\rho(CbODown(\langle A_1, B_1\rangle, o), CbODown(\langle A_2, B_2\rangle, o)) = \rho(\langle A_1, B_1\rangle, \langle A_2, B_2\rangle), \text{ если } o \notin A_1 \text{ и } o \notin A_2.$$

Нам теперь понадобится один известный класс распределений вероятностей (см. [4]):

Определение 2. Целочисленная случайная величина T имеет ***геометрическое*** распределение вероятностей, если $P[T = k] = (1-p)^{k-1} \cdot p$, где $0 < p \leq 1$ и $k > 0$.

Лемма 2. *Время* T_n *ожидания уменьшения расстояния с* $\rho(\langle O, \varnothing\rangle, \langle \varnothing, F\rangle) = n$ *до* n-1 *имеет геометрическое распределение с* $p = 1$. *Время* T_j *ожидания уменьшения расстояния с* j *до* j-1 *имеет геометрическое распределение с* $p = \frac{j}{n}$.

Доказательство следует из предыдущей леммы и разбора случаев выбора для следующего замыкаемого объекта или признака.

Сформулируем и докажем теперь классическую лемму:

Лемма 3. *Для целочисленной случайной величины* T *с геометрическим распределением вероятностей выполнено* $E[T] = \frac{1}{p}$ *и* $D[T] = \frac{(1-p)}{p^2}$.

Доказательство. Воспользуемся производящей функцией моментов

$$\psi_T(z) = E[z^T] = \frac{pz}{1-(1-p)z}.$$

Имеем

$$E[T] = \psi'(1)_T = \frac{1}{p}$$

и

$$D[T] = \psi_T''(1) + \psi_T'(1) - (\psi_T'(1))^2 = \frac{(1-p)}{p^2}.$$

Теорема 1. *Среднее время склеивания для* n*-мерного гиперкуба равно*

$$E\left[\sum_{j=1}^{n} T_j\right] = \sum_{j=1}^{n} \frac{n}{j} \approx n \cdot \ln(n) + n \cdot \gamma + \frac{1}{2}.$$

Доказательство. Из-за линейности среднего имеем $E\left[\sum_{j=1}^{n} T_j\right] = \sum_{j=1}^{n} E\left[T_j\right]$. По Лемме 3 имеем $E\left[T_j\right] = \frac{n}{j}$.

Вспомним следующую классическую лемму П.Л.Чебышева:

Лемма 4. $P\left[\left|T - E[T]\right| \geq \varepsilon\right] \leq \frac{D[T]}{\varepsilon^2}$.

Теорема 2. $P\left[\sum_{j=1}^{n} T_j \geq (1+\varepsilon)\cdot n \cdot \ln(n)\right] \to 0$ *при* $n \to \infty$ *для любого* $\varepsilon > 0$.

Доказательство. Из-за независимости T_j имеем $D\left[\sum_{j=1}^{n} T_j\right] = \sum_{j=1}^{n} D\left[T_j\right]$. По Лемме 3 имеем $D\left[T_j\right] \approx \frac{n^2}{j^2}$. Поэтому $D\left[\sum_{j=1}^{n} T_j\right] = O\left(n^2\right)$. Для по Лемме 4 при достаточно больших n имеем

$$P\left[T \geq (1+\varepsilon)\cdot n \cdot \ln(n)\right] \leq$$
$$\leq P\left[\left|T - E[T]\right| \geq \varepsilon n \cdot \ln(n)\right] \leq \frac{D[T]}{\varepsilon^2 n^2 \cdot \ln(n)}.$$

При $n \to \infty$, имеем требуемый результат.

Хотелось бы обратить внимание читателя на следующие замечательные обстоятельства:

- для 32-мерного гиперкуба среднее время склеивания $E[T] = 32 \cdot \sum_{j=1}^{32} \frac{1}{j} \leq 130$. Чтобы выбрать случайное подмножество из 32 признаков, нужно использовать 32 раза датчик случайных чисел, так что наша оценка не сильно (только логарифмически) хуже;
- но в 32-мерном гиперкубе 4294967296 сходств, подавляющее большинство которых не будет вычисляться в процессе вероятностного поиска сходств;
- эксперименты показывают, что и в общем случае время спаривания мало по сравнению с числом различных ВКФ-кандидатов.

Теперь мы сформулируем открытую проблему, исследование которой может провести сам читатель:

Получить оценку среднего времени склеивания в общем случае. Полезно указать, что метрика Хэмминга между верхним и нижним ВКФ-кандидатом не является функцией Ляпунова (может возрастать) в спаривающей цепи Маркова для произвольного формального контекста. Общее определение спаривающей цепи Маркова и соответствующий пример убывания расстояния Хэмминга есть у автора в статье [5].

ЗАКЛЮЧЕНИЕ

В настоящей работе описана программная реализация ВКФ-метода интеллектуального анализа данных. Эта система была успешно применена к двум массивам SPECT Hearts и Mushrooms из репозитария данных для тестирования алгоритмов машинного обучения. Более того, упрощенный вариант ВКФ-системы для поиска случайного подмножества продемонстрировал впечатляющую скорость работы. Этот феномен получил свое теоретическое обоснование через две теоремы. Первая утверждает, что среднее время склеивания растет как $O\left(n \cdot \ln(n)\right)$. Вторая же доказывает сильную концентрацию времени склеивания около своего среднего. К сожалению, общий случай пока не поддается такому анализу, хотя на практике скорость работы и в общем случае очень высока.

* * *

Автор благодарит проф. В.К. Финна за внимание к работе, проф. Е.М Бениаминова за полезные обсуждения и своих коллег по лаборатории 35 ФИЦ ИУ РАН за поддержку и полезные дискуссии.

СПИСОК ЛИТЕРАТУРЫ

1. Аншаков О.М., Скворцов Д.П., Финн В.К. О дедуктивной имитации некоторых вариантов ДСМ-метода автоматического порождения гипотез // Семиотика и информатика. – Вып. 33. – 1993. – С. 164–233.
2. ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания / ред.: В.К. Финн, О.М. Аншаков. – М.: URSS, 2009. – 432 с.
3. Vinogradov D.V. VKF-methodofhypothesesgeneration // Communication in Computer and Information Science. – 2014. – Vol. 436. – P. 237–248.
4. Феллер В. Введение в теорию вероятностей и ее приложения. В 2–х томах. Т. 1: Пер. с англ. – М.: Мир, 1984. – 528 с.
5. Виноградов Д.В. Вероятностное порождения гипотез в ДСМ-методе с помощью простейших цепей Маркова // Научно-техническая информация. Сер. 2. – 2012. – № 9. – С. 20–27; Vinogradov D.V. Random Generation of Hypotheses in the JSM Method using Simple Markov Chains // Automatic Documentation and Mathematical Linguistics. – 2012. – Vol. 46, № 5. – P. 221-228.

Материал поступи в редакцию 26.01.17.

Сведения об авторе

ВИНОГРАДОВ Дмитрий Вячеславович – кандидат физико-математических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» РАН и Российского государственного гуманитарного университета, Москва
e-mail: vinogradov.d.w@qmail.com

УДК 004.774

Н.С. Редькина

Направления развития инструментов веб-аналитики

Рассмотрены возможности счетчиков веб-сайтов и способы анализа тенденций развития различных инструментов веб-аналитики с помощью сервиса Google Trends и представлены перспективные подходы к более детальному исследованию трендов в области веб-аналитики, многоаспектному изучению данных по счетчикам, сервисам и лог-анализаторам, ориентированным на оптимизацию и продвижение сайтов различных организаций.

Приведен обзор источников сравнительного анализа инструментов веб-аналитики, рейтингов и данные лаборатории Ruward:Tracko о количестве установленных счетчиков на сайтах в Рунете.

Ключевые слова: *веб-аналитика, счетчики веб-сайтов, веб-анализаторы, оценка, веб-метрика, Google Trends*

ВВЕДЕНИЕ

Стремительное развитие веб-технологий и информационной сферы получает отклик практически во всех сферах человеческой деятельности. Увеличивается количество сайтов и пользователей сети. По данным Всероссийского центра изучения общественного мнения в 2016 г. в России 70% граждан в возрасте от 18 лет и старше пользуются Интернетом [1]. Число сайтов в Интернете достигло к началу 2017 г. 1, 8 млрд [2]. Веб-сайты являются неотъемлемой частью деятельности любой организации. Их владельцы стали все больше внимания уделять не только контенту, улучшению пользовательских интерфейсов, поисковых систем, навигационных функций, но и сбору данных о своих сайтах с помощью инструментов веб-аналитики для определения информативности, удобства и популярности у пользователей, выявления наиболее востребованных разделов, а также причин отказов от использования.

Системы веб-аналитики и разного вида счетчики позволяют обрабатывать большое количество данных о пользователях, таких как тип браузера, скорость соединения, размер экрана, тип, возраст посетителей и т.д. Собранные сведения становятся полезными для решения задач оптимизации сайта и динамичного реагирования на постоянно изменяющиеся условия внешней среды, информационные предпочтения и потребности пользователей. Инструменты веб-аналитики позволяют адаптировать услуги к потребностям пользователей, способствуют формированию благоприятного имиджа и положительной репутации организации в виртуальном мире, а также продвижению ее ресурсов и услуг в сети, добавлению новых каналов распространения информации и улучшению сервисного обслуживания пользователей.

Основная цель настоящей работы – изучить подходы к оценке потенциальных возможностей и функциональных характеристик систем веб-статистики.

РАЗВИТИЕ СИСТЕМ ВЕБ-СТАТИСТИКИ

Системы статистики сайтов в своем развитии насчитывают несколько десятилетий – от инструментов, позволяющих считать только количество пользователей и показанных страниц, до сложных систем аналитики. Зарождение веб-аналитики относят к 1990-м гг. [3, с. 26-27]. Именно в этот период пользователи Интернета стали проявлять интерес к веб-статистике. Журналы серверов фиксировали обращение к веб-сайту и некоторую дополнительную информацию, включая имя файла, время, реферрер (веб-сайт или страница, с которой сделан запрос), IP-адрес, идентификатор браузера, операционную систему и т.д. Коммерческая веб-аналитика появилась позже, по мере усовершенствования стандартных анализаторов файлов серверных журналов и добавления функции представления данных в виде таблиц и графиков. К 2000 гг. с экспоненциальным ростом популярности Интернета, веб-аналитика твердо укрепилась как направление, в котором разрабатываются все более сложные решения для оценки эффективности и качества веб-ресурсов. Рост популярности инструментов веб-аналитики связан с доступностью, функциональностью, простотой и эффективностью. В последние годы количество инструментов веб-аналитики увеличивается, расширяются их возможности, растет популярность тех или иных счетчиков.

Основываясь на данных исследовательской лаборатории Ruward:Track, которая проводит аналитические исследования, основанные на собственной технологии количественных замеров по различным сегментам Рунета, и публикует ежеквартальный независимый рейтинг/исследование счетчиков посещаемости и систем веб-аналитики Рунета, можно сделать вывод о популярности в зоне .RU таких систем как Яндекс.Метрика, LiveInternet и Google Analytics (рис. 1). В основе их разработки лежит уникальная методика, позволяющая зафиксировать на-

хождение сервиса по его «отпечаткам». В анализе участвует около 5 млн сайтов в зоне .RU. В июне 2016 г. было опрошено 5,236 млн доменов зоны RU, из них 65,4% доменов успешно ответили [4]. Счетчики и системы веб-аналитики были обнаружены на 42,7% из числа ответивших доменов.

Чтобы получить более достоверную информацию, организации используют внутренние и внешние системы статистики посещений сайта, либо два и более варианта внешних инструментов, предоставляемых бесплатно или на коммерческой основе. Данные лаборатории Ruward:Track подтверждают этот факт (рис. 2). Более 32% анализируемых сайтов имеют два и более счетчика. Однако следует учитывать, что установка нескольких счетчиков на сайте имеет недостатки, в частности, влияет на замедление доступа посетителей к сайту. Кроме того, программные средства, имеющие разные алгоритмы сбора данных и их результаты, могут существенно отличаться. Например, в Google Analytics отказом считается просмотр пользователем только одной страницы сайта за время визита, а в системе Яндекс.Метрика – просмотр лишь одной страницы, продолжавшийся менее 15 секунд.

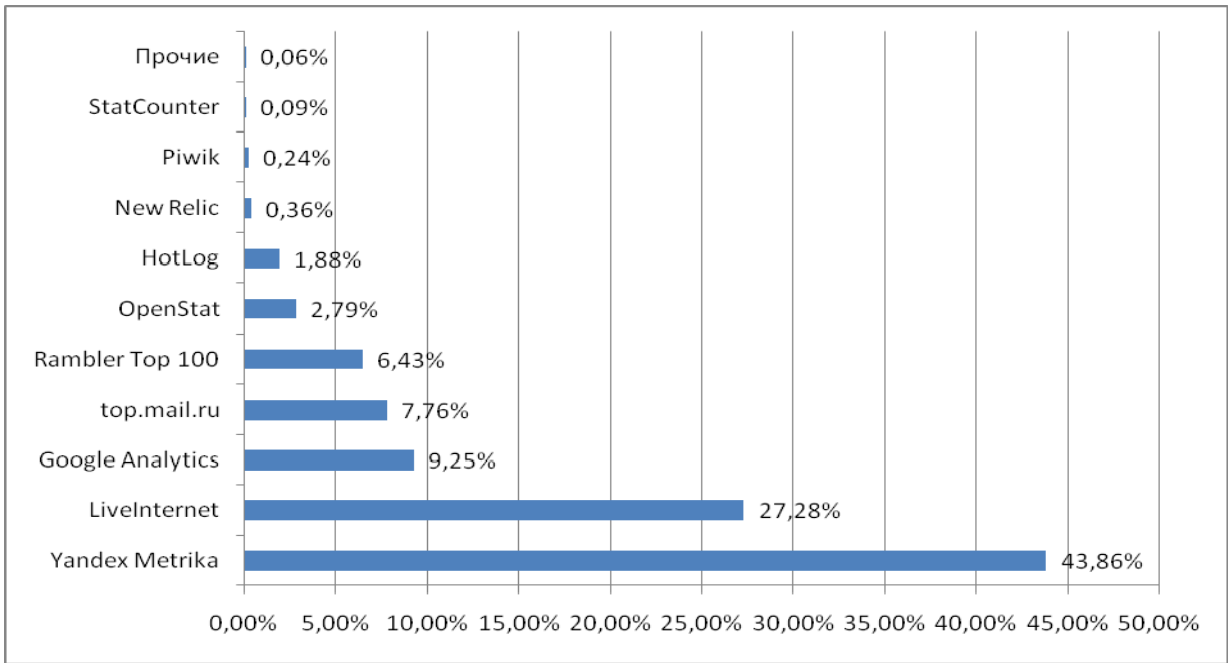

Рис. 1. Рейтинг инструментов веб-аналитики в Рунете в 2016 г.

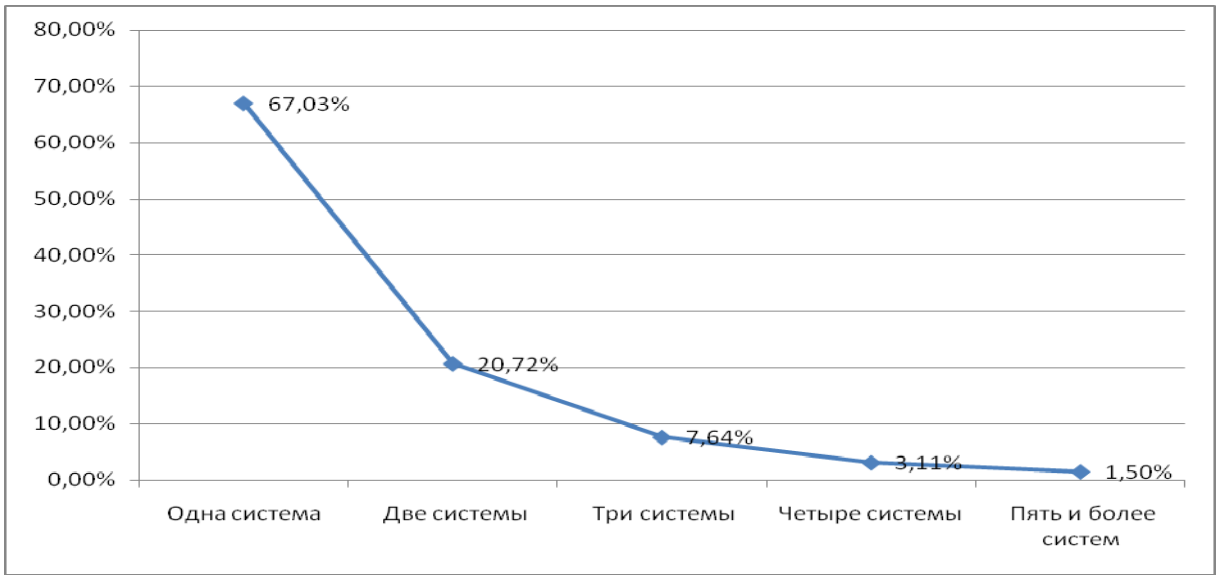

Рис. 2. Количество установленных на сайте инструментов веб-аналитики в Рунете в 2016 г.

ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ И СРАВНИТЕЛЬНЫЕ ХАРАКТЕРИСТИКИ ИНСТРУМЕНТОВ ВЕБ-АНАЛИТИКИ

Дать четкую классификацию инструментов веб-аналитики представляется сложным, учитывая, что в рейтинговой оценке и аналитических системах используют счетчики. Вместе с тем, для более четкого понимания и дальнейшего анализа современной структуры инструментов веб-аналитики, целесообразно условно выделить основные способы определения статистики посещений, используемые в веб-аналитике: плагины (к примеру, плагин статистики Word Press Stats), лог-анализиторы (Analog, Webalizer, AWStats и др.), счетчики (OpenStat (ранее Spylog), 24log.ru), рейтинги (LiveInternet, Рамблер.Топ100, top.mail.ru, Ruward, Directrix) и сервисы (Google Analytics, Яндекс.Метрика, Hotlog).

