5. *Labsker L.G.* Teoriya kriteriyev optimal'nosti i ekonomicheskie resheniya. M.: Knorus, 2012. 744 s.
6. *Arrow K.J., Hurwitz L.* An optimality criterion for decision making under ignorance // Uncertainty and expectations in economics. Oxford: Basil Blackwell and Mott, 1972.
7. *Savage L.J.* The foundation of statistics. N.Y.: Wiley, 1954.
8. *Wald A.* Contribution of the theory of statistical estimation and testing hypothesis // Annuals Math. Statist. 1939. Vol. 10. P. 299–326.

DOI: 10.25586/RNU.V9187.19.02.P.080 УДК 004.8

А.С. Башков, Я.К. Соломенцев

ИСПОЛЬЗОВАНИЕ ВЕКТОРНЫХ МЕТОДОВ ПРЕДСТАВЛЕНИЯ СЛОВ В ЗАДАЧАХ ВЫЯВЛЕНИЯ ТРЕНДОВ

Описываются методы обработки текстов естественного языка на основе нейронных сетей. Обрабатываются научные статьи для выявления тенденций развития научных направлений. Приводятся примеры разбора текста посредством морфологического анализатора Pullenti. Анализируется метод word2vec, основанный на нейронных сетях, описывается применение алгоритма этого метода Skip-gram. Приводятся результаты использования метода word2vec для построения трендов развития научных направлений.
Ключевые слова: интеллектуальный анализ данных, приоритетные направления, прогнозирование, нейронные сети, векторное представление слов.

A.S. Bashkov, Ya.K. Solomentsev

THE USE OF VECTOR METHODS OF REPRESENTING WORDS IN THE TASKS OF REVEALING TRENDS

This article describes natural language text processing techniques based on neural networks. Scientific articles are being processed to identify trends in the development of scientific fields. Examples of text parsing via morphological analyzer Pullenti are given. The word2vec method based on neural networks is analyzed, the application of the algorithm of the Skip-gram method is described. In conclusion, the results of using the word2vec method for building trends in the development of scientific fields are presented.
Keywords: data mining, priority directions, prediction, neural networks, vector word representation.

Введение

Выявление тенденций (трендов) в научном прогрессе является важной задачей: она решается международными организациями, государствами, научными учреждениями и крупным бизнесом. Определение трендов важно для построения прогнозов, на основе которых принимаются решения о дальнейшем развитии государства, общества, компании, выделяются финансовые и другие ресурсы [4].

В научной среде все значимые результаты исследований и информацию об открытиях принято в первую очередь публиковать в рецензируемых научных журналах. Затем происходит обсуждение опубликованных результатов, выходят публикации в средствах массовой информации. Поэтому мониторинг трендов в науке целесообразно проводить путем анализа опубликованных научных статей как первоисточника новых знаний.

С учетом растущего объема научных публикаций (рис. 1) задача по их мониторингу усложняется и требует применения методов автоматизированного анализа.