В работах отечественных и зарубежных исследователей описываются широкие возможности отдельных инструментов веб-статистики и метриксайтов [3, 5–9]. Представляют интерес публикации, в которых дано сравнение функциональных возможностей, достоинств и недостатков, показателей различных счетчиков. Изучению наиболее востребованных систем Яндекс.Метрика и Google Analytics посвящены работы К. Р. Еникеева, И.А. Задорожного, Е.С. Медведевой, М.А. Олевинского, А.В. Радзевича и др. [10–13]. В статье [14] приведено сравнение четырех систем сбора и анализа статистики посещения сайтов (Яндекс.Метрика, Google Analytics, Hotlog, LiveInternet). Более детальный анализ представлен в работе Е.Ю. Васьковского и Ю.М. Брумштейна [15], где приведена характеристика функциональности платных и бесплатных счетчиков, используемых в России: SimilarWeb, Alexa, GoogleAnalytics, OpenStat (ранее SpyLog), LiveInternet, Яндекс.Метрика, Hotlog, Clicky, Woorpa, GoSquared, Gauges, Chartbeat, StatCounter. Для перечисленных счетчиков авторы сравнивали их функциональность с разбивкой на ряд подгрупп: сводная статистика; анализ содержания и навигации; параметры веб-дизайна; анализ географии и поведения посетителей; анализ коммерческих показателей, а также дополнительные функции (список страниц, ссылающихся на ресурс, возможность определения поисковых роботов и пр.).

В каждом из рассматриваемых инструментов веб-аналитики определены достоинства и недостатки. К примеру, в Google Analytics невозможно отследить трафик в случае, если у пользователя отключены (не сохраняются) cookie-файлы; невозможно повторно обработать данные, если потерян профиль пользователя с настроенными фильтрами; есть ограничения по количеству отчетов и «целей». Увеличение числа установок Яндекс.Метрики связывают с наличием уникального инструмента аналитики – Вебвизора, позволяющего просматривать видео с действиями каждого посетителя на страницах сайта. Недостатки Яндекс.Метрики заключаются в недостаточно точной оценке трафика, невозможности настроить один счетчик для группы сайтов на поддоменах и сохранять колонки в отчетах. Популярная в России система LiveInternet дает возможность объединять все сайты в тематические группы и составлять рейтинги по посещаемости, однако, счетчик не осуществляет учет посетителей с отключенный поддержкой java-скриптов.

Можно констатировать, что пока не существует единого средства, которое могло бы комплексно решать веб-аналитические задачи и в полной мере удовлетворять потребности владельцев сайтов. При выборе инструментов веб-аналитики важно детально изучить характеристики различных счетчиков и сервисов; метрик, используемых системами; формируемые ими отчетов и ответить на следующие вопросы: Какие инструменты веб-аналитики наиболее часто устанавливаются? Каково количество сайтов, на которых установлены те или иные счетчики? Какие связки счетчиков чаще используются? Какие счетчики набирают популярность?

ПОПУЛЯРНОСТЬ ИНСТРУМЕНТОВ ВЕБ-АНАЛИТИКИ ПО ДАННЫМ *GOOGLE TRENDS*

Google Trends является публичным web-приложением корпорации Google, которое показывает, насколько популярно определенное слово или словосочетание по сравнению со всеми запросами за какой-либо период в различных регионах мира и на разных языках (https://www.google.ru/trends/). Следует учитывать, что в тренды попадают только те запросы, которые пользователи вводят в поисковой строке достаточно часто, одинаковые запросы от одного и того же пользователя за короткий промежуток времени исключаются, не учитываются запросы с апострофами или подобными знаками.

При поиске в Google Trends практически в реальном времени можно просматривать график, отражающий популярность запроса и получать сведения о показателях в тот или иной момент. Числа на графике отображают популярность слова или словосочетания по сравнению с общим количеством поисковых запросов в Google за определенный промежуток времени. Если график нисходящий, то можно констатировать снижение интереса к теме, но не во всех случаях, есть вариант, при котором популярность других растет быстрее, а устоявшиеся термины, по которым проводится анализ, растворяются в лавинно нарастающем потоке запросов. Кроме того, Google Trends позволяет с помощью вычислений предугадать популярность того или иного запроса в ближайшем будущем, т.е. прогнозировать интерес к той или иной теме. Получаемые в ходе анализа данные обезличены и нормализованы (разделены на масштабирующую переменную), значения приводятся в интервале от 0 до 100. Алгоритмы Google определяют точку на графике за выбранный период, когда запрос был наиболее популярен, и принимает его за 100. Все остальные точки на графике определяются в процентном отношении к максимуму. Если данных недостаточно, значение равно нулю. Google Trends позволяет сравнивать статистику по поисковым

запросам, регионам, периодам, на разных языках и практически в реальном времени.

Контент-анализ публикаций по теме «Самые популярные счетчики посещаемости сайтов», проведенный на базе первичных и вторичных источников информации, позволил выявить наиболее используемые в Рунете инструменты веб-аналитики. Далее этот список был проанализирован в системе Google Trends, выявлены запросы, по которым дается корректный результат с учетом географического фильтра, и определен перечень терминов для более детального анализа.

Одновременно имеется возможность сравнивать до пяти наборов слов или словосочетаний, каждый из которых содержит не более 25 поисковых запросов. Для формирования поискового предписания и дальнейшего сравнительного анализа опытным путем нами было определено пять основных терминов на английском языке: «LiveInternet», «Google Analytics», «Awstats», «Spylog» и «MetrikaYandex». Иные варианты словосочетаний, связанных с инструментами веб-аналитики, например, на русском языке, как «Яндекс.Метрика» или «Гугл Аналитикс» и другие, не позволяли получить корректные результаты в аналитической системе Google Trends. В первом поиске ограничение по регионам не устанавливалось, анализ проводился по всему охвату запросов (с 2004 г. по ноябрь 2016 г.) во всех категориях.

Наибольший интерес к «Google Analytics» зафиксирован в апреле 2013 г. Этот результат стал самым высоким показателем за весь период исследования. Далее, после повышенной заинтересованности, наблюдается тенденция стабильного интереса пользователей к Google Analytics. С 2004 г. по 2009 г. наблюдался умеренный интерес пользователей в системе «Awstats» (максимум был достигнут в ноябре 2005 г. и составил 18). Количество поисковых запросов к «LiveInternet», «Spylog» и «YandexMetrika» в общем потоке запросов не позволяет корректно выявить тренды.

Во втором случае для анализа было установлено ограничение по регионам России, поиск велся также по всему охвату запросов (с 2004 г. по ноябрь 2016 г.) во всех категориях. Среднее значение динамики популярности за рассматриваемый период составило: «LiveInternet» (33), «GoogleAnalytics» (22), «Awstats» (9), «Spylog» (5) и «YandexMetrika» (0). Корректировка запроса «YandexMetrika» на «ЯндексМетрика», «Яндекс.Метрика», «Metrika.yandex» и др., не изменил результата. Повышенный интерес был у пользователей к системе Awstats в период с 2004 по 2006 г., затем наблюдается рост запросов по LiveInternet по 2010 г., далее лидирует Google Analytics.

Более детальная оценка популярности запросов по географии пользователей представлена в табл. 1. Топ регионов и городов по запросам «Spylog» и «YandexMetrika» определить не удалось, так как данных для получения статистических форм слишком мало.

Данные свидетельствуют, что большой интерес к «LiveInternet» наблюдается у пользователей из Латвии, России, Украины, Беларусии и Казахстана (ранжирование по убывающему индексу запросов/нормализованных баллов). Сервисом «Google Analytics» интересуются жители Сан-Франциско, Лондона, Барселоны, Сиднея и Дублина (топ городов). Однако следует учитывать, что эти данные также относительны для разных регионов. Если показатели, связанные с поисковым запросом, равны в двух регионах, это не означает, что они одинаково популярны. Например, если в Германии и на Тайване доли запросов по названию «Awstats» близки, то это не означает, что пользователи из этих стран одинаковое число раз набирали это слово в поиске, а указывает на популярность терминов в общем количестве поисков. Если для какого-либо региона указан нулевой результат, это не означает, что выбранный запрос здесь совсем не использовался – он просто недостаточно популярен.

Таблица 1

Популярность основных терминов по странам и городам

Поисковые запросы					
LiveInternet		Awstats		Google Analytics	
Топ стран (балл)	*Топ городов (балл)*	*Топ стран (балл)*	*Топ городов (балл)*	*Топ стран (балл)*	*Топ городов (балл)*
Латвия (100)	Москва (100)	Швейцария (100)	Пекин (100)	Эстония (100)	Сан-Франциско (100)
Россия (88)	Донецк (77)	Нидерланды (99)	Амстердам (71)	Нидерланды (87)	Лондон (92)
Украина (81)	Нижний Новгород (72)	Китай (96)	Шанхай (67)	Великобритания (83)	Барселона (91)
Беларусь (68)	Запорожье (71)	Тайвань (81)	Франкфурт (45)	Ирландия (79)	Сидней (86)
Казахстан (52)	Санкт-Петербург (70)	Германия (80)	Брюссель (42)	Чехия (76)	Дублин (85)

ПОХОЖИЕ И СВЕРХПОПУЛЯРНЫЕ ЗАПРОСЫ

Google Trends позволяет также анализировать похожие запросы и запросы, набирающие популярность. Эти запросы включают слова и словосочетания, которые в последнее время чаще всего пользователи искали вместе с указанным словом. По каждому запросу видно, на сколько процентов он стал популярнее по сравнению с предыдущим периодом. В случае, если интерес к слову вырос более чем на 5 тыс. процентов, вместо процента появляется значение «Сверхпопулярность». Пример по анализируемому запросу «Google Analytics» приведен в табл. 2. По словосочетанию «Google Analytics» похожими запросами являются «analytics», «googleanalytics» и инструмент для создания рекламных сообщений «adwords»[1]. Наибольшее количество значений «сверхпопулярности» с трендами запросов «googletagmanager» (диспетчер тегов Google), «searchconsole» и «googlesearchconsole» (многофункциональный инструмент, позволяющий устранять технические ошибки и работать над оптимизацией сайта, отслеживании видимости и представление сайта в поисковой выдачи Google). Эта информация позволяет прогнозировать заинтересованность термином.

Таблица 2

Список похожих и трендовых запросов с терминов «GoogleAnalytics»

Топ запросов	Балл	Тренд запросов	Процент
analytics	100	googletagmanager	Сверх-популярность
googleanalytics	65	searchconsole	Сверх-популярность
adwords	5	googlesearchconsole	Сверх-популярность
webmaster	0	keywordplanner	+4 300%
google a	0	analytics.js	+3 300%

ВЫВОДЫ

Современные системы веб-аналитики – активно развивающиеся и востребованные инструменты анализа сайтов с богатым функционалом. Используя различные источники информации о тенденциях развития и возможностях различных счетчиков, сервисов и лог-анализаторов, можно выбрать наиболее оптимальный инструмент для изучения сайта конкретной организации.

[1] Используется стобалльная шкала, где 100 – самые популярные темы; 50 – темы, которые ищут в два раза реже, чем самые популярные; 0 – темы, которые ищут на 99% реже, чем самые популярные

СПИСОК ЛИТЕРАТУРЫ

1. Новое о цифровой грамотности, или россияне осваиваются в сети // Пресс-выпуск ВЦИОМ. – 2016. – № 3084. – URL: http://wciom.ru /index.php?id=236&uid=115657 (дата обращения: 13.09.2016).
2. January 2017 Web Server Survey. – URL: http://news.netcraft.com/ (дата обращения: 25.01.2017).
3. Кошик А. Веб-аналитика: анализ информации о посетителях веб-сайтов / пер. с англ. – М. : Диалектика, 2009. – 462 с.
4. Системы веб-аналитики. Июнь 2016. – URL: http://track.ruward.ru/analytics (дата обращения: 21.11.2016).
5. Яковлев А., Довжиков А. Веб-аналитика: основы, секреты, трюки. – СПб.: БХВ-Петербург, 2010. – 266 с.
6. Шляхтина С. Обзор решений для анализа посещаемости сайта. – URL: http://compress.ru/ Article.aspx?id=18178 (дата обращения: 21.11.2016).
7. Егорова И.Н., Кадушкевич О.Н. Методика эффективного использования инструментов Google Analytics // ScienceRise. – 2016. – № 1/2(18). – С. 40-43.
8. Андреева С.Л., Енгибарова М.М. Веб-аналитика как интеллектуальный инструмент // BI-технологии в оптимизации бизнес-процессов: Материалы Международной научно-практической очно-заочной конференции. – Российско-Армянский (Славянский) университет, Уральский государственный экономический университет, 2014. – С. 60-62.
9. Скородумов П.В., Холодев А.Ю. Анализ популярности веб-сайта научной организации с помощью различных систем сбора статистических данных // Вопросы территориального развития. – 2016. – № 1 (31). – С. 11-13.
10. Еникеев К.Р. Реализация системы анализа и мониторинга Web сайта // 8 Всероссийская зимняя школа-семинар аспирантов и молодых ученых «Актуальные проблемы науки и техники», Уфа, 19-20 февр., 2013. – Уфа, 2013. – Информационные и инфокоммуникационные технологии, Т. 1. – С. 127-130.
11. Задорожный И.А., Медведева Е.С. Сравнительный анализ Яндекс метрики и Google Analytics как инструментов оценки эффективности диджитал-маркетинговых коммуникаций // Научные исследования: от теории к практике. – 2015. – Т. 2, № 4 (5). – С. 161-164.
12. Олевинский М.А. Веб аналитика. Сравнение систем веб аналитики // InSitu. – 2015. – № 4. – С. 46-48.
13. Радзевич А.В. Веб-аналитика для бизнеса: как сделать правильные выводы об эффективности работы сайта // Интернет-маркетинг. – 2012. – № 4. – С. 218-225.
14. Скородумов П.В., Холодев А.Ю. Анализ подходов и инструментальных средств анализа

статистики посещения веб-сайта научной организации // Вопросы территориального развития. – 2015. – № 9 (29). – С. 11-12.
15. Васьковский Е.Ю., Брумштейн Ю.М. Системный анализ функциональных возможностей счетчиков посещаемости сайтов // Прикаспийский журнал: управление и высокие технологии. – 2015. – № 3. – С. 96-113.

Материал поступил в редакцию 31.01.17.

Сведения об авторе

РЕДЬКИНА Наталья Степановна – доктор педагогических наук, заместитель директора по научной работе Государственной публичной научно-технической библиотеки Сибирского отделения Российской академии наук (ГПНТБ СО РАН), профессор Новосибирского государственного педагогического университета
e-mail: to@spsl.nsc.ru

УДК 004.6.056 : 004.75

А.В. Зайцев, С.С. Гостев, П.А. Черкашин, А.Ю. Щербаков

О технологии распределенного хранения конфиденциальной информации в центрах обработки данных общего назначения*

Обсуждается проблема разработки проекта архитектуры и математической модели хранения конфиденциальной информации в центрах обработки данных общего назначения.

Ключевые слова: *центр обработки данных, распределенное хранение данных, информационная безопасность*

ВВЕДЕНИЕ

Специалисты исследовательской лаборатории армии США Александр Котт, Анантрам Свами и Брюс Вест в ноябре 2016 г. опубликовали статью «Туман войны в киберпространстве» (Kott A., Swami A., West Bruce. J. The Fog of War in Cyberspace // Article in Computer. – November, 2016. DOI: 10.1109/MC.2016.333), в которой анонсировали подход к обеспечению безопасности, когда данные разделяются на многочисленные фрагменты и постоянно распределяются по нескольким конечным вычислительным устройствам. Этот подход, как указывают авторы, может не только обеспечить большую отказоустойчивость информационных систем и предотвратить атаки, но и предоставить огромные технические проблемы. Название этой статьи породило название технологии «Кибертуман», которая нуждается в осмыслении и моделировании, а также в реальной статистической оценке. Выдающийся военный теоретик Карл фон Клаузевиц писал о «тумане войны», понимая его как фундаментальную неопределенность информации в сложной состязательности воюющих сторон (Клаузевиц К. О войне. – М.: Изд-ва: Эксмо, Мидгард, 2007. – 458 с.)

ПАРАМЕТРЫ И ЭЛЕМЕНТЫ АРХИТЕКТУРЫ ХРАНЕНИЯ

Система распределенного хранения информации (СРХИ, *далее* – система хранения) состоит из поставщиков информации, потребителей информации (в общем случае поставщики и потребители могут быть одинаковы), Центра управления распределенным хранением (ЦУРХ, *далее* – центр управления) и центров обработки данных (ЦОД). Поставщики направляют в центр управления информационные массивы, которые он распределяет по центрам обработки данных. При запросе на информационный массив центр управления «собирает» его из фрагментов, хранящихся в центрах обработки данных, и направляет потребителям. Информационный массив (или распределяемый информационный массив – РИМ) определяется именем I и длиной L, с ним также связаны свойства – конфиденциальность содержащейся в нем информации и необходимость зашифрования данного массива.

Система хранения характеризует следующие ключевые параметры:

1) общее число центров обработки данных – N;

2) число активных ЦОД на текущий момент времени – n;

3) ЦОД идентифицируется номером i, i= 1,…, N, множество активных ЦОД образует выборку мощностью n из N возможных и характеризуется вектором $A = (a_1, a_2,\ldots, a_N)$, где a_k принимает значение 0, если k-й ЦОД неактивен и 1 – если он активен, сумма компонентов вектора равна n;

4) безопасная доля информации B, $0<B<1$ (понятие, связанное с грифом конфиденциальности информации – по аналогии с инструкцией по закрытому делопроизводству: если документ имеет менее 10% (0,1) объема информации j-го грифа, то он относится к j–1-му грифу, из этого следует, что минимальное количество m ЦОД для распределения m = [1/B]+1, где в квадратных скобках результат вычисления выражения является целой частью числа, находящегося в квадратных скобках;

5) длина единичного блока e байт, на которые разбивается распределяемый информационный массив;

6) Центр управления – обязательный участник схемы, хранящий Главный файл распределенного массива (*далее* – главный файл) и описывающий расположение и свойства распределяемого информационного массива;

7) главный файл содержит: реальное имя распределяемого информационного массива, гриф конфиденциальности, ссылку на ключ шифра (если он тре-

* Публикация подготовлена в рамках работ по программе фундаментальных исследований Отделения математических наук РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения», а также работ, поддержанных РФФИ, грант № 15-07-08522.

буется), цепочку номеров центров обработки данных и идентификатор текущего обрабатываемого блока информации – ID в рамках этих центров (это может быть имя файла, в который записан текущий блок), а также индикаторы «архитектурной» избыточности.

В эту схему входят два датчика случайных чисел (ДСЧ): первый, генерирующий равномерно распределенную последовательность чисел, каждое из которых принимает значение от 1 до N, и второй, назначение которого мы поясним далее, а также код, исправляющий ошибки (КИО), который эмпирически должен создавать избыточность порядка одной трети длины e в байтах. Обозначим преобразование этого кода как $h=KIO(e)$.

С учетом возможной недоступности некоторых центров обработки данных в распределяемую информацию необходимо вводить «архитектурную» избыточность, т.е. дублировать распределяемый массив в нескольких ЦОДах. Статистическую оценку этой величины представим как $(N-n)/N$ – вероятность того, что в некоторый момент времени центр обработки данных недоступен и записанный в его массивах хранения блок не может быть прочитан – это минимальная вероятность потери блока. Соответственно, приблизительно с такой же вероятностью очередной блок надо записывать еще в один ЦОД. Таким образом, второй датчик случайных чисел (ДСЧ-2) в каждом текущем распределяемом блоке с заданной вероятностью $p>(N-n)/N$ должен давать команду на дублирование этого блока в другом центре обработки данных.

Примерный алгоритм работы Центра управления распределенным хранением:

1) информационный массив разделяется на x блоков по e байт;

2) циклически (x циклов) проводится процедура обработки массива по блокам;

3) текущему блоку вырабатывается корректирующая последовательность h;

4) вырабатывается случайное число i центров обработки данных и приводится к модулю n;

5) выбирается i-й ЦОД и в него записываются блоки e и h;

6) на i-м центре проверяется h и записывается блок e, центру управления передается ID блока, сохраненного в i-м центре обработки данных;

7) с помощью ДСЧ-2 с заданной вероятностью p вырабатывается команда архитектурного дублирования и повторяются шаги 4-6 для другого центра обработки данных;

8) в главный файл распределенного массива добавляется номер центра обработки данных и ID блока (при архитектурном дублировании два раза).

При этом должны быть предусмотрены счетчики, которые предупреждают излишнюю запись в центрах обработки данных (когда превышена безопасная доля для конкретного центра), в этом случае i-й ЦОД в векторе a становится «временно неактивным» ($a_i=0$). В случае включения шифра вопрос о безопасной доле, естественно, снимается. Описанный алгоритм и технологию далее мы предлагаем называть «Кибертуман».

ОСНОВНЫЕ ТЕНДЕНЦИИ РАЗВИТИЯ СОВРЕМЕННЫХ ТЕХНОЛОГИЙ ХРАНЕНИЯ ДАННЫХ

Более 2-х квинтиллионов байт (эксабайт) генерируются каждый день по всему земному шару, из них 85% – неструктурированные (медиа-файлы: аудио, фото, видео, электронная почта, коммерческая документация, социальные сети и т.д.). Безопасный и надежный способ хранения постоянно растущих объектов данных является одним из основных вопросов ИТ-директоров. Для решения этой проблемы многие делают выбор в пользу облачных решений.

Объем рынка облачного хранения данных увеличится с $23,76 млрд в 2016 г. до $74,94 млрд к 2021 г. (исследование Markets and Markets). Таким образом, этот рынок будет расти на 25,8% ежегодно.

Большинство современных технологий хранения данных основано на увеличении количества копий данных. Компании делают несколько резервных копий для увеличения надежности. Для предотвращения утечек конфиденциальной информации их шифруют и устанавливают к ним многоуровневый контроль доступа.

Снижение затрат

Помимо видимых преимуществ в отношении надежности, безопасности и скорости, система распределенного хранения информации экономит еще и объем памяти. Увеличение количества копий является обычной практикой для компаний, понимающих недостаток применения обычных RAID-массивов (Redundant Array of Independent Disks – избыточный массив независимых дисков).

Копии стоят дорого: каждая копия требует более 133% дополнительного объема дискового пространства, в случае использования стандартной конфигурации RAID 6. Корпорации часто прибегают к территориально-распределенному хранению данных. Некоторые компании делают по 2, 3 или даже 4 копии.

При классическом подходе увеличение затрат на 500% повышает надежность так же, как и при использовании системы распределенного хранения информации.

Превосходя крупных провайдеров облачных систем хранения данных в надежности, безопасности и скорости, система распределенного хранения информации сравнима по цене с двумя или более узлами хранения.

Масштабируемость хранилищ данных

Основной проблемой большинства облачных хранилищ является ограничение размера одного файла, как правило, это сотни мегабайт. В свете растущих потребностей хранения больших данных (BIGDATA), система распределенного хранения информации позволяет без изменения существующей инфраструктуры получить практически не ограниченное по объёму и скорости хранилище данных.

Повышение надежности и доступности систем хранения данных

В классических системах хранения данных каждая дополнительная копия дает приближение к 100% надежности. Добавление еще 30% избыточности описан-

ного алгоритма работы центра управления распределенным хранением означает, что данные будут доступны и защищены от потерь, даже если 30 из 100 узлов памяти одновременно выйдут из строя. Это соответствует 24-м девяткам надежности (99,99999999999999999999%) и 12-ти девяткам доступности.

Узлы хранения данных могут быть распределены по десяткам и сотням географических местоположений. Эта отказоустойчивая система обеспечивает беспрецедентную защиту от какого-либо внешнего или внутреннего отказа.

Безопасность

Типичные системы хранения данных применяют сложный контроль доступа и шифрование для обеспечения безопасности и предотвращения утечек. В случае несанкционированного доступа на уровне администратора все данные в системе находятся под угрозой.

«Кибертуман» гарантирует защиту данных от кражи, утечки или несанкционированного доступа. Любой файл воспринимается программой как поток байтов. Этот поток разделяется на блоки данных, которые смешиваются с закодированными маркерами для обеспечения избыточности и «распыляются» на сотни территориально-распределенных узлов. Отдельный узел хранит не весь исходный файл – а только небольшую закодированную его часть. Каждый блок имеет закрытый ключ доступа. Если случайно злоумышленники смогут взломать один или несколько узлов, то они получат только набор блоков (фрагментов) данных, недостаточных для понимания полного содержания файла. Реконструкция всех блоков (фрагментов) может быть выполнена только владельцем данных, который имеет главный ключ. Без ключа злоумышленник не сможет узнать алгоритм соединения блоков (фрагментов), так что даже если найти соответствующие фрагменты в других узлах, их корректное соединение будет крайне сложным.

Увеличение скорости передачи данных

Достижение высокой скорости передачи данных при перемещении их от одного узла хранения или от центра обработки данных маловероятно. Причина проста - один центр обработки данных имеет много клиентов. «Кибертуман» уравнивает этот дисбаланс: один клиент обслуживается десятками узлов. Каждый узел, используемый алгоритмом с точки зрения пропускной способности канала не лучше или не хуже, чем любой другой типичный современный центр обработки данных. Но при их объединении они обеспечивают увеличение скорости более чем в 10 раз по сравнению с любыми другими системами, даже если они будут использоваться независимо друг от друга.

Практические тесты скорости облачных хранилищ показали, что увеличение скорости осуществляется не только при передаче данных через множество разных каналов связи, но и при передаче файла частями с использованием множества параллельных потоков.

Дополнительный прирост скорости можно получить применив несколько операторов связи при передаче данных. Перехватить весь объём данных при передаче через один канал значительно проще, чем при передаче данных через несколько независимых каналов операторов.

Причем можно использовать гибридные виды связи (GPRG, LTE, WIFi, 3G и прочие) одновременно, что особенно актуально для регионов, в которых отсутствуют высокоскоростные каналы связи. Применение такой технологии позволяет одновременно использовать все доступные виды связи и многократно поднять скорость и безопасность передачи данных.

Преимущества технологии «Кибертуман»

1. Система хранения данных на базе технологии «Кибертуман» может быть создана с использованием существующих ресурсов государственных и коммерческих центров обработки данных или узлов хранения.
2. Для достижения максимальной территориальной распространенности описанный алгоритм работы центра управления распределенным хранением сохраняет различные блоки (фрагменты) кодированной информации в разных узлах.
3. Каждый узел содержит минимум 5 серверов хранения данных, которые изначально оснащены 5 ТБ каждый, объем легко модифицируется до 45 ТБ.
4. Общий объем 200 таких узлов хранения данных будет составлять 5 петабайт, которые легко масштабируются до 45 петабайт без добавления новых серверов.
5. Узлы хранения территориально распределены и будут обеспечивать стабильный быстрый доступ к хранилищу данных в режиме 24/7.

ЗАКЛЮЧЕНИЕ

Рассмотренный набор параметров и архитектурных решений может быть основой для разработки общих и ведомственных положений и классификаций, а также выбора и оценки технических решений в области архитектуры и математической модели хранения конфиденциальной информации в центрах обработки данных общего назначения для широкого круга информационных систем.

Материал поступил в редакцию 20.03.17.

Сведения об авторах

ЗАЙЦЕВ Андрей Викторович – заместитель директора по технической части ООО «МЕДИА», Москва
e-mail: avzajcev@mail.ru

ГОСТЕВ Сергей Сергеевич – кандидат технических наук, заместитель генерального директора по науке, «Концерн ГРАНИТ», Москва
e-mail: gostevss@mail.ru

ЧЕРКАШИН Павел Александрович – заместитель генерального директора по информационным технологиям, «Концерн ГРАНИТ», Москва
e-mail: cherkashin@granit-concern.ru

ЩЕРБАКОВ Андрей Юрьевич – доктор технических наук, профессор НИУ ВШЭ, главный научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН, Москва
e-mail: x509@ras.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322’373 : 004.85

В.А. Можарова, Н.В. Лукашевич

Исследование признаков для извлечения именованных сущностей из текстов на русском языке*

Рассматриваются различные признаки для извлечения именованных сущностей из текстов на русском языке, применяемые в рамках подходов на основе машинного обучения, включая признаки собственно токена (лексемы), а также словарные, контекстные, кластерные и двухэтапные признаки. Исследуется вклад каждого признака в улучшение качества извлечения именованных сущностей. В качестве метода машинного обучения в экспериментах, описанных в данной статье, используется CRF-классификатор. Сравнение вкладов признаков произведено на двух открытых коллекциях с помощью F-меры.

Ключевые слова: *именованная сущность, извлечение информации, машинное обучение*

ВВЕДЕНИЕ

Извлечение информации из текстов – одна из самых востребованных задач в обработке естественного языка. Наиболее частым источником для ее решения являются новостные тексты, из которых извлекаются несколько типов информации. Первый тип – это именованные сущности, такие как имена людей, названия компаний и географических объектов. Следующий тип – это отношения между именованными сущностями, например, должность сотрудника в организации. Третий тип – это события, которые происходят с объектами, соответствующими именованным сущностям, например, слияние компаний, покупка акций или бизнес-встречи. Такая информация используется в разного рода информационно-поисковых и информационно-аналитических системах, а также в качестве предварительного этапа для других типов автоматической обработки текстов.

Для извлечения именованных сущностей используются два основных подхода. Первый – называется инженерным подходом, поскольку он основан на создании словарей и правил извлечения вручную ([1-3]).

Второй подход использует методы машинного обучения. Для таких методов необходима подготовка обучающей коллекции, в которой размечены необходимые для выделения сущности. Для работы систем машинного обучения размеченные данные должны быть преобразованы в наборы признаков для каждого токена (лексемы) (его написание, контекст и др.).

Большое количество статей было посвящено исследованию вклада различных признаков в извлечение именованных сущностей для английского языка. В то же время для русского языка многие аспекты использования этих признаков оказались не изучены. Цель настоящей работы – рассмотрение важности различных признаков для достижения лучшего качества извлечения именованных сущностей на основе тестирования на открытых текстовых коллекциях. В качестве метода машинного обучения был выбран метод *CRF* (*Conditional Random Fields*) [4], который предназначен для обработки последовательных данных и часто используется для извлечения имен из текстов.

БЛИЗКИЕ РАБОТЫ

Методы машинного обучения являются часто применяемыми подходами в решении задач извлечения именованных сущностей. Признаки токенов (лексем), которые используются в этих методах, можно подразделить на следующие категории: признаки собственно токена (символьный состав, часть речи, написание и др.); контекстные признаки, кото-

* Работа частично поддержана фондом РФФИ (проекты 15-07-09306 и 16-29-09606).

рые используют информацию о соседних токенах; признаки, основанные на неоднократном упоминании токена в других частях текста или коллекции, а также признаки, основанные на знаниях, т.е. на собранных заранее словарях или автоматически порожденных кластерах близких по смыслу слов и т.п. [5, 6].

В работе [5] для английского языка исследуется вклад признаков, основанных на знаниях (словари и автоматически порожденные кластеры слов), а также два варианта двухэтапного анализа текстов для получения дополнительной информации о классификации других вхождений данного токена. Используя двухэтапный подход, авторы достигли 90,57% *F*-меры (из них вклад двухэтапных признаков составил 2,88%) на текстовой коллекции *CoNLL03*[1] [7], в которой были размечены четыре вида именованных сущностей (персоны, локации, организации и «другие» именованные сущности). Описанная система также показала 89,19% *F*-меры (вклад второго этапа анализа составил 2,33%) на тестовой коллекции *MUC7*[2], в которой были размечены три вида именованных сущностей (персоны, локации и организации).

В [6] авторы проводили эксперименты для чешского языка с 42 типами именованных сущностей. Чтобы извлечь их была использована система, основанная на методе максимальной энтропии. Система включала в себя два этапа классификации: второй этап использовал результаты первого. Чтобы сформировать признаки, авторы задействовали большое количество словарей и кластеры типа *Brown clusters* [8], автоматически порожденные на текстовом корпусе.

В работе [9] представлена система для английского языка, в которой токены, классифицированные как именованные сущности на первом проходе, использовались для порождения признаков – на втором. Авторы не обнаружили улучшений при добавлении двухэтапного подхода и достигли 91,02% *F*-меры на тестовой коллекции CoNLL03.

Для открытых коллекций на польском языке *CZER*, *CEN* и *CPR* авторы [10] применили метод *CRF* для распознавания пяти типов именованных сущностей (имена, фамилии, страны, города и дороги). Они использовали признаки, основанные на орфографии, морфологии, семантической сети *WordNet*[3] и словарях.

Существует несколько работ для русского языка, основанных на методе *CRF*.

В [11] авторы представили результаты применения *CRF* в различных задачах, включая распознавание именованных сущностей. Эксперименты проводились на их собственном русскоязычном корпусе текстов, который состоял из 71 тыс. предложений. Использовались только признаки, основанные на *N*-граммах и орфографии, без дополнительных словарей. Был получен результат 89,89% *F*-меры на трех типах именованных сущностей: имена людей (93,15%), географические объекты (92,7%) и организации (83,83%).

В работе [12] эксперименты основывались на русской текстовой коллекции *Persons-600* для задачи распознавания имен. Авторы также выбрали для этой цели метод *CRF* и использовали признаки, основанные на информации о токене, контексте и словарях, содержащих знания о людях (профессии, роли, должности и т. д.). Было достигнуто 88% *F*-меры задачи извлечения имен людей.

В [13] эксперименты проводились на русской текстовой коллекции, содержащей 97 документов. Авторы применили два подхода к извлечению именованных сущностей: подход, основанный на знаниях, и подход, основанный на *CRF*. При использовании метода машинного обучения были взяты признаки токена и признаки, основанные на кластеризации (*Brown clusters* [8], LDA [14], *Clark clusters* [15]). Было достигнуто 75,05% значения *F*-меры на двух типах именованных сущностей: персоны (84,84%) и организации (71,31%).

ТЕКСТОВЫЕ КОЛЛЕКЦИИ

Для русского языка нет достаточного количества текстовых коллекций, размеченных именованными сущностями. Те коллекции, которые открыты для использования, либо содержат немного типов именованных сущностей, либо имеют небольшой общий объем, что не подходит для методов машинного обучения и качественного тестирования. Так, в недавно организованном тестировании подходов по извлечению информации из текстов *FactRuEval* , в качестве набора данных для настройки алгоритмов участникам были выданы только 122 размеченных новостных документа (табл. 1) [16].

Таблица 1

Количество именованных сущностей в различных коллекциях

Тип именованной сущности	*FactRuEval Dev*	*FactRuEval Test*	*Persons1000*	*Persons1111*
Person	728	1350	10623	5693
Organization	942	1537	8541	-
Location	661	1284	7244	-
Overall	2331	4171	26408	-

Чтобы извлечь именованные сущности, в настоящей работе были проведены эксперименты на двух открытых текстовых коллекциях. Первая коллекция *Persons-1000* содержит 1 тыс.[4] новостных документов с размеченными именами персон. Эта коллекция

[1] http://www.cnts.ua.ac.be/conll2003/ner/

[2] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

[3] https://wordnet.princeton.edu/

[4] http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000

была размечена в Исследовательском центре искусственного интеллекта [17].

Рассматриваемая в нашей работе коллекция была дополнительно размечена другими типами именованных сущностей:

- Организации (*ORG*);
- Медиа-организации, имеющие функции информирования (*MEDIA*);
- Географические объекты (*LOC*);
- Геополитические объекты (*GEOPOLIT*) – страны и столицы, выступающие в роли правительства (например, «Москва анонсировала»).

Особенностью новой разметки является то, что в подавляющем числе случаев размеченное имя должно начинаться с заглавной буквы. Важными принципами разметки, которая сделана в соответствии с принципами разметки конференций *MUC* и *CONLL*, являются следующие:

- в разметке нет вложенных именованных сущностей,
- именованные сущности не могут пересекаться,
- каждому токену соответствует не более одного класса разметки.

Для доразметки коллекции «Persons-1000», был использован инструмент *Brat annotation tool*[5] (рис. 1), который удобен для разметки текста. Правила, согласно которым были размечены новые именованные сущности, подробно описаны в [18].