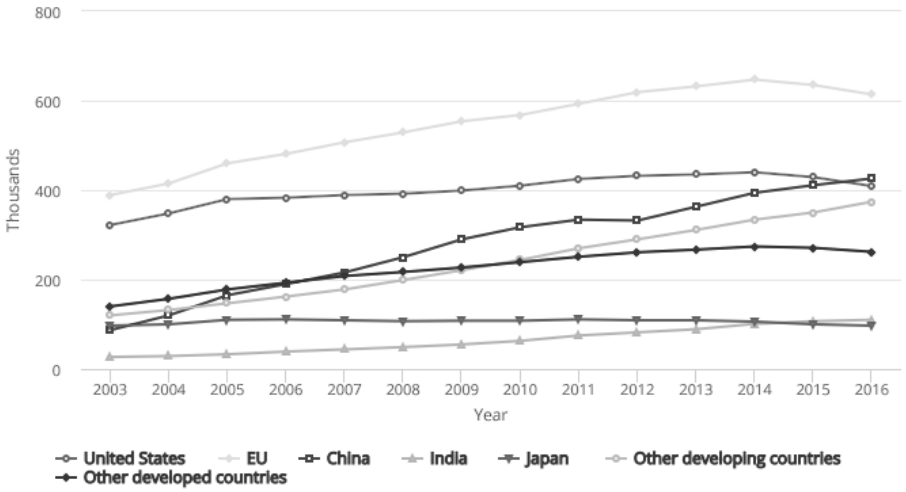

Рис. 1. Количество научных статей по отдельным странам в 2003–2016 гг. (данные Национального совета по науке США) [8]

В 2018 г. в базе данных Web of Science зарегистрировано 1,62 млн научных статей [7]. Это число на 5% выше, чем в 2017 г., и является самым высоким за всю историю.

В базе данных Scopus содержится 83 тыс. статей российских ученых за 2017 г., Россия занимает 12-е место по количеству статей после США, Китая, Индии, Японии, Канады, Австралии и ряда европейских стран [10].

Машинная обработка текста

При машинной обработке текста необходимо закодировать текст в том виде, который может обрабатываться компьютером. Самый простой способ – пронумеровать все имеющиеся слова номерами по порядку. Но в этом случае номер слова не связан с его значением. Допустим, что слова в словаре следуют по алфавиту. Тогда близкие номера будут у слов, имеющих одинаковые буквы в начале слова. Чтобы получить модель, отражающую семантическую близость слов, необходимо сопоставить со словом вектор, который отображал бы его в пространстве по смыслу.

Возьмем вектор размерностью, равной количеству слов в словаре, и закодируем каждое слово проставлением единицы в одной позиции, соответствующей порядковому номеру слова в словаре. Такое представление называется one-hot encoding – OHE. Векторы в таком пространстве не отражают смысловой близости слов [11].

Смысл слова может определяться его контекстом, т.е. словами, стоящими с ним рядом. Приведем пример из задания «Вставьте в предложения пропущенные слова» учебника для начальных классов:

По _____ мчатся машины.

Ученики ждали учителя у дверей _____.

В указанных примерах можно вставить различные слова, например: «дороге», «шоссе», «улице» и «класса», «школы», «автобуса». В первом примере все слова близки по смыслу, во втором – более далекие. Таким образом, использование контекста задает с некоторой вероятностью смысл слова, а в случае большого набора контекстов такие вероятности можно рассчитать достаточно точно.

Существуют различные методы задания векторного представления слов. Также используются представления в виде «слово-документ», которые порождают терм-документную матрицу для корпуса текстов.

До появления нейронных сетей для анализа близости слов составляли матрицы частотности каждого слова. В такой матрице по горизонтали и по вертикали были отложены слова, а в ячейках указывалась частота появления слова в указанной строке со словом, указанным в столбце.

Для реализации указанного подхода потребуется построить матрицу встречаемости слов, размерность которой будет равна $N \times N$, где N – количество слов в языке. Расчет такой матрицы – слишком сложная вычислительная операция, поэтому были найдены другие методы, которые позволяют анализировать контекст [1].

Метод word2vec

В основе методов word2vec лежит предположение о том, что необходимо учитывать только наиболее близкий к слову контекст, т.е. не далее чем на 2–5 слов по тексту. Это предположение существенно упрощает задачу анализа контекста в текстах больших объемов.

Вначале word2vec необходимо обучить на определенном корпусе текстов. При обучении по корпусу текстов проходит скользящее окно задаваемого размера, в рамках которого и учитывается контекст. Для слов рассчитывается их векторное представление. При этом слова, встречающиеся в тексте близко, будут иметь близкие векторы (по косинусной мере) [9].

Полученная модель в дальнейшем используется для решения двух типов задач:

1. На вход задаем контекст (несколько слов), на выходе получаем наиболее подходящее слово в контексте (CBOW).