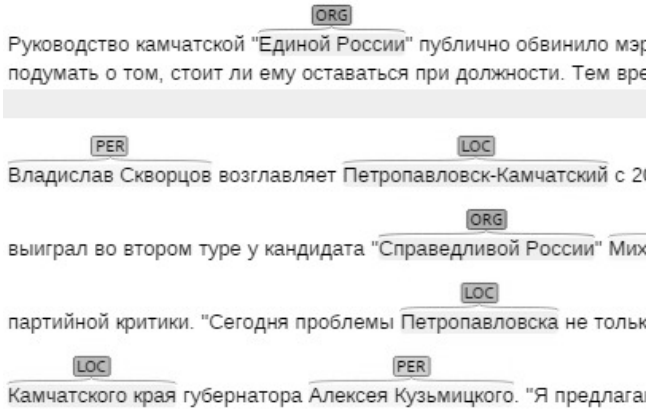

Рис. 1. Интерфейс инструмента Brat Annotation Tool

Для экспериментов, представленных в настоящей статье, были использованы только три вида именованных сущностей: персоны, организации и географические объекты[6].

Вторая коллекция *Persons-1111-F*[7] содержит 1111 новостных документов с размеченными именами персон. Эта коллекция была специально собрана из документов, в которых упоминаются сложные для анализа восточные имена, такие как арабские, индийские, китайские и японские, что может негативно повлиять на качество распознавания именованных сущностей.

Данные по количеству размеченных именованных сущностей в коллекциях *Persons-1000* и *Persons-1111-F* представлены в табл. 1.

Для извлечения именованных сущностей из текстов на русском языке в качестве метода машинного обучения был использован *CRF*-классификатор, который часто используется в приложениях, связанных с извлечением этих сущностей. Этот метод был создан специально для классификации последовательных данных. В его основе лежит марковская сеть, в которой вершины делятся на скрытые (x_t) и наблюдаемые (y_t) (рис. 2).

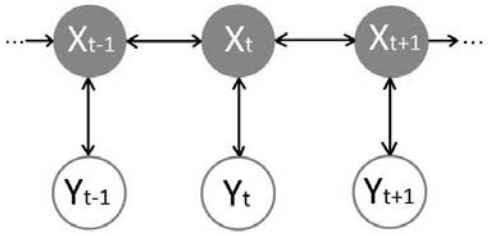

Рис 2**.** Схема зависимостей переменных в методе *CRF*

Скрытые состояния соответствуют типам именованных сущностей, т. е. тому, что мы хотим предсказать; а наблюдаемые состояния – признакам токенов. В нашей работе был использован *CRF++*[8], являющийся готовой реализацией этого метода с легко и быстро настраиваемым открытым кодом.

ПРИЗНАКИ ДЛЯ ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ

Перед извлечением именованных сущностей текстовые коллекции обрабатывались морфологическим анализатором. Для каждого токена были определены его лемма, часть речи, а также грамматические характеристики (род, число, падеж и др.). Эта информация используется в дальнейшем для формирования признаков. В нашей работе для классификации токена используются пять типов признаков: признаки собственно токена; контекстные признаки; признаки, основанные на повторах токенов в разных контекстах; признаки, основанные на больших словарях имен и других типов слов, а также автоматически сформированные кластеры слов, которые употребляются в похожих контекстах.

Признаки токена

Признаки токена являются самыми употребительными при извлечении именованных сущностей. Мы используем следующий набор признаков токена.

1. Начальная форма слова, например: *сказал – сказать*, *оборонительных – оборонительный*.

[5] http://brat.nlplab.org/
[6] http://www.labinform.ru/pub/named_entities/collection3.zip
[7] http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000
[8] https://taku910.github.io/crfpp

2. Количество символов в токене. Например: *девушка* – 7, *возле* – 5.

3. Регистр букв. Если токен начинается с большой буквы, а остальные буквы маленькие, тогда значение этого признака будет *BigSmall.* Если все буквы большие, тогда значение – *Big*. Если все буквы маленькие, значение признака – *Small*, в остальных случаях – *Fence*.

4. Тип токена. Для лексем значение этого признака будет приравниваться к части речи. Для пунктуационных знаков – к типу пунктуации.

5. Наличие гласной (бинарный признак).

6. Является ли токен концом предложения (бинарный признак).

7. Содержит ли токен известную *N*-грамму из специального списка:

а) окончания фамилий (*-ов, -ова, -енко, -швили* и т.д.),

б) начала имен,

в) N-граммы, свойственные организациям (*-ком-, -орг-, -деп-* и т. д.).

Признаки, основанные на словарях

Чтобы улучшить качество распознавания, были добавлены словари, хранящие списки полезных объектов. Объектом может быть как слово, так и словосочетание.

Для каждого токена система определяет, является ли он знакомым словом или входит в известное словосочетание. Длина словосочетания также включается в значение признака чтобы различать, когда слово вошло в длинное словосочетание, а когда оно встретилось одиночно, что может говорить о случайном сопоставлении из-за омонимии. Например, в списках организаций есть название партии «Яблоко», однако, необязательно, что слово «яблоко», встретившееся в тексте, будет принадлежать классу организаций. В то же время, если в тексте встретилось словосочетание «Московский государственный университет им. Ломоносова», которое также есть и в списках организаций, то вероятность того, что это действительно организация, будет гораздо выше, чем в первом примере. Именно учет длины словосочетания помогает разделить эти два случая и придать им разную степень уверенности.

Наибольшие и значимые словари представлены в табл. 2. Общий размер словарей составляет около 335 тысяч объектов. Для создания этих словарей использовались телефонные справочники, Википедия и тезаурус *РуТез* [19].

Признаки, основанные на контексте употребления токена

При классификации каждого слова огромную роль может сыграть его контекст. Он поможет системе учитывать статистику совместного появления слов и их признаков. Например, в табл. 2, посвященной применяемым словарям, упоминается список глаголов информирования. Если проанализировать новостные документы, можно заметить, что глаголы из этого словаря часто встречаются вместе с персонами и медийными организациями. Рассматриваемые признаки помогают выявить такие закономерности в текстовой коллекции.

Группа контекстных признаков включает два признака. Признак ***context*** учитывает значения всех признаков соседних токенов. Например, для рассматриваемого слова будут приниматься во внимание все части речи, в написание которых вошли предыдущие и следующие два слова.

Второй контекстный признак ***bigram*** содержит информацию об уже вычисленном классе предыдущего токена, что помогает точнее определять границы именованных сущностей. Например, если фамилия человека сложна для распознавания, то наличие имени перед ней облегчает эту задачу.

Кластеризация слов с помощью WORD2VEC

При применении программы извлечения именованных сущностей может оказаться, что рядом с именем встречаются слова, которые ни разу не встречались в обучающей размеченной выборке. Это может привести к ухудшению качества извлечения. Чтобы снизить влияние этого фактора применяются различные подходы, позволяющие группировать (кластеризовывать) слова, которые употребляются в похожих контекстах. При этом предполагается, что употребление слов в похожих контекстах отражает их смысловое сходство.

Таблица 2

Словари

Словарь	Размер	Примеры
Известные люди	31482	*Владимир Путин, Ангела Меркель*
Имена	2773	*Василий, Анна, Том*
Фамилии	66108	*Кузнецов, Грибоедов*
Профессии	9935	*министр, китаевед*
Глаголы информирования	1729	*высказать, признаться, отпроситься*
Компании	33380	*Сбербанк*
Типы компаний	6774	*организация, ООО, авиафирма*
Медиа организации	3909	*РИА Новости, Первый канал*
Географические объекты	8969	*Балтийское море, Владивосток*
Географические прилагательные	1739	*финский, томский, югославский*
Частотные слова	58432	*автомобиль, падать, желтый*
Устройства	44094	*компьютер, телефон*

В настоящей работе для кластеризации слов применяется относительно новый, но уже очень популярный пакет *word2vec*[9]. Этот пакет производит сокращение матрицы слов и их контекстов текстовой коллекции путем представления этой матрицы заданным числом скрытых измерений (*word embeddings*) на основе применения специализированной нейронной сети. Каждое слово представляется как вектор в сокращенном пространстве. Семантическое сходство между словами вычисляется как косинусная мера между векторами. Пакет имеет встроенную возможность разбиения всего множества слов коллекции на кластеры.

Пакет *word2vec* был применен к коллекции в 2 млн новостных документов. Была использована модель *cbow* с размерностью векторов 300. Таким образом, каждый токен имеет теперь дополнительный признак – номер кластера, в котором он состоит. В табл. 3 указаны примеры из трех кластеров, полученных на этой коллекции.

Таблица 3

Примеры кластеров

Кластер 1	Кластер 2	Кластер 3
США	*Москва*	*Художественный*
Американский	*Санкт-Петербург*	*Авторский*
Американец	*Смольный*	*Мастер-класс*
Барак	*Мэр*	*Грандиозный*
Обама	*Горадминистрация*	*Фотовыставка*
Клинтон	*Госстройнадзор*	*Конкурсант*
Вашингтон	*Столица*	*Выставка-ярмарка*

Поскольку с помощью такой кластеризации вычисляются совокупности близких слов, возникает важный вопрос, насколько нужны ручные словари и основанные на них признаки, так как, по сути, ручные словари также представляют собой некоторые смысловые группировки слов.

Признаки двухэтапного предсказания

При формировании признаков двухэтапного предсказания было сделано предположение, что для лучшей классификации полезно учитывать предыдущий опыт системы и запоминать статистику классов для дальнейшего использования.

На первом этапе классификатор извлекает именованные сущности. Далее система собирает статистику классов, полученных после первого этапа, и преобразовывает ее в новый набор признаков, который используется новым классификатором наряду со старыми признаками. Эту статистику можно собирать не только по всей коллекции, но и по определенным ее фрагментам. Нами были изучены следующие типы признаков второго этапа, отличающиеся исключительно участками текстовой коллекции, использованными для сбора статистики: история предсказаний, использование статистики внутри документа и использование статистики внутри всей текстовой коллекции.

Признак **история предсказаний** формируется исходя из предположения, что в начале текста имена обычно употребляются в полной форме, и классификатору легче распознать именованную сущность. Например, фамилия рядом с именем или отчеством легче определяется, если она стоит отдельно. Если рассмотреть текст, в котором словосочетание «Геннадий Столяр» встретилось раньше, чем отдельно слово «Столяр», то второе упоминание будет проще распознать как персону, потому что оно уже было так классифицировано ранее.

Для каждого токена система находит все предыдущие его вхождения в текст и считает статистику присвоенных ему классов. Основываясь на этой статистике, система создает для каждого класса дополнительные признаки, которые получают одно из трех значений: *no_one* (если токен ни разу не был отнесен к этому классу), *best* (если токену был проставлен этот класс более чем в 50% случаев) и *rare* (если этот токен попадал в этот класс недостаточно часто). Например, если токен «Россия» встретился пять раз в тексте, и классификатор определил его два раза в класс организаций и три раза – в класс географических объектов, то значения признаков для шестого токена «Россия» будут следующими: PER – *no_one*, ORG – *rare*, LOC – *best*.

Признак **использование статистики документа** похож на предыдущий, и отличается от него только тем, что статистика считается в целом документе, а не только в предшествующей токену части документа.

Признак **статистики текстовой коллекции** вычисляется на основе применения такого же подхода к некоторому заданному набору документов.

ЭКСПЕРИМЕНТЫ

В качестве представления разметки для обучения классификатора использовалось *BIO*-представление, в котором токены размечаются тремя типами меток для каждой категории извлекаемых имен: начало именованной сущности, продолжение именованной сущности и неименованная сущность [5, 18]. Например, в предложении «Владимир Путин поздравил россиян с праздником» токен «Владимир» получит метку «B-PER» (начало именованной сущности), токен «Путин» получит метку «I-PER» (продолжение именованной сущности), а все остальные токены будут помечены меткой «OUT» (не принадлежит никакой именованной сущности). В работе [18] было проведено сравнение этого представления вместе с *IO*-представлением и было показано, что представление *BIO* больше подходит для текстов на русском языке, и с помощью него можно достичь лучших результатов в задаче извлечения именованных сущностей.

Целью нашей работы было протестировать различные виды признаков для извлечения именованных сущностей. Для этой задачи были использованы две открытые коллекции, упомянутые ранее. Система

[9] https://github.com/dav/word2vec

по очереди запускалась с разными типами признаков, а затем с их комбинациями. Сначала система была протестирована на коллекции *Persons-100* с использованием кросс-валидации и соотношением обучающей и тестовой выборки 3:1. Затем, чтобы проверить возможность перенесения система извлечения имен на другие коллекции, классификатор был обучен на коллекции *Persons-1000* и применен на коллекции *Persons-1111*, особенностью которой является то, что она составлена из текстов, содержащих восточные имена, относящиеся к разным странам и народам.

Результаты работы системы (*F*-мера) на разных наборах признаков представлены в табл. 4. В качестве базового, самого простого уровня классификатора (1), рассматривается набор признаков токена без учета проверки на вхождение заданных *n*-грамм, и контекстный признак, который рассматривает эти же признаки у двух соседних слов – слева и справа.

На следующем уровне добавляется биграммный признак, который учитывает класс предыдущего токена (2). Видно, что происходит значительный рост качества выделения имен, т.е. используя его, классификатору легче выделять границы именованных сущностей и распознавать незнакомые и нетривиальные части именованных сущностей. Далее в набор признаков (2) добавляется учет автоматически полученных кластеров (3) и видно, что дальнейший рост качества извлечения значителен.

Добавление отдельных признаков (4-6) второго этапа анализа также немного увеличивает качество извлечения для обеих коллекций и типов имен. Комбинация всех признаков второго этапа (7) в среднем немного лучше добавления каждого отдельного такого признака. Признаки второго этапа вносят сравнительно небольшой вклад, но, если просмотреть обработанные тексты, можно увидеть, что исчезли многие именованные сущности, которые были неправильно выделены на первом этапе. Таким образом, двухэтапный подход к извлечению именованных сущностей, в первую очередь, улучшает точность классификации [20].

Также, если подробнее рассмотреть виды признаков двухэтапного подхода, можно заметить, что для распознавания персон, лучше всех работает история предсказаний, т.е. учет классификации токена в предшествующей части текста. Это обусловлено тем, что статистику имен людей лучше считать внутри рассматриваемого документа, так меньше шансов перепутать похожие имена. Более того, как говорилось ранее, этот признак учитывает, что в начале документа обычно есть упоминание персоны в полной форме. Для организаций лучший результат показывает признак, соответствующий статистике внутри всей коллекции. Это можно объяснить тем, что на всей коллекции статистику организаций посчитать легче, так как внутри одного новостного документа одна и та же организация встречается не так часто, как персоны.

Отметим, что набор признаков (7) представляет собой лучший результат извлечения имен, полученный только на основе автоматически вычисляемых признаков, без составленных экспертами словарей.

Дополнение словарей (набор признаков (8)) к простому набору признаков (2) сразу дало качество лучше, чем все автоматические признаки. Дальнейшее добавление к набору признаков со словарями автоматических кластеров (набор признаков (10)) улучшает качество извлечения.

Наборы признаков (11-14) показывают результаты добавления к признакам (10) признаков второго этапа анализа отдельно и всех вместе. Качество еще немного возрастает, причем лучшим признаком второго этапа при условии применения словарей и кластеров является история предсказаний. При этом на коллекции *Person-1111* качество выросло более чем на 1%.

Полученные результаты можно сравнить с результатами работы [21], в которой описывается инженерно-лингвистическая система извлечения имен персон, основанная на словарях и правилах без использования машинного обучения. Тестирование в упомянутой работе проводилось на коллекциях *Person-1000* и *Person-1111*, при извлечении имен персон из которых получены результаты соответственно 96,96 и 73,71 *F*-меры.

Таблица 4

Результаты, полученные в коллекциях *Persons-1000* и *Persons-1111*

	Persons-1000				***Persons-1111***
Признаки	***PER***	***ORG***	***LOC***	***Micro***	***PER***
1) Признаки токена +Контекст	82,78	62,57	78,77	75,09	55,51
2) (1) + Bigram	91,56	74,40	88,39	86,75	75,6
3) (2) + word2vec	95,92	84,53	93,50	91,32	80,99
4) (3) + история предсказаний	96,14	84,79	93,77	91,82	83,30
5) (3) + статистика документа	95,92	84,52	93,62	91,62	82,76
6) (3) + статистика коллекции	96,00	84,62	93,83	91,74	82,07
7) (3) + все признаки второго этапа	96,28	84,75	93,71	91,86	83,34
8) (2) + словари	96,61	85,19	94,94	92,49	84,82
9) (8) + все признаки второго этапа	97,33	85,75	95,26	93,05	85,98
10) (2) + dictionaries + word2vec	97,10	87,27	95,43	93,48	86,54
11) (10) + история предсказаний	**97,62**	87,25	**95,60**	**93,73**	**87,71**
12) (10) + статистика документа	97,27	87,26	95,49	93,56	86,66
13) (10) + статистика коллекции	97,20	87,40	95,32	93,53	86,72
14) (10) + все признаки второго этапа	97,40	**87,42**	95,43	93,64	87,20

Из табл. 4 видно, что близкий результат извлечения имен персон из коллекции *Person-1000* (96,28 *F*-меры – набор признаков (7)) может быть получен только на основе различных признаков из коллекции, без использования ручных словарей. Полный набор предложенных признаков достигает несколько более высокой величины *F*-меры. Что касается коллекции *Person-1111*, то использование только признаков из коллекции дает уже существенно большую величину *F*-меры для извлечения имен персон – 83,34 (набор признаков (7)). Качество извлечения возрастает при добавлении словарных признаков, достигая 87,71 *F*-меры (набор признаков (11)). При этом применяемая нами модель для извлечения имен людей из коллекции *Person-1111* обучена на коллекции *Person-1000*. Таким образом, подход на основе машинного обучения показал более высокую способность по переносу на текстовую коллекцию с другими типами имен персон.