2. На вход задаем одно слово, на выходе получаем возможный контекст к этому слову (Skip-gram).

Алгоритм Skip-gram может использоваться в задачах по выявлению трендов, поэтому рассмотрим его реализацию подробнее.

Особенностью использования метода word2vec на русскоязычных текстах является необходимость нормализации слов (приведения всех форм слова к одной). Такая задача не возникает в случае работы с текстом на английском языке. Кроме того, перед обучением необходимо исключить из текста стоп-слова, не несущие смысла. Обе эти задачи могут быть решены с использованием открытого программного решения Pullenti.

Предобработка русскоязычного текста с помощью Pullenti

Программа Pullenti позволяет:

- разбивать текст на слова;
- производить морфологический анализ: определять все возможные части речи слова (независимо от контекста), нормализовать слова к единому падежу, роду, числу;
- выделять именованные сущности;
- производить ряд функций с числовыми, именными и глагольными группами, скобками, кавычками [2].

Приведем пример использования Pullenti через Python (Jupyter Notebook). Нижеприведенная программа выделяет именные группы из произвольного текста.

Пример текста: «Масштабный железнодорожный проект столичных властей заработает уже в 2019 году, заявил заммэра. Движение по первым центральным диаметрам запустят "в ускоренном режиме". Прежде ввести их в эксплуатацию планировалось в конце 2018-го года».

В результате Pullenti выделила и привела к нормальной форме следующие именованные группы:

МАСШТАБНЫЙ ЖЕЛЕЗНОДОРОЖНЫЙ ПРОЕКТ
ЖЕЛЕЗНОДОРОЖНЫЙ ПРОЕКТ
ПРОЕКТ
СТОЛИЧНАЯ ВЛАСТЬ
ВЛАСТЬ
УЖ
ГОД
ДВИЖЕНИЕ
ЦЕНТРАЛЬНЫЙ ДИАМЕТР
ДИАМЕТР
УСКОРЕННЫЙ РЕЖИМ
РЕЖИМ
ЭКСПЛУАТАЦИЯ
КОНЕЦ
2018 ГОД
ГОД

Как видно из приведенного списка именованных групп, алгоритм включил слово «уже» в этот список, поскольку слово «уже» может принимать именованную группу от слова «уж». Pullenti не включает в себя функции определения контекста, поэтому в результате возможны такие погрешности.

Применение алгоритма Skip-gram

Рассмотрим простейший случай алгоритма Skip-gram, когда нужно предсказать одно соседнее слово, если дано одно слово. Далее можно распространить этот случай на несколько слов.

Для обучения нейронной сети на вход нужно подавать пары слов. Для этого нужно выбрать размер окна, пройтись окном по предложению и перебрать все пары слов в данном окне. Если выбрать окно, равное одному, то окно будет содержать одно слово слева

от целевого и одно слово справа от целевого. В случае если окно было бы равным двум, то слева и справа от целевого слова было бы по два слова. Ниже приведен пример для окна, равного двум:

Старый	*заброшенный*	*дом*	*стоит на опушке леса*

Тренировочные фразы*: (Старый, заброшенный), (Старый, дом)*

Старый	*заброшенный*	*дом*	*стоит*	*на опушке леса*

Тренировочные фразы*: (заброшенный, Старый), (заброшенный, дом), (заброшенный, стоит)*

Старый	*заброшенный*	*дом*	*стоит*	*на*	*опушке леса*

Тренировочные фразы*: (дом, Старый), (дом, заброшенный), (дом, стоит), (дом, на)*

Старый	*заброшенный*	*дом*	*стоит*	*на*	*опушке*	*леса*

Тренировочные фразы*: (стоит, заброшенный), (стоит, дом), (стоит, на), (стоит, опушке)*

Нейронная сеть обучится статистике частоты появления каждой пары слов [12].

Каждое слово нужно преобразовать в цифровой вид. Представим его в виде one-hot encoding:

$$\vec{x} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}.$$

Здесь наше слово, которое мы представляем в виде вектора, занимает второе место в словаре.