ЗАКЛЮЧЕНИЕ

В настоящей статье были рассмотрены разные виды признаков для извлечения именованных сущностей на основе метода машинного обучения: признаки токена, контекстные признаки, словарные признаки, признаки, полученные на основе кластеризации слов, а также признаки второго этапа: история предсказаний, статистика документа и статистика коллекции. Было проведено сравнение этих признаков для каждого типа именованных сущностей и показано, что их комбинация показывает высокое качество классификации. Также было произведено сравнение с системой, основанной на правилах, на открытой текстовой коллекции.

Из полученных результатов стоит отметить большой вклад контекстного биграммного признака. Из проведенных исследований можно заключить, что, используя признаки, которые не зависят от экспертных знаний, можно получить достаточно хорошую классификацию. При добавлении словарей ее качество улучшается.

Представленная система машинного обучения при извлечении восточных имен достигла значительно более высокого качества, чем инженерно-лингвистическая система, основанная на словарях и правилах.

СПИСОК ЛИТЕРАТУРЫ

1. Ермаков А.Е. Извлечение знаний из текста и их обработка: состояние и перспективы // Информационные технологии. – 2009. – №. 7. – С. 50−55.
2. Kuznetsov I.P., Kozerenko E.B., Kuznetsov K.I., Timonina N.O. Intelligent system for entities extraction (ISEE) from natural language texts. Proceedings of the International Workshop on Conceptual Structures for Extracting Natural Language Semantics-Sense, 2009. − № 9. – P. 17−25.
3. Khoroshevsky V.F. Ontology driven multilingual information extraction and intelligent analytics // Web Intelligence and Security. – 2010. − P. 237−262.
4. Lafferty J., McCallum A., Pereira F. Conditional random fields: Probabilistic: models for segmenting and labeling sequence data // Proceedings of the International Conference on Machine Learning ICML. – 2001.
5. Ratinov L., Roth D. Design challenges and misconceptions in named entity recognition // Proceedings of the 13th Conference on Computational Natural Language Learning CoNLL. – ACL, 2009. – P. 147–155.
6. Straková J., Straka M., Hajič J. A New state-of-the-art. Czech named entity recognizer // Proceedings of the 16th International Conference «Text, Speech, and Dialogue» TSD 2013. – Berlin-Heidelberg: Springer, 2013. – P. 68–75.
7. Tjon Kim Sang Erik, F., Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition // Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003. – ACL. – 2003. – № 4. – P. 142–147.
8. Brown P.F., Della Pietra V.J., Desouza P.V., Lai J.C., Mercer R.L. Class-based n-gram models of natural language // Computational Linguistics. – 1992. – №. 18 (4). – P. 467−479.
9. Tkachenko M., Simanovsky A. Named Entity Recognition: Exploring Features // Proceedings of the 11th Conference on Natural Language Processing KONVENS 2012. – Eigenverlag ÖGAI, 2012. – P. 118−127.
10. Marcin`czuk M., Stanek M., Piasecki M., Musiał A. Rich Set of Features for Proper Name Recognition in Polish Texts // Proceedings of the International Joint Conferences «Security and Intelligent Information Systems» SIIS 2011. − Springer Berlin Heidelberg. 2012. – P. 332–344.
11. Antonova A.Y., Soloviev A.N. Conditional random field models for the processing of Russian // Proceedings of the International Conference «Dialog 2013». − RGGU, 2013. – P. 27–44.
12. Подобряев А.В. Поиск упоминаний лиц в новостных текстах с использованием модели условных случайных полей // Труды всероссийской конференции «Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции» RCDL-2013 – ЯРГУ им. Демидова, 2013. – С. 255–258.
13. Gareev R., Tkachenko M., Solovyev V., Simanovsky A., Ivanov V. Introducing baselines for Russian named entity recognition // Proceedings of the 14th International Conference CICLing 2013. − Springer Berlin Heidelberg, 2013. – P. 329–342.
14. Chrupala G. Efficient induction of probabilistic word classes with LDA // Proceedings of the 5th International Joint Conference on Natural Language Processing IJCNLP 2011. − Asian Federation of Natural Language Processing, 2011. − P. 363−372.
15. Clark A. Combining distributional and morphological information for part of speech induction // Proceedings of the 10th Conference on European Chapter of the Association for Computational Lin-

guistics EACL 2003. – ACL, 2003. – № 1. − P. 59−66.
16. Bocharov V., Starostin A., Alexeeva S., Bodrova A., Chunchunkov A., Dzhumaev, S., Efimenko I, Granovsky D., Khoroshevsky V., Krylova I., Nikolaeva M., Smurov I., Toldova S. FactRuEval 2016: Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian// Proceedings of International Conference on Computational Linguistics Dialog-2016, 2016. – P. 702−720.
17. Власова Н.А. К проблеме разметки текстов на русском языке для задачи извлечения фактографической информации // Труды конференции TEL, 2014. − Fan, 2014. – С. 36–40.
18. Mozharova V., Loukachevitch N. Combining Knowledge and CRF-based Approach to Named Entity Recognition in Russian // Proceedings of the 5th International Conference on Analysis of Images, Social Networks, and Texts AIST'2016. − 2016.
19. Loukachevitch N., Dobrov B. RuThes linguistic ontology vs. Russian wordnets // Proceedings of the Global WordNet Conference GWC, 2014. − Tartu, 2014.
20. Mozharova V., Loukachevitch N. Two Stage Approach In Russian Named Entity Recognition // Proceedings of the International FRUCT conference on Intelligence, Social Media and Web ISMW, 2016. − 2016.
21. Трофимов И.В. Выявление упоминаний лиц в новостных текстах // Программная инженерия. − 2015. – №6.– С. 41−47.

Материал поступил в редакцию 07.12.16.

Сведения об авторах

МОЖАРОВА Валерия Александровна – студент факультета вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова.
e-mail: valerie.mozharova@gmail.com

ЛУКАШЕВИЧ Наталья Валентиновна – кандидат физико-математических наук, ведущий научный сотрудник НИВЦ Московского государственного университета имени им. М.В. Ломоносова
e-mail: louk_nat@mail.ru

УДК 81’322.2/.3’366

Л.Ю. Ковригина

Компьютерная морфология для исследований вариативного текста*

Описаны принципы и результаты разработки свободно распространяемой компьютерной морфологии для проведения морфологического анализа средневековых рукописных текстов на русском языке, а также принципы разработки компьютерной морфологии для корпуса вариативного текста и приведены примеры словарей и грамматик NooJ для обработки текстов с большим содержанием графических и грамматических вариантов.

Ключевые слова: *диахронический корпус вариативного текста, компьютерная морфология для обработки текстов на русском языке XVI-XVII вв., автоматический морфологический анализ рукописных текстов*

ВВЕДЕНИЕ

Актуальность разработки лингвистических *open source* модулей для морфологического анализа текстов на русском языке XVI-XVII вв. обусловлена тем, что сбор и предварительная обработка материала в большинстве проектов по исторической лексикологии и лексикографии русского языка производится вручную на основании картотек, на создание которых уходит не одно десятилетие, в то время как разработаны технологии корпусной лингвистики, которые могут эту задачу упростить и ускорить.

Корпус представляет собой массив текстов, снабженный разметкой (морфологической, синтаксической, семантической, прагматической и т.п.). Для поиска интересующих данных в корпусе используется специальное программное обеспечение – корпус-менеджер [1]. Поскольку задачи исследователя могут быть узкоспециальными, то необходимо выбрать такое программное обеспечение, которое позволяло бы и исследователю-неспециалисту в области прикладной лингвистики и автоматической обработки текста пополнять корпус и модули лингвистического анализа[1].

Для решения поставленных задач был выбран лингвопроцессор NooJ[2], в котором можно разрабатывать модули для создания автоматической морфологической (словари и морфологические грамматики), синтаксической (контекстно-свободные и контекстно-связанные формальные грамматики) и семантической (перифразирование, задание семантических свойств и ограничений лексем) разметки текста.

Автоматический морфологический анализ заключается в приписывании каждой словоформе входного текста правильной морфологической информации [3]. В корпусных исследованиях эта морфологическая информация необходима для поиска по заданным грамматическим характеристикам. Автоматический морфологический анализ рукописного текста до сих пор является не вполне решенной научной задачей. Это отчасти обусловлено разными форматами входных данных, трудоемкостью получения исходных данных, обилием графических вариантов, встречающихся в текстах для одной и той же словоформы. С проблемой обработки и «сведения» к одной словоформе или лексеме их разных графических вариантов сталкиваются все исследователи, занимающиеся составлением и разметкой диахронических корпусов [4–8]. При такой краеугольности проблемы вариативности в задачах автоматической обработки рукописного текста важно иметь репрезентативный корпус для установления многообразия графических и грамматических вариантов, а также пределов варьирования плана выражения одной и той же словоформы. Представляется, что идеальным эмпирическим материалом для такого рода исследований является корпус вариантов одного и того же текста, поскольку, несмотря на варьирование словника в разных редакциях и списках текста, отследить тенденции варьирования плана выражения словоформ и лексем полнее можно именно в корпусе вариативного текста. В то же время, ранее не ставившейся проблемой является создание корпуса одного текста, имеющего богатую литературную историю и представленного множеством своих вариантов.

Компьютерная морфология для исследования рукописного текста была разработана и протестирована на корпусе вариативного текста XVI-XVII вв. «Ска-

* Работа выполнена при поддержке Российского гуманитарного научного фонда, проект № 15-04-00521.

[1] Разработка именно компьютерной морфологии для анализа рукописного текста не является абсолютно новой задачей. Например, инструменты для автоматического морфологического анализа славянских и русских рукописей разработаны в проекте «Манускрипт» [2], однако они не являются открытыми и свободно пополняемыми, отсутствует возможность проводить морфологический анализ фрагментов рукописей, не входящих в число источников проекта.

[2] http://nooj4nlp.net/pages/nooj.html

зание о Мамаевом побоище», первоначально составленном для исследования темпоральных изменений вариантов текста статистическими методами (пояснения о термине «вариативный текст» и описание корпуса приводится далее).

КОРПУС ВАРИАТИВНОГО ТЕКСТА ДЛЯ РАЗРАБОТКИ И ТЕСТИРОВАНИЯ МОРФОЛОГИИ

Вариативный текст как предмет исследования

При всем разнообразии текстов [9], наиболее изучаемыми являются авторские художественные тексты нового и новейшего времени, тиражируемые с помощью традиционной гутенберговской полиграфии, обеспечивающей идентичность разных отпечатков одного тиража.

Ситуация же с устными и рукописными текстами, произведениями самиздата, современной городской (авторской) песней, текстами, циркулирующими в Интернете, и т.д. оказывается совсем иной. Для таких текстов не существует эталонного, окончательного, подлинного авторского варианта, который и должен быть предметом исследования с позиций лингвистики.

Прецеденты рассмотрения таких текстов были и ранее [10–12]. Однако, целенаправленно концентрируясь на изучении таких текстов, Ю.В. Доманский обосновывает представление о том, что они должны рассматриваться в качестве особого класса текстов – вариативных текстов, трактуемых им следующим образом: «категория варианта … важна уже потому, что в неклассической художественности произведение не реализуется только в каком-то одном варианте, а представляет из себя совокупность текстуальных (в широком смысле) манифестаций, каждая из которых обладает относительно самостоятельными смыслами. Это сближает словесность парадигмы неклассической художественности с фольклором и древней литературой» [13]. Такими текстами являются и сказки (см. на эту тему работу В.Я. Проппа [12]), и рукописные тексты Средневековья, и многие тексты замкнутых профессиональных и полупрофессиональных сообществ и т.д.

Текст, существующий в нескольких вариантах, каждый из которых является полноправным представителем данного текста, будет квалифицироваться как вариативный. Варьирование формы текста обеспечивается с помощью его сокращения, расширения, правки, внесения содержательных, стилистических, структурных, грамматических изменений, фонетических изменений, получивших отражение в рукописном тексте, интенсивности интертекстуальных связей и т.д. При этом *все такие варианты функционируют в культуре как единый текст*, без явного предпочтения одного из его вариантов.

В силу культурных предпосылок семантически полноценные письменные вариативные тексты значительного объема обнаруживаются только в средневековой литературе, но их существование затруднено в условиях профессиональной литературы Нового времени, выделяющей только один вариант текста в качестве «канонического». В Новое время культурная норма Средневековья сменяется представлением о допустимости единственной окончательной версии авторского текста, что сопровождается одновременным вытеснением коллективного авторства индивидуальным, поэтому закономерно сужаются диапазон и объем допустимых редакционных изменений. С появлением и распространением Интернета вновь актуализируется коллективное авторство и сосуществование разных вариантов текста, но уже в других формах и жанрах. В целом, можно заметить, что в постмодернизме возникает вопрос о принципиальной неразличимости авторского текста и плагиата.

«Сказание о Мамаевом побоище» как эталонный вариативный текст

Среди памятников русской литературы примером такого эталонного материала для проведения лингвостатистических (изучение изменения структурных характеристик вариативного текста) и лексикологических исследований (изучение границ варьирования лексических единиц) является «Сказание о Мамаевом побоище», поскольку: 1) текст переписывался и неоднократно редактировался на протяжении 400 лет вплоть до XIX в. и сохранился в большом количестве списков, 2) объем каждой редакции текста (в среднем) составляет 10 тыс. словоупотреблений, что вполне приемлемо для стандартной статистической обработки.

Этот литературный памятник сохранился более чем в 100 списках, причем со времени появления произведения – конец XIV – начало XV вв. – до появления самых поздних версий – XIX в. – было создано 8 редакций и обработок, не считая компилятивные версии.

При отборе списков для исследования использовались указанные Л.А.Дмитриевым палеографические данные [14]. Отбор производился по следующим критериям, которые можно разделить на 2 группы:

1. Требования к характеристикам отдельного текста:

а) список должен содержать типовой текст редакции.

2. Требования к совокупности выделенных текстов:

а) наличие в репертуаре списков контрастных вариантов редакций с целью выделения формальных различительных показателей текста: (1) полного и сокращенного, (2) полного и беллетризованного, (3) близкого к авторскому, и компилятивного;

б) среди исследуемых рукописей не должно быть поздних списков первых редакций и списков поздних редакций, которые датируются раньше, чем списки более ранних редакций (т.е. установленная Л.А. Дмитриевым очередность редакций должна быть изоморфна последовательности списков «Сказания…», упорядоченных по дате их написания). Это требование, которое нам удалось выдержать, обусловлено стремлением «синхронизировать» языковые и структурно-сюжетные изменения текста.

Материалом исследования послужили следующие 8 списков:

1) Основная редакция,
2) Летописная редакция,

3) Киприановская редакция,
4) Распространенная редакция,
5) редакция в составе Киевского Синопсиса,
6) сокращенный список Распространенной редакции,
7) компилятивный список Распространенной редакции,
8) беллетризованный список Основной редакции (см. таблицу).

Эти списки включают пять основных редакций (1-5) и списки, представляющие интерес для изучения статистических различий между: а) полным текстом редакции и текстом, представляющим ее последовательное сокращение (4 и 6); б) полным текстом и текстом, сокращенным с целью беллетризации (1 и 7); в) текстом, близким к авторскому, и компилятивным текстом (4 и 8), нумерация дана по табл. 1.

Корпус из восьми указанных списков «Сказания…» в машиночитаемой форме опубликован в открытом доступе[3]. Там же опубликованы лингвистические ресурсы (словари и грамматики NooJ) для морфологической разметки корпуса. Общий объем корпуса составляет 78 629 словоупотреблений.

Рукописный текст воспроизводился в соответствии с правилами, принятыми в проекте «СКАТ» [15], в рамках которого на кафедре математической лингвистики СПбГУ на протяжении нескольких десятилетий создается база русских агиографических текстов XV–XVII вв. Участники проекта разработали «оригинальный компьютерный шрифт для отображения особенностей древнерусской графики, позволяющий воспроизводить текст рукописи с достаточно высокой степенью приближения к оригиналу. Отображены графические начертания всех древнерусских букв и их семантически значимых вариантов (узкое и широкое «о»; узкое, широкое, якорное «е» и т. п.). Воспроизводятся титла, титловые покрытия, паерки, выносные буквы и буквосочетания, а также знаки придыхания и акцентные знаки» [16]. Методика подготовки корпуса «Сказания…» в машиночитаемой форме подробно описана в работе [17].

Списки «Сказания о Мамаевом побоище», привлеченные для исследования

№	Название редакции	Время возникновения редакции	Датировка и шифр списка	Особенности редакции
Типовые списки редакций в хронологическом порядке				
1	Основная (ОР)	не ранее 1-й четверти XV в.	XVI в., РНБ О.IV.22	Редакция, наиболее близкая к протографу
2	Летописная (ЛР)	конец XV – нач. XVI вв.	XVI в., СПбО-ИИ №251	Летописная повесть + ОР
3	Киприановская (КР)	1526-1530 гг.	XVI в., БАН 32.14.8	Структура Летописной повести + сокращенная ОР, возникла в церковной среде, прославление митрополита Киприана
4	Распространенная (РР)*	до начала XVII в.	XIX в., РНБ Q.IV.354	ОР + 2 самостоятельные повести «О посольстве Захария», «О новгородцах»
5	Редакция в составе Киевского Синопсиса (редакция Синопсиса)*	ок. 1680 г.	конец XVIII в., РНБ Собр. Колобова, №336	Последовательное сокращение ОР, местами вставки
Нетиповые и компилятивные списки				
6	Сокращенный список Распространенной редакции*	не установлено	XVII в., РНБ Q.XVII.70	Сквозное сокращение текста РР
7	Компилятивный список Распространенной редакции*	не установлено	XIX в., РНБ О.IV.46	Компилят Распространенной редакции, в котором прочитывается 11 других источников
8	Беллетризованный список Основной редакции	не установлено	XIX в., РНБ Собр. Михайловского, № Q.509	ОР, опущены молитвы + поздние вставки

Примечание: Списки, отмеченные *, ранее не публиковались и не переводились в машиночитаемую форму.

[3] https://github.com/kamivao/oldrus_morf/

ОПИСАНИЕ КОМПЬЮТЕРНОЙ МОРФОЛОГИИ

Формальные модели компьютерной морфологии в лингвопроцессоре NooJ

NooJ – это свободно распространяемое (лицензия GNU Affero GPL) программное обеспечение для обработки естественного языка[4], которое эффективно применяется для составления морфологических, синтаксических, семантических формализованных описаний языковых единиц и использовать в качестве корпус-менеджера, поддерживающего, в том числе, и язык регулярных выражений, для извлечения данных с учетом имеющейся разметки. NooJ допускает введение любого неограниченного количества пользовательских обозначений для лингвистических категорий и неограниченный набор других меток. Следует отметить, что в NooJ нет встроенных статистических алгоритмов, алгоритмов машинного обучения, средств интеграции с онтологиями и т.д., как, например, в средах автоматической обработки естественного языка GATE[5] и NLTK[6], но технологий NooJ вполне достаточно для проведения работ по исторической лексикологии и лексикографии и аннотированию диахронических корпусов.

В Nooj используется несколько структур данных: словари лексем (с расширением .nod), словари с описанием словоизменительных и словообразовательных характеристик (с расширением .nof), грамматики для описания морфологических (с расширением .nom) свойств лексем, грамматики для описания синтаксических конструкций (с расширением .nog). Допустимые грамматические значения для классов слов определяются в отдельном файле (_properties.def). В начале каждого словаря .nod указывается, какие именно словари и грамматики и какой файл _properties.def использовать для морфологического анализа.

Основные формальные модели, используемые в NooJ для представления лингвистических ресурсов и проведения морфологического, синтаксического и семантического анализа, – это конечные автоматы, конечные преобразователи и (расширенные) рекурсивные сети перехода [18, 19].

Конечные преобразователи являются одной из самых популярных математических моделей для морфологической и синтаксической разметки [20–23]. Морфологический конечный преобразователь распознает цепочку букв, образующих словоформу, и порождает для этой словоформы морфологическую информацию (например, ее лексему, грамматические характеристики, и т.д.).

В лексических словарях NooJ для каждой входной лексемы может быть указана, помимо модели словоизменения и словообразования, любая семантическая информация. Эти метаданные могут быть в дальнейшем использованы для конвертации размеченного корпуса в семантический формат (например, в открытые связные данные), что, в свою очередь, создает возможность их применения в сторонних базах данных. Словари NooJ легко пополнимы, и любой исследователь может добавить собственные лингвистические ресурсы к существующему словарю или создать свой словарь. После компиляции словари NooJ можно использовать в качестве корпус-менеджеров и писать поисковые запросы с использованием регулярных выражений. Входные единицы лексических словарей NooJ имеют следующий формат:

«Лексическая_единица+
+Словоизменительная_модель+
+Словообразовательная_модель+
Семантические_характеристики».

Таким образом, за минимальный срок с момента начала работы исследователь, не прибегая к посторонней помощи, может создать свой корпус и лингвистическое обеспечение к этому корпусу. Все это крайне актуально в силу отсутствия единых стандартов на представление рукописного текста в машиночитаемом виде и разработки лингвистических инструментов для его автоматического анализа.

В NooJ имеются функции, которые представляются перспективными для обработки вариативного текста.

При создании формализованных описаний можно объединить подобные варианты в суперлексему. Единицы одной суперлексемы образуют класс эквивалентности, что позволяет производить поиск в корпусе по любой из этих единиц. В выдачу по поисковому предписанию, содержащему суперлексему, попадут все единицы, «привязанные» к одной суперлексеме, с учетом приписанных им словоизменительных характеристик. Суперлексемы значительно облегчали работу с вариантами и активно использовались при составлении словаря .nod, в особенности для объединения вариантов написания собственных имен (см. пример далее). Единицы одной суперлексемы образуют класс эквивалентности, что позволяет производить поиск в корпусе по любой из этих единиц. В выдачу по поисковому предписанию, содержащему суперлексему, попадут все единицы, «пристегнутые» к одной суперлексеме, с учетом приписанных им словоизменительных характеристик. Приведен пример объединения вариантов в суперлексему:

евдокия, N (суперлексема)
evдокия, евдокия, N (графический вариант + суперлексема)
евдокэя, евдокия, N
евдокея, евдокия, N
еовдокэа, евдокия, N
еовдокия, евдокия, N
овдотя, евдокия, N

Написание словоизменительных парадигм с помощью набора встроенных операторов Nooj также не является затруднительным [18, с. 94]. Например, оператор <B> (Backspace) удаляет символы с конца лексемы, количество удаляемых символов указывается внутри скобок при этом операторе: <B2>, <B5>. После закрывающей скобки без пробела указывается последовательность символов, добавляемая к урезанной лексеме до получения определенной словоформы. Через слеш (/) записывается морфологическая

[4] www.nooj-association.org
[5] https://gate.ac.uk/
[6] http://www.nltk.org/

информация, соответствующая полученной словоформе. Как видно, в Nooj не применяется представление о псевдоосновe, сопряженной с набором допустимых аффиксов и словоизменительным типом, хотя в морфологических грамматиках NooJ можно задавать морфемы. На названия грамматических классов, категорий и значений программа не накладывает никаких ограничений, они определяются пользователем.

Обозначения граммем и лексико-грамматических категорий в компьютерной морфологии

В разработанной компьютерной морфологии применяется следующий набор тегов для аннотирования лексико-грамматических категорий:

ADJ – прилагательное;
ADV – наречие;
APRO – местоименное прилагательное;
CONJ – союз;
DPRO – указательное местоимение;
INTJ – междометие;
N – существительное;
NUM – числительное;
PART - частица.
PPRO – притяжательное местоимение;
PREP – предлог;
RPRO – возвратное местоимение;
SPRO – личное местоимение;
V – глагол.

Имена собственные размечены в корпусе как существительные с пометой “N+prop”.

Приведем теги для обозначения грамматических категорий и соответствующих граммем и их группировку в файле _properties.def:

SPRO_Case = nom | gen | dat | acc | ins | loc | voc ; # обозначение граммем падежа
SPRO_Pers = 1p | 2p | 3p; # обозначение граммем лица
SPRO_Nb = sg | pl | dual; # обозначение граммем числа
SPRO_Gender = m | f | n; # обозначение граммем рода

APRO_Case = nom | gen | dat | acc | ins | loc ; # обозначение граммем падежа
APRO_Anim = An | Inan ; # обозначение категории одушевленности
APRO_Nb = sg | pl | dual; # обозначение граммем числа
APRO_Gender = m | f | n; # обозначение граммем рода

V_Repres = inf | partcp | l_partcp | sup | fin ; # обозначение форм глагола
V_Tense = praes | praet | fut | aor | imperf_f | imperf_c | perf |; # обозначение граммем категории времени
V_Pers = 1p | 2p | 3p; # обозначение граммем лица
V_Nb = sg | pl | dual; # обозначение граммем числа
V_Gender = m | f | n; # обозначение граммем рода
V_Voice = act | pass | refl; # обозначение граммем залога

Для причастий (partcp) дополнительно указываются следующие граммемы:

V_Type = long | short ; # обозначение формы причастий
V_Case = nom | gen | dat | acc | ins | loc | voc ; # обозначение граммем падежа

N_Gender = m | f | n; # обозначение граммем рода
N_Nb = sg | pl | dual; # обозначение граммем числа
N_Case = nom | gen | dat | acc | ins | loc | voc; # обозначение граммем падежа
N_Distribution = abst | anim | animcoll | coll | conc | prop | propcoll | inanimate | inanimatecoll; # обозначение лексико-грамматических разрядов имен существительных

ADJ_Gender = m | f | n; # обозначение граммем рода
ADJ_Nb = sg | pl | dual; # обозначение граммем числа
ADJ_Case = nom | gen | dat | acc | ins | loc | voc;
ADJ_Distribution = Qual | Quan | Rel; # обозначение лексико-грамматических разрядов имен прилагательных
ADJ_Type = long | short; # обозначение формы имен прилагательных
ADJ_Compare = Pos | Comp | Sup; # обозначение степени имен прилагательных

NUM_Type = Card | Ord | Coll ; # обозначение лексико-грамматических разрядов числительных
NUM _Gender = m | f | n; # обозначение граммем рода
NUM _Nb = sg | pl | dual ; # обозначение граммем числа
NUM _Case = nom | gen | dat | acc | ins | loc | voc ; # обозначение граммем падежа

Текстовое описание словоизменительных моделей было сделано для следующих лексико-грамматических категорий:

1) существительные: описываются основные типы склонения и подтипы (например, если флексии предшествует «к», «г», «х»), указываются падеж (nom|gen|dat|acc|ins|loc|voc) и число (sg|pl|dual), род существительного указывается в словаре .nod, так как к одному словообразовательному типу могли относиться существительные разных родов;

2) прилагательные: описываются краткие и местоименные формы прилагательных для основ, заканчивающихся на твердую и мягкую согласную, указываются падеж (nom|gen|dat|acc|ins|loc|voc), род (m|f|n) и число (sg|pl|dual);

3) местоимения: описываются личные, указательные, притяжательные, вопросительные, неопределенные, относительные, отрицательные, определительные местоимения, указываются падеж (nom|gen|dat|acc|ins|loc|voc), род (m|f|n) для неличных местоимений, лицо (1p|2p|3p) для личных местоимений и число (sg|pl|dual);

4) «нетематические» глаголы настоящего времени (быти, дати, ведати, имети, ясти): указываются время (praes), лицо (1p|2p|3p) и число (sg|pl|dual);

5) числительные.

Словоизменительные модели можно задавать в виде морфологической грамматики (.nom). Такой способ является более наглядным и, кроме того, позволяет строить гипотезы о морфологических характеристиках лексем, не входящих в словарь, и генерировать для них леммы, впрочем, в некоторых случаях они могут генерироваться неправильно, и нужно задавать ограничивающие условия, что значительно сужает множество словоформ, которые могли бы быть проанализированы с помощью построенной грамматики. Однако именно морфологические грамматики обеспечивают быстрое и компактное описание вариантов словоформ.

Примеры грамматик для анализа форм имперфекта основ, заканчивающихся на гласный «а», и форм сигматического аориста приведены на рис. 1 и 2.

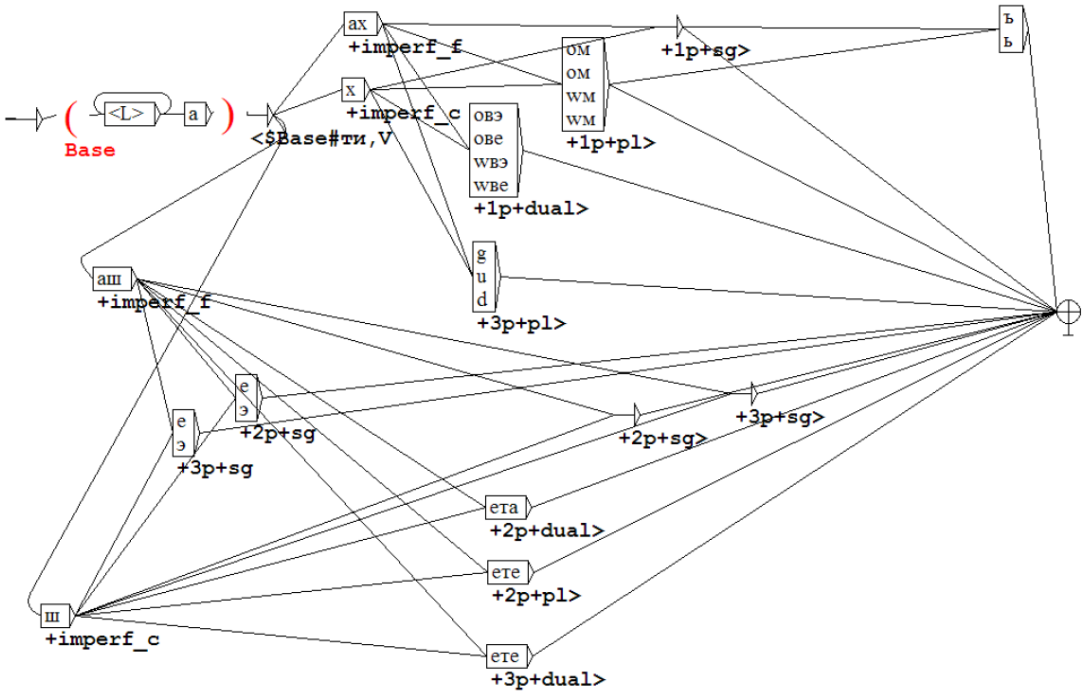

Рис.1. Морфологическая грамматика NooJ для анализа форм имперфекта основ, заканчивающихся на гласную «а»

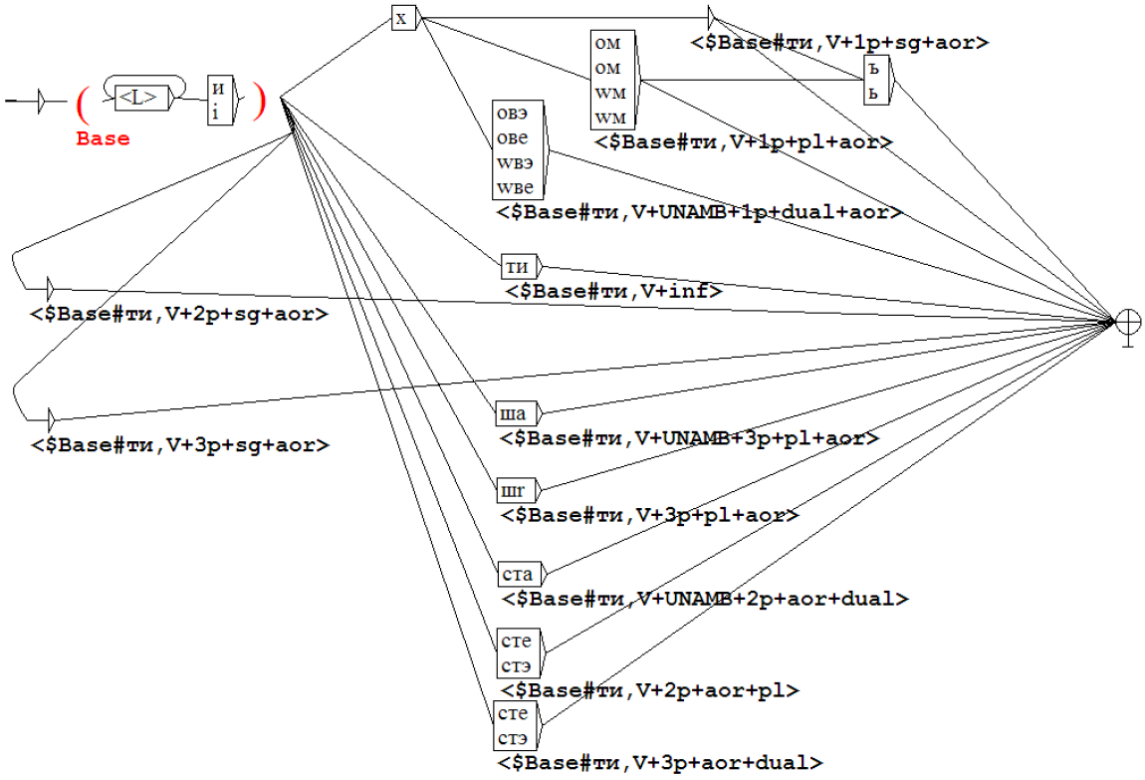

Рис. 2. Морфологическая грамматика NooJ, обрабатывающая формы сигматического аориста основ на «и»

Морфологическая грамматика представляет собой конечный преобразователь, в узлах записаны распознаваемые символы, а после «+» указана морфологическая информация, соответствующая распознанной цепочке.

В грамматике дополнительно указывается, является ли проанализированное слово стяженной (imperf_c) или нестяженной формой имперфекта (imperf_f).

Приведем результаты анализа словоформ «**вызнаахомъ**», «**восплакахоу**», «**бесѣдшваховѣ**» с помощью указанной грамматики:

LU LEMMA="вызнаати" CAT=
"V" Tense="imperf_f" Pers=
"1p" Nb="pl">вызнаахомъ</LU></LU>

<LU LEMMA="восплакати" CAT="V" Tense=
"imperf_c" Pers="3p" Nb="pl">восплакаху</LU>

<LU LEMMA="бесэдwвати" CAT=
"V" Tense="imperf_c" Pers="1p" Nb=
"dual">бесэдwваховэ</LU>

Словоизменительные модели описаны нами с помощью морфологических грамматик для:

1) глаголов (настоящее время (praes), аорист (aor), имперфект (imperf_f, imperf_c), будущее время (fut), выраженное формами настоящего времени, повелительное наклонение (imper), указываются время, лицо и число);

2) причастий (указываются, действительное это или страдательное причастие (act|pass), время (praes|praet), род, число, падеж);

3) числительных, записанных буквами;

4) отдельных лексем, написанных под титлом или имеющих большое количество вариантов в основе.

Варианты лексемы, которая пишется под титлом, помечаются тегом «+titlo» в грамматике, и этот тег можно использовать в поисковых запросах к корпусу (например, запрос <V+titlo> вернет все глаголы, написанные под титлом). В NooJ имеются хорошо разработанные механизмы поиска по регулярным выражениям и по метаданным (разметке).

Грамматика, анализирующая написания под титлом лексемы «глаголати», встречающиеся в корпусе «Сказания…», приведена на рис. 3.