Преобразования при помощи нейронной сети представлены на рисунке 2.

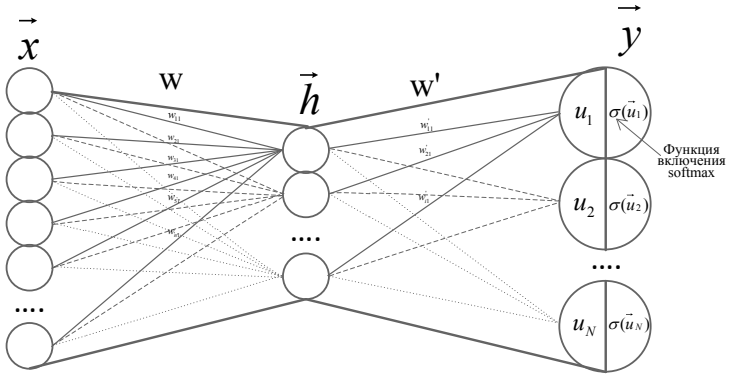

Рис. 2. Принципиальная схема нейронной сети

Здесь x – входное слово (или несколько слов), по которому (которым) мы хотим предсказать y – слово (или несколько слов);

H (скрытый слой нейронной сети) – вектор, получаемый при умножении вектора-слова x на матрицу весовых коэффициентов w:

$$\vec{h} = w^T \vec{x},$$

где w – матрица, содержащая весовые коэффициенты, она имеет размерность: (длина словаря) × (количество признаков). Количество признаков задается один раз перед запуском нейронной сети, оно подбирается для получения лучшего результата. Например, Google использовал 300 признаков для обучения нейронной сети на множестве данных новостей Google. Весовые коэффициенты в начальный момент времени принимают случайные значения, далее они корректируются в соответствии с методом обратного распространения ошибки [5].

После скрытого слоя h с учетом другой матрицы весовых коэффициентов w' образуется вектор u:

$$\vec{u} = w'^T \vec{h} = w'^T w^T \vec{x}.$$

Размерность вектора $\vec{u}$ совпадает с размерностью вектора $\vec{x}$.

Чтобы нормализовать выходной вектор $\vec{y}$ в диапазоне [0; 1], применим функцию softmax (используем в качестве функции активации, см. $\sigma(\vec{u}_i)$ на рис. 2):

$$y_i = \sigma(u_i) = \frac{e^{u_i}}{\sum_{k=1}^{N} e^{u_k}},$$

где N – количество признаков.

В результате получим, что y_i – вероятность наблюдения (предсказания) i-го слова (или фразы) в словаре при входящем слове (контексте) x.

Целью нейронной сети (см. рис. 2) является определение весовых коэффициентов w и w'. Критерием схождения вычислений выступает максимизация вероятности y при всех возможных выходных словах (фразах). В результате математических преобразований (взятие логарифма вероятности y, далее вычисление производной логарифма вероятности y по переменной w') получится уравнение, у которого невозможно найти оптимум. Поэтому придется воспользоваться численными методами.

Одним из лучших численных методов является метод градиентного спуска. В результате получится, что нужно решить рекурсивную задачу:

$$w'^{new} \leftarrow w'^{old} - G(w(1-y)),$$

где G – функция градиентного спуска [6].

Таким образом, если вероятность для перебираемого выходного слова максимальна, то скобка близка к нулю и $w'^{new} \leftarrow w'^{old}$. В ином случае, когда вероятность выхода слова очень мала, то от w' отнимается доля значений w. Таким образом, матрица w' приближается к матрице w.

Аналогичным образом можно приблизить w к w':

$$w^{new} \leftarrow w^{old} - G(w'(1-y)).$$

Но описанный метод не применяется на практике, поскольку вычисление функции softmax затратно по продолжительности. Поэтому в алгоритме используется поправка «Negative Sampling» (отрицательная выборка).