На выходе грамматика порождает такие аннотации к словоформам под титлом (сама словоформа находится перед закрывающим тегом </LU>):

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" Repres="inf" TYPE="titlo">ГЛАТИ</LU>

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" Repres="inf" TYPE="titlo">ГЛГОЛАТИ</LU>

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" Pers="3p" Nb="sg" TYPE="titlo">ГЛЕТ</LU>

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" Pers="3p" Nb="sg" TYPE="titlo">ГЛЕТЪ</LU>

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" Pers="3p" Nb="sg" TYPE="titlo">ГЛЕТЬ</LU>

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" Pers="2p" Nb="sg" TYPE="titlo">ГЛЕШИ</LU>

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" TYPE="imper" TYPE="titlo">ГЛИ</LU>

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" Pers="3p" Nb="sg" TYPE="titlo">ГЛЛЕТЪ</LU>

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" Repres="inf" TYPE="titlo">ГЛТИ</LU>

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" Pers="1p" Nb="sg" TYPE="titlo">ГЛЮ</LU>

<LU LEMMA="ГЛАГОЛАТИ" CAT=
"V" Pers="3p" Nb="pl" TYPE="titlo">ГЛЮТЬ</LU>

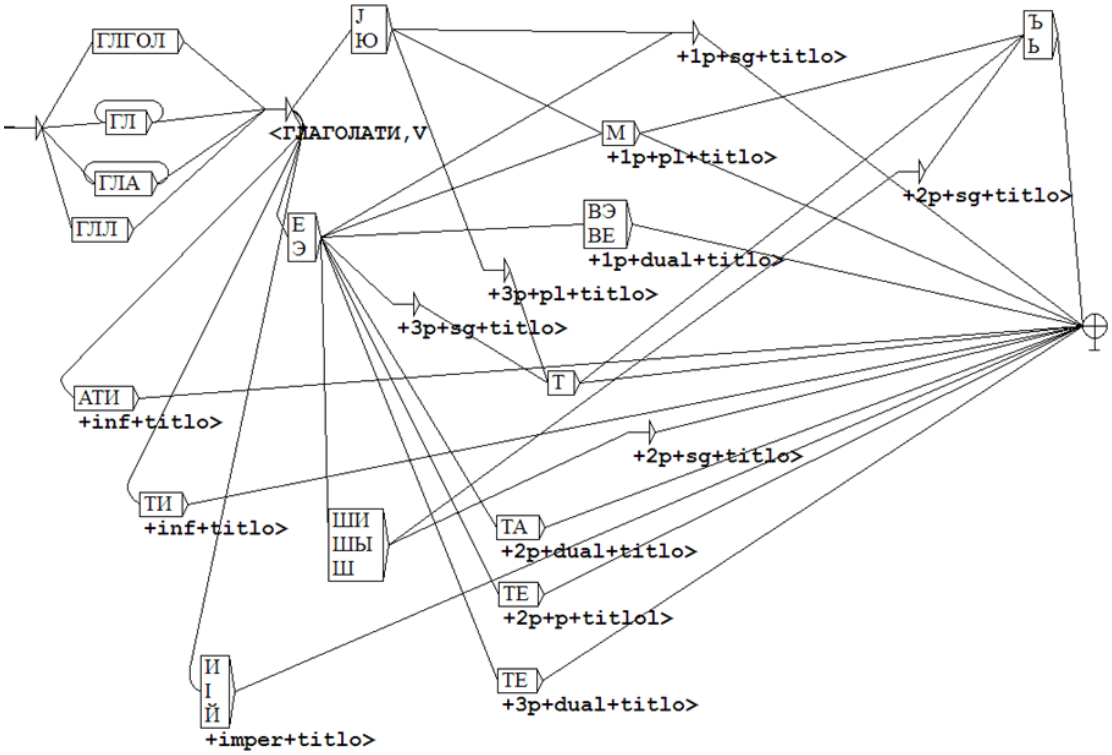

Рис. 3. Морфологическая грамматика NooJ, обрабатывающая написания лексемы «глаголати» под титлом

При использовании морфологических грамматик для анализа текста можно устанавливать очередность их применения, что позволяет избежать неправильной генерации лексем при срабатывании более общей грамматики.

Для разметки форм будущего сложного времени и перфекта мы использовали синтаксические грамматики NooJ.

При разработке словоизменительных моделей к множеству флексий были добавлены их варианты, возникшие в результате мены еров, юсов, «ять» и «е», пропуска еров на конце строки и в выносных буквах, фонетических изменений (например, -аго > -ого > -ово > -ова> в родительном падеже единственного числа мужского рода у прилагательных), унификации типов склонения. В результате, например, при порождении только части парадигмы прилагательного (кратких и местоименных форм) с основой на твердый согласный генерируется в общей сложности 290 словоформ.

Для описания парадигм использовались монографии В.В. Иванова, В.В. Колесова и А.М.Селищева [24-26]. Файлы компьютерной морфологии опубликованы в открытом доступе[7].

СПИСОК ЛИТЕРАТУРЫ

1. Захаров В.П. Корпусная лингвистика. – СПб : Изд-во Санкт-Петербургского гос. ун-та, 2005.
2. Манускрипт: славянское письменное наследие. – URL: http://manuscripts.ru/ (дата обращения: 01.12.2016).
3. Коваль С. А. Лингвистические проблемы компьютерной морфологии. – СПб. : Изд-во Санкт-Петербургского гос. ун-та, 2005. – 150 с.
4. Алексеева Е.Л., Лаврентьев А.М., Азарова И.В., Захарова Л.А. Разметка корпуса древнерусских текстов // Труды международной конференции «Корпусная лингвистика 2004». – 11-14 октября 2004 г. – СПб., 2004. – С. 16-24.
5. Amoia M. Martinez J.M. Using Comparable Collections of Historical Texts for Building a Diachronic Dictionary for Spelling Normalization // Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@ACL 2013, August 8, 2013. – Sofia, Bulgaria, 2013. – P. 84—89.
6. van Dalen-Oskam K.H. Authors, scribes, and scholars: Detecting scribal variation and editorial intervention via authorship attribution methods // Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches / eds. T. L Andrews, C. Macé. – Turnhout: Brepols, 2014.
7. van Dalen-Oskam K.H. In praise of the variant analysis tool. A computational approach to Medieval literature // Texts, Transmissions, Receptions: modern approaches to narratives / eds. André Lardinois, Sophie Levie, Hans Hoeken, Christoph Lüthy. – Leiden: Brill, 2015. – P. 35-54.
8. Zampieri M., Malmasi Sh., Dras M. Modeling Language Change in Historical Corpora: The Case of Portuguese // Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC. — 2016, Portorož, Slovenia, May 23-28, 2016.
9. Филиппов К.А. Лингвистика текста : Курс лекций. – СПб, 2003. – 333 с.
10. Лихачев Д.С. Текстология. На материале русской литературы X-XVII в. / Д. С. Лихачев при участии А. А. Алексеева и А. Г. Боброва. – М., 2001. – 758 с.
11. Корона В.В. Поэзия Анны Ахматовой: поэтика автовариаций. Екатеринбург, 1999. – 263 с.
12. Пропп В.Я. Морфология волшебной сказки. – М., 2003. – 143 с.
13. Доманский Ю.В. Вариативность и интерпретация текста (парадигма неклассической художественности): автореф. дис. … д-ра филол. наук. – М., 2006. – 43 с.
14. Дмитриев Л.А. Сказания и повести о Куликовской битве. – Л., 1982. – 422 с.
15. СКАТ : Санкт-Петербургский корпус агиографических текстов. – URL: http://project.phil.spbu.ru/scat/page.php?page=project (дата обращения 20.12.2016).
16. Герд А.С., Алексеева Е.Л., Азарова И.В., Захарова Л.А. Электронный корпус текстов по памятникам древнерусской агиографической литературы // Научно-техническая информация. Сер. 2. – 2004. – №. 9. – С.16-20.
17. Ковригина Л.Ю. Негауссовое моделирование лексико-статистической структуры вариативного текста (на примере «Сказания о Мамаевом побоище»)»: дис. … канд. филол. наук. – СПб., 2015. – 272 с. – URL: ttp://spbu.ru/ disser2/ 246/disser/covrigina_dis.pdf.
18. Silberztein M. NooJ Manual. 2003.– URL: www.nooj4nlp.net.
19. Silberztein M. Formalizing Natural Languages: the NooJ Approach. – Wiley, – 2016. – 346 p.
20. Roche E., Schabes Y. Deterministic Part-of-Speech Tagging with Finite-State Transducers // Computational Linguistics. – 1995. – Vol.21, №2. – P. 227-253.
21. Linden K., Axelson E., Drobac S., Hardwick S., Silfverberg S., Pirinen T.A. HFST – Framework for Compiling and Applying Morphologies // Systems and Frameworks for Computational Morphology – Second International Workshop, SFCM 2011, August 26. – Zurich, Switzerland, 2011. – P. 67-85.
22. Linden K., Axelson E., Drobac S., Hardwick S., Silfverberg S., Pirinen T.A. Using HFST for Creating Computational Linguistic Applications // Computational Linguistics Applications. – 2013. – P. 3- 25.
23. Луканин А.В. Автоматическая обработка естественного языка . – Челябинск: Изд. центр ЮУрГУ, 2011. – 70 с.

[7] https://github.com/kamivao/oldrus_morf/

24. Иванов В.В. Историческая грамматика русского языка : учеб. для филол. спец. ун-тов и пед. ин-тов; 2-е изд., испр. и доп. – М : Просвещение, 1983. – 399 с.
25. Колесов В.В. История русского языка: учеб.пособие. – СПб. : СПбГУ, 2005. – 672 с.
26. Селищев А.М. Старославянский язык : в 2 ч. - 2-е изд. - М : Эдиториал УРСС, 2001. – 205 с.

Материал поступил в редакцию 14.02.17.

Сведения об авторе

КОВРИГИНА Любовь Юрьевна – кандидат филологических наук, доцент кафедры информатики и прикладной математики Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики
e-mail: lyukovriguina@corp.ifmo.ru

УДК 81’276’322

Е.П. Буторина, Е.О. Губанова

Выявление употребительности коллокаций в деловых текстах

Рассматривается алгоритм выявления неоднословных сочетаний и производных предлогов, характерных для русской деловой и официальной речи. Построение такого алгоритма можно рассматривать как необходимый этап изучения деловой и официальной речи. Описываемый механизм может быть полезен при составлении учебников, а также обучающих и контролирующих сервисов как для иностранцев, так и для носителей русского языка. На основе ядра словосочетаний, уже внесенных в существующие учебники, создается словник в виде файла Excel. В качестве текстов, по которым производится поиск, использованы приказы органов государственной власти, опубликованные в открытом доступе в Интернете.

Ключевые слова: *неоднословные единицы, деловая речь, официальная коммуникация, русский язык, словник, база данных*

ВВЕДЕНИЕ

Письменные деловые тексты составляются большинством людей постоянно. Нередко возникают вопросы о правильности формулировок и точности употребления терминологических сочетаний, юридических конструкций и этикетных формул. Но в настоящее время для текстов на русском языке не всегда можно опираться на единое общепризнанное представление о том, какие словосочетания входят в число таких формулировок. Поэтому важно найти способ, который позволит объективно оценить, какие из них действительно являются часто встречающимися и употребительными в актуальных текстах.

Цель настоящей статьи – описать создание программы, выявляющей наиболее характерные для современной русской деловой и официальной речи неоднословные единицы. Работа имеет прикладную направленность: актуальность исследуемых в ней проблем связана, прежде всего, с необходимостью выявления наиболее часто встречающихся в деловой и официальной речи на русском языке неоднословных единиц и их систематизации для лингводидактических и лексикографических целей.

Деловое взаимодействие часто относится к сфере, регулируемой правом, поэтому оно нередко носит официальный характер, т.е. в нём используются средства официальной коммуникации. Для обозначения множества таких средств обычно употребляется термин «официально-деловой стиль» [1], включающий признаки принадлежности его единиц и к деловой, и к официальной сферам коммуникации. Пересечение деловой и официальной сфер коммуникации не является обязательным, так как существует неофициальное деловое общение (например, различные виды частных услуг), а официальная коммуникация не всегда носит деловой характер (например, законодательное регулирование семейных отношений).

Официально-деловой стиль используется в таких сферах, как государственное управление, законодательство и судопроизводство, административная деятельность, официальное делопроизводство, образование, массовая информация, наука и др.

Неоднословные единицы, характерные для русской официальной и деловой речи, необходимы для правильного составления различных документов, для ведения переписки и т.п. Это актуально не только для граждан России, например, для специалистов (особенно в нестандартных, не предусмотренных регламентом ситуациях) или для русскоязычных учащихся, которым нужно освоить принципы ведения деловой переписки на русском языке. Эти знания необходимы и тем, кто изучает русский язык как неродной или иностранный. При этом носителям языка все же легче справиться с усвоением соответствующих единиц и текстов, используя справочную литературу, свой практический опыт или даже интуитивное представление об особенностях деловой и официальной речи. Сложнее освоить такие единицы иностранцам, изучающим русский язык. Для этой цели существуют учебные пособия (например, [2]), но подобные пособия часто выпускаются небольшими тиражами, и их не всегда легко найти. Кроме того, более серьёзная проблема видится в том, что единого общепризнанного представления о том, какие слова и словосочетания должны изучаться как типовые неоднословные единицы официальной речи, еще не существует, так как они пока не описаны и не кодифицированы. Как правило, каждый автор составляет учебник, опираясь на собственный опыт и личные представления. Поэтому очень важно найти способ, который позволит объективно оценить, какие из этих словосочетаний являются действительно часто встречающимися и употребительными в современных текстах.

Для получения статистических данных о текстах нередко используются языковые корпусы. Но в данном случае нет возможности опереться на материал корпусов, поскольку деловая и официальная речь в Национальном корпусе русского языка (НКРЯ) [3] представлена небольшим количеством в значительной степени устаревших документов. Немалая их часть относится еще к XIX в., и основная масса внесенных в НКРЯ документов была составлена не позднее начала 2000-х годов.. В Корпусе Университета Лидс [4] содержится массив деловых, медиа- и интернет-текстов на русском языке, но отсутствует отдельный подкорпус официальных документов.

Поэтому наиболее достоверным представляется анализ документов, размещаемых на официальных сайтах в больших количествах в актуальные сроки и доступных для скачивания и последующей обработки. Если документы выкладываются в виде нередактируемых pdf-файлов, то можно конвертировать эти файлы в удобный для работы формат (обычно doc или docx) с помощью программы ABBYY Fine Reader или бесплатных онлайн-сервисов с аналогичными возможностями.

Для статистического исследования документов, которые находятся в общем доступе, можно применять уже существующие программы, однако проблема заключается в том, что эти программы, во-первых, не являются полностью бесплатными (имеют ограниченный срок бесплатного пользования), во-вторых, требуют ручного ввода текста.

В качестве альтернативы возможно получение статистических данных с помощью Microsoft Word и Microsoft Excel. Их удобство в том, что они имеются практически на каждом компьютере и знакомы большинству пользователей. Пакет Excel подходит для обработки больших объемов информации, позволяя автоматизировать математические вычисления, при этом можно пользоваться уже встроенными формулами или самостоятельно написать необходимые для конкретных вычислений. Также Excel позволяет составить программу на языке VBA (Visual Basic for Application), что существенно расширяет диапазон предусмотренных разработчиками операций, в том числе и в работе с текстами. В частности, становится возможным автоматически импортировать и обрабатывать тексты из файлов Word.

Теоретическая значимость нашего исследования заключается в углублении и систематизации представлений о функционировании русского языка в современной деловой и официальной речи.

ОПИСАНИЕ ПРОГРАММЫ

Для создания программы, выявляющей наиболее характерные для современной русской деловой и официальной речи неоднословные единицы, были поставлены и решены следующие задачи:

1) сформировать предварительный список неоднословных единиц, опираясь на множество сочетаний, уже приведенных в существующих учебниках и справочниках по деловой речи;

2) оформить этот список в виде одного или нескольких раскрывающихся списков в Excel, что позволит быстро выбирать неоднословные единицы для статистического анализа;

3) найти оптимальный способ поиска заданных неоднословных единиц в сформированной базе текстов.

В списке неоднословных единиц могут быть указаны:

- вербономинанты [5], т.е. сочетания глагола с ослабленным семантическим значением и существительного, например: *достигнуть договоренности, обеспечить сохранность*;
- сочетания глагола с существительным, носящие характер несвободной коллокации, когда нельзя заменить один глагол другим (например, *занять должность*);
- атрибутивные сочетания, которые могут представлять собой как свободные словосочетания, так и фразеологизированные [6]: *денежные средства, платежное поручение* и др.

В число подобных единиц могут быть включены производные отыменные предлоги, характерность которых для официально-деловой речи подчёркивается, например, в работе [7].

Чтобы иметь возможность рассчитывать значения статистических мер, был создан файл Excel, содержащий словник, состоящий из трех таблиц. В каждую таблицу были занесены слова, образующие неоднословную единицу, следующим образом: в заголовке таблицы проставлены главные слова словосочетаний (или непроизводные предлоги в составе производных), в столбце под заголовком записаны зависимые компоненты.

Каждая из таблиц размещена на отдельном рабочем листе[1] книги (файла) Excel «Статистика_Словник.xls». Нами разработаны рабочие листы:

- «Основной», в который заносится бóльшая часть неоднословных единиц, описываемых авторами существующих пособий как стандартные. Например, *контроль выполнения*, *предоставление полномочий*. В основном слова заносятся в лист в их начальной форме. В некоторых случаях имеет смысл оставить слова в той форме, в которой они употребляются в неоднословной единице (например, компонент «необходимым» в косвенном падеже в словосочетании *считать необходимым*);
- «Этикет», в который по тому же принципу записываются этикетные формулы (например, *просим разрешить*);
- «Предлоги», в который заносятся производные отыменные предлоги (например, *при условии*), а также несколько рабочих листов, куда выводится информация, необходимая для проверки и отладки программы, например «Лист2» и «Лист4».

Кроме того, создан рабочий лист «Лист1», на котором сформированы выпадающие списки[2] для трех

[1] Рабочий лист – это элемент рабочей книги (файла) Excel, предназначенный для ввода, хранения информации и выполнения вычислений. Рабочий лист является электронной таблицей, состоящей из ячеек.

[2] Выпадающий список – элемент графического интерфейса пользователя для выбора одного из нескольких заранее определенных значений. В Excel – это список из множества пунктов, встроенный в одну ячейку.

перечисленных таблиц Excel (рабочие листы «Основной», «Этикет» и «Предлоги»). Это позволяет выбрать неоднословную единицу для работы с ней. Сначала нужно определить главное слово, тогда во втором списке для выбора появляются только зависимые слова, относящиеся к выделенному главному слову словосочетания. Подставить вручную данные в выпадающих списках на рабочем листе «Лист1» уже нельзя. Но в таблицы-словники можно дописывать как новые главные слова (тогда к ней прибавится новый столбец), так и новые зависимые слова. Диапазоны будут автоматически расширяться и отображаться в выпадающих списках на рабочем листе «Лист1».

Мы рассматриваем следующие виды неоднословных единиц.

1. Этикетные формулы в соответствии с дефиницией Н.И. Формановской [8]. Автор определяет их как стереотипные, устойчивые сочетания, которые используются для выражения речевых интенций приветствия, благодарности и т.п.

2. Сочетания слов для «основного» словника

2.1. Сочетания модальных слов со значением волеизъявления с инфинитивами (например, *необходимо направить*)

2.2. Сочетания слов с глаголами в форме настоящего времени со значением долженствования (например, *осуществляют контроль*)

2.3. Сочетания слов, близкие по значению к юридическим или экономическим терминам или совпадающие с ними (например, *доля рынка*).

3. Производные отыменные предлоги, включающие непроизводные предлоги (например, *в, на, к*) и слова, соотносимые с отдельными частями речи, в основном с именами существительными (например, *течение, продолжение*). Эти предлоги также заносятся в отдельную таблицу. Фрагмент этой таблицы:

во	**за**
благо	далеко
времена	годы
взаимосвязи	вслед
вкусе	бортом
избежание	время
изменение	вычетом

Как источник неоднословных единиц, характерных для деловой и официальной коммуникации, мы использовали словник, в который включены неоднословные единицы, взятые из справочников по русской деловой речи, таких как, например, [2]. При этом не учитывались обозначения дат и времени, персональные имена, географические названия, названия органов власти и наименования ее представителей.

Потребность в составлении подобных таблиц заключается в следующем:

- выбор производится из заданных неоднословных единиц, что делает более удобной работу пользователя. Необходимость вносить неоднословную единицу заново появляется, только если она ещё не внесена в одну из таблиц;
- группировка неоднословных единиц по трем таблицам Excel («Основной», «Этикет» и «Предлоги») и, соответственно, по трем спискам облегчает поиск нужной неоднословной единицы в зависимости от предмета поиска: этикетная формула, производный предлог и т.д.;
- таблицы являются основой для работы программы и позволяют соблюдать единообразие запросов при поиске в разных массивах документов;
- для нерусскоязычного пользователя возможность выбирать неоднословную единицу из предложенного числа вариантов может сократить время на формулировку запроса и помочь избежать ошибок;
- в таблицы заносится информация из справочников и учебников по официально-деловой речи, поэтому нет необходимости каждый раз к ним обращаться. Если требуется только проверка встречаемости, а не какая-то иная информация, таблиц для этого будет достаточно;
- сохраняется возможность править таблицы, добавляя в них неоднословные единицы.

Для получения статистических данных нами была разработана программа на VBA Excel, которая позволяет вычислять меры MI и t-score для заданных неоднословных единиц в заданных документах. При этом формируется запись, в которой отражено, в каких документах была обнаружена определенная неоднословная единица, каков объем этих документов по количеству слов и чему равны MI и t-score в данном случае. Программа может обрабатывать любые документы, которые открываются с помощью Microsoft Word. Были проверены форматы doc, docx, rtf, txt.

Выявить неслучайность совместной встречаемости компонентов словосочетаний можно с помощью статистических мер, позволяющих сделать это с достаточно высокой точностью. В настоящий момент существует немало статистических мер, дающих возможность выявить несвободные сочетания слов, но чаще других используются меры MI и t-score.

Мера MI (mutual information) сравнивает зависимые контекстно-связанные частоты с независимыми. Она наиболее чувствительна к низкочастотным сочетаниям, составляющие которых гораздо чаще встречаются вместе, нежели по отдельности. Мера t-score также учитывает частоту совместной встречаемости слов и позволяет выяснить, насколько устойчиво рассматриваемое словосочетание. Она, наоборот, направлена на вычленение сочетаний слов, присущих всему массиву коллекции[9]. В русском языке сочетание слов будет считаться статистически значимым, если MI больше или равно трем.

Мера t-score также учитывает частоту совместной встречаемости ключевого слова и зависимого, отвечая на вопрос, насколько неслучайной является сила ассоциации (связанности) между ними.

Для этих мер ассоциации характерны следующие особенности.

1. Мера t-score позволяет выявлять несвободные сочетания, характерные для данного типа текстов вне зависимости от их тематики [10].

2. Мера MI более всего подходит для выявления низкочастотных специальных терминов.

Работа с неоднословными единицами, занесенными в таблицу «Словник_Основной», оказалась сложнее, чем предполагалось изначально, так как пришлось учитывать то, что слова, составляющие неоднословную единицу, могут встречаться в разных грамматических формах (роде, числе, падеже и т.д.). Для точного подсчета требуется учесть все возможные варианты. Решить эту проблему с помощью средств Excel и VBA достаточно сложно. Удобнее воспользоваться ресурсами, где уже перечислены все возможные формы каждого слова. В качестве такого ресурса можно выбрать русский Викисловарь [11]. Создаваемая нами программа выбирает нужные формы слова, сравнивая содержимое таблиц на странице Викисловаря с заданной пользователем основой слова, при этом она учитывает то, что в таблицах могут быть указаны два варианта одной формы слова, разделенные запятой или пробелом, а также то, что в словах обычно отмечены ударные слоги. Без этого работа с Викисловарем оказалась бы невозможной. Если одно из слов было неизменяемым или было задано как неизменяемое (например, если нам не нужны все формы слова «необходимый» в словосочетании «считать необходимым», а только данная форма в косвенном падеже), то в соответствующем столбце таблицы будет заполнена только первая ячейка заданной формой слова.

Для обработки любого количества файлов и получения информации о них, нужно будет сохранить эти файлы следующим образом: все документы, подготовленные для обработки, должны быть занесены в папки, названные в соответствии с типом документов (например, папка «Приказ»), каждому документу будет дано оригинальное название.

При работе программы на рабочем листе «Лист3» того же файла появляются сведения об обрабатываемых документах:

- тип документа – определяется по названию папки;
- название документа;
- ключевое слово и коллокат неоднословной единицы, полученные из соответствующего списка;
- встречаемость каждого из этих слов по отдельности и неоднословной единицы в целом в обрабатываемых документах;
- результаты расчетов мер MI и t-score.

Количество обрабатываемых с помощью нашей программы документов может быть так велико (десятки тысяч), что позволит с большой точностью определять встречаемость неоднословных единиц.

ЗАКЛЮЧЕНИЕ

Разработанная нами программа для выявления наиболее характерных для современной русской деловой и официальной речи неоднословных единиц может использоваться для того, чтобы можно было выяснить, вносить ли каждую данную неоднословную единицу в пособие по официальной и/или деловой речи, а также проверить, не является ли эта неоднословная единица уже неупотребимой в данное время в данном типе документов.

В качестве перспектив дальнейшей работы можно рассматривать:

- определение оптимального способа вывода, хранения и обработки полученных данных;
- нахождение возможности исправления и добавления форм слов, составляющих неоднословную единицу на рабочем листе «Лист3», предположительно с помощью InputBox. Это необходимо в тех случаях, когда основа слова при склонении или спряжении меняется. Также может потребоваться удаление ненужных форм слова, если мы имеем дело с омонимами.

В сфере преподавания языка результаты, полученные с помощью разрабатываемой нами программы, могут использоваться для наглядной демонстрации того, насколько в официальной или деловой речи характерны или нехарактерны те или иные неоднословные единицы. При этом возможно уточнить, в каких именно документах встречаются эти единицы и с какой частотой. Полученные результаты можно использовать и для составления учебных пособий, проверяя встречаемость неоднословных единиц, оценивая их частотность, составляя графики и диаграммы. В итоге можно найти статистическое подтверждение тому, что именно эти неоднословные единицы являются наиболее часто встречающимися в официально-деловой речи, поэтому именно их учащимся нужно освоить в первую очередь.

Составитель документа может сверяться с нашими данными, чтобы быть уверенным, что подобное сочетание слов уместно в данном типе документов и не является устаревшим, редким или не соответствующим документу.

СПИСОК ЛИТЕРАТУРЫ

1. Кожина М.Н. К основаниям функциональной стилистики. – Пермь: Пермский ун-т, 1968. – С. 251.
2. Романова С.В., Маркина Н.А. Русский язык для делового общения: Пособие для изучающих русский язык как иностранный. – М.: Русский язык. Курсы, 2013. – С. 264.
3. НКРЯ – Национальный корпус русского языка. – 2016. – URL: ruscorpora.ru (дата обращения 15.12.2016)
4. КУЛ – Корпус Университета Лидс. – 2016. – URL: http://corpus.leeds.ac.uk/ruscorpora.html (дата обращения 15.12.2016)
5. Буре Н.А., Волкова Л.Б., Косарева Е.В. Основы русской деловой речи. Учебное пособие для студентов высших учебных заведений. – ООО Центр «Златоуст», 2012. – С. 448.
6. Гусева О.Н. Фразеология научной и деловой речи // Труды БГТУ. – 2014. –№ 5. – С. 152.
7. Сологуб О.П. Официально-деловой текст в системно-структурном аспекте. – 2009. – URL: http://siberia-expert.com/publ/oficialno_delovoj_ tekst_ v_sistemno_strukturnom_aspekte_op_sologub/ 3-1-0-57 (дата обращения: 05.12.2016)
8. Формановская Н.И. Речевой этикет и культура общения. – М.: Высшая школа, 1989. – С. 159.

9. Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке. – 2010. – URL: http://www.dialog-21.ru/digests/dialog2010/materials/html/22.htm (дата обращения: 15.12.2016)
10. Буторина Е.П., Соловьева К.В. Выявление несвободных сочетаний слов как основы институционального дискурса // Международный аспирантский вестник. Русский язык за рубежом. – 2012. – № 2. – С. 11–14.
11. Русский Викисловарь. – 2016. – URL: https://ru.wiktionary.org/wiki (дата обращения 18.12.2016).

Материал поступил в редакцию 10.02.17.

Сведения об авторах

БУТОРИНА Елена Петровна – кандидат филологических наук, доцент кафедры русского языка Института лингвистики Российского государственного гуманитарного университета (РГГУ), Москва
e-mail: ep_butorina@il-rggu.ru

ГУБАНОВА Елизавета Олеговна – аналитик некоммерческой организации Фонд «Московский предприниматель», Москва
e-mail: kharibdaa@yandex.ru

ИНФОРМАЦИОННОЕ ПИСЬМО И ПРИГЛАШЕНИЕ

МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ К 65-ЛЕТИЮ ВИНИТИ РАН

«ИНФОРМАЦИЯ В СОВРЕМЕННОМ МИРЕ»

Москва, 25-26 октября 2017 г.

подробная информация на сайте: http://www.viniti.ru

Главный организатор:

Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН)

Соорганизаторы:

Российская академия наук
Федеральное агентство научных организаций
Российский фонд фундаментальных исследований
Министерство образования и науки РФ

Проблемно-тематическое направление конференции: современный издательский процесс, интеллектуальная собственность, научные библиотеки, информационное обеспечение научной и инновационной деятельности, информационные технологии для научной и библиотечной отрасли, информационная безопасность, международное сотрудничество и информационный обмен, инфометрия, классификации, стандартизация, образование для отрасли, экономика информации

Основные вопросы, предлагаемые к обсуждению:

● Популяризация научных знаний: Новые модели распространения научной информации
● Редакционно-издательская деятельность в цифровой среде: продукты и сервисы
● Издательские стандарты и технологии
● Перспективы развития книжного дела. Проекты и программы
● Взаимодействие цифровых и печатных ресурсов в научно-технической библиотеке
● Информационно-библиотечное обслуживание: сервисный подход
● Управление данными и навигация в современной научной библиотеке
● Научные библиотечные консорциумы – основные подписчики на научную литературу
● Перспективы развития национальных систем научно-технической информации
● Государственные проекты и программы поддержки информационного обеспечения научно-образовательной деятельности
● Тенденции развития региональных аналитических центров
● Информационное обеспечение экспертной деятельности. Использование информационно-аналитических систем для управления наукой и образованием
● Формальные и неформальные каналы развития современных научных коммуникаций

● Современные агрегаторы научной литературы открытого доступа как источник научной информации
● Машинная обработка данных и аналитические исследования: Приоритеты и сотрудничество
● Использование специальных сервисов компании CrossRef для идентификации научных публикаций
● Роль поисковых систем в современном издательском процессе
● Защита данных от несанкционированного использования. Маркеры безопасности. Политика безопасности открытых систем
● Вопросы достоверности и доверенности при обработке информационного потока
● Межгосударственный обмен научно-технической информацией на евразийском пространстве
● Информационное взаимодействие в рамках СНГ
● Международное партнерство при хранении и обработке больших массивов данных
● Современное состояние систем классификации знаний как инструмента индексирования и поиска данных по перспективным направлениям науки и критическим технологиям
● Современные библиометрические методы определения научных лидеров: Новые математические модели
● Анализ читательской аудитории научной литературы путем вебометрического анализа
● Подготовка специалистов в сфере научно-информационной деятельности
● Мастер-класс по работе с классификационными системами (УДК, ГРНТИ)
● Информация как источник цифрового капитала и фактор социальных изменений
● Информационная деятельность как фактор национальной экономики
● Новейшие бизнес-модели для публикаций открытого и закрытого доступа

На конференции планируются доклады представителей ведущих информационных центров и научно-технических библиотек России, СНГ и дальнего зарубежья.

В рамках юбилейной конференции состоится научно-практический семинар по классификационным системам «Перспективные направления научных исследований и критические технологии в классификационных системах». Предполагается проведение специализированных обучающих мероприятий по УДК индексированию. Запланировано заседание методического совета пользователей ГРНТИ и УДК. Участники конференции получат свидетельства о повышении квалификации.

Материалы конференции будут опубликованы в сборнике Трудов и на CD-ROM, основные – в сборнике **«Научно-техническая информация»**.

Доклады

Принимаются оригинальные работы, имеющие научное и прикладное значение, соответствующие тематическим направлениям конференции и НЕ ОПУБЛИКОВАННЫЕ ГДЕ-ЛИБО РАНЕЕ.

Предлагаемый доклад должен отвечать следующим требованиям:

1. Необходимо указать название доклада, фамилию, имя, отчество (полностью) авторов/соавторов, название организации, город, страну, выделить автора, который будет представлять доклад.

2. Необходимо наличие аннотации, раскрывающей содержание доклада. Размер аннотации - не более 850 знаков (включая пробелы).

3. Доклады принимаются только в электронной форме; тексты – в формате MS Word; схемы, диаграммы, фотографии, сканированные виды экранов и т. п. - в формате JPG. Объем доклада вместе с аннотацией, рисунками, приложениями и т.п. не более 10 страниц формата А4.

4. Доклад необходимо выслать по электронной почте до 11 сентября 2017 г. в адрес оргкомитета: conf@viniti.ru

Доклады, не соответствующие вышеуказанным требованиям,
НЕ РАССМАТРИВАЮТСЯ.

Программный комитет оставляет за собой право определять статус доклада (пленарный доклад, доклад, стендовый доклад), включать принятые доклады в те или иные секции.

Время для выступления: пленарные доклады – 15–20 мин., доклады на отдельных мероприятиях – до 10 мин. Доклады включаются в Труды на основании решения экспертов оргкомитета.

Контакты: 125190, Москва, ул. Усиевича, 20, ВИНИТИ РАН
Телефоны: 8 (499) 152 61 13, 8 (499) 155 42 52, 8 (499) 151 02 61. Факс 8 (499) 943 00 60
Интернет-сайт: http://www.viniti.ru Эл. почта: conf@viniti.ru

База данных (БД) ВИНИТИ РАН

Федеральная база отечественных и зарубежных публикаций по естественным, точным и техническим наукам, генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. Тематическое наполнение соответствует реферативному журналу ВИНИТИ. Для поиска одновременно по всем или нескольким тематическим фрагментам генерируется единая Политематическая БД.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ - http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов за любой период с 1981 г., а также **проблемно-ориентированные выборки** из БД ВИНИТИ по актуальным направлениям научных исследований могут быть предоставлены на договорной основе **в поисковой системе (ИПС) "Сокол"**, работающей под управлением Microsoft Windows и обеспечивающей следующие возможности:

- **Чтение** документов в режиме последовательного просмотра или выборочно по оглавлению за весь период заказанной ретроспективы
- **Поиск** документов по автору, заглавию, источнику, ключевым словам или словосочетаниям, реферату, рубрикам, году издания, стране, языку и т.д. (всего более 20 признаков)
- **Словарь** системы поможет правильно подобрать термины для поиска и выбрать глубину их усечения.
- **Для уточнения поиска** можно дополнительно использовать год издания документа, язык текста документа, рубрики, шифры тематических разделов БД.
- Выполненные **запросы можно сохранять** для их последующего использования и/или редактирования.

125190, г. Москва, ул. Усиевича, 20, БД ВИНИТИ РАН.
Отдел взаимодействия с потребителями – (499) 155-45-25, (499) 152-58-81
E-mail: csbd@viniti.ru, sales@viniti.ru

WWW: http://www.viniti.ru

УВАЖАЕМЫЕ КОЛЛЕГИ!

ВИНИТИ РАН предлагает Вашему вниманию Реферативный Журнал в электронной форме

РЖ в электронной форме (ЭлРЖ) выпускается по всем разделам естественных, технических и точных наук.

Каждый номер ЭлРЖ является полным аналогом печатного номера РЖ по составу описаний документов, их оформлению и расположению. Он сопровождается оглавлением, указателями.

ЭлРЖ представляет собой информационную систему, снабженную поисковым аппаратом и позволяющую пользователю на персональном компьютере:

- читать номер РЖ, последовательно листая рефераты;
- просматривать рефераты отдельных разделов по оглавлению;
- обращаться к рефератам по указателям авторов, источников, ключевых слов;
- проводить поиск документов по словам и словосочетаниям;
- выводить текст описаний документов во внешний файл.

ЭлРЖ в версии Windows Вы можете получить за текущий год с любого номера, а также за предыдущие годы.

Подробную информацию Вы можете получить:

Адрес: 125190, Россия, Москва, ул. Усиевича, 20, ВИНИТИ РАН

Коммерческое управление

Телефон/Факс: 8 (499) 155-45-25, 8 (499) 152-58-81

E-mail: contact@viniti.ru, sales@viniti.ru