Пример работы алгоритма

Алгоритм был запущен на корпусе статей по медицинской тематике.

В исходных текстах было всего 218 137 слов.

После обработки Pullenti стало 165 590 слов.

После обработки модулем word2vec корпуса текстов после проведения обучения получились следующие контексты к ключевым словам:

ключевое слово «лечение» – контекст: [('медикаментозное', 3.296358e-05), ('головной', 3.296348e-05), ('лечение', 3.2963446e-05), ('хирургическое', 3.2963435e-05), ('проводится', 3.296343e-05), ('нарушений', 3.2963384e-05), ('местное', 3.2963366e-05), ('методами', 3.2963337e-05), ('головные', 3.2963333e-05), ('комплексное', 3.2963275e-05)];

ключевое слово «профилактика» – контекст: стаканом, трава, настоять, головные, клиническая, процедить, течение, путей, перечной, можно;

ключевое слово «болезнь» – контекст: гипертоническая, желудка, язвенная, области, дней, давления, Бехтерева, двенадцатиперстной;

ключевое слово «синдром» – контекст: болевой, нефротический, раствора, заболеваниях, смеси, кипятка, кровообращения, быть, использовать, воды.

Использование word2vec в задаче выявления трендов

Рассмотрим использование описанного выше метода Skip-gram для выявления трендов в статьях по мединской тематитике.

Выявление трендов состоит в необходимости анализа изменений, происходящих с корпусом статей по медицинской тематике с течением времени. Для этого необходимо выбрать временной интервал t, на котором будет проводиться анализ. Репрезентативные результаты на выборке статей даст применение метода при анализе корпуса статей не менее чем за один год. Можно разбить корпусы статей при $t = 3$ или $t = 5$ лет, для того чтобы изменения в тематиках статей были более заметными. Таким образом, при принятии промежутка $t = 5$ лет будут анализироваться следующие корпусы статей: с 31.12.2018 по 01.01.2014, с 31.12.2013 по 01.01.2009 и т.д.

Все статьи на исследуемом промежутке объединяются в один корпус, на нем производится обучение word2vec. После проведения обучения при помощи алгоритма Skip-gram устанавливаем контекст по ключевым для предметной области словам. Например, для слов: «лечение», «болезнь», «синдром», «диагностика», «профилактика», «рак».

Аналогичная операция выполняется для корпусов текстов других временных интервалов. При этом для каждого из них важно заново обучить word2vec.

Сравнение полученных слов, контекстных со словами-триггерами, в разные временные промежутки с учетом веса их встречаемости позволит отследить изменение трендов в исследованиях.

Заключение

Векторные методы анализа текстов могут успешно применяться в задачах выявления трендов. В качестве объекта исследований выбираются корпусы научных статей по ка-

кой-либо определенной тематике, разбитые на репрезентативные временные интервалы (от года до пяти лет).

Программа Pullenti позволяет провести нормализацию текста для анализа без учета окончаний слов. При этом она не учитывает контекст слова при его нормализации, из-за чего в редких случаях может обрабатывать слова некорректно. Однако на больших корпусах текстов такие погрешности не должны влиять на конечный результат.

Алгоритм Skip-gram word2vec позволяет найти контекст к ключевым словам в тематике, которые являются ключевыми для определения трендов. Недостаток этого подхода заключается в необходимости задавать ключевые слова самостоятельно, в результате чего возникает возможность упустить из анализа тренды, не связанные с ключевыми словами.

Литература

1. *Золотарев О.В., Шарнин М.М., Еромасова А., Тезадова Ф.М.* Современные подходы к обработке многоязычных текстов, основанные на методах дистрибутивной семантики // Сборник трудов международной научной конференции по физико-технической информатике – СРТ2018 (Пущино, 28–31 мая 2018 г.). Протвино, 2018. С. 43–47.

2. *Золотарев О.В., Шарнин М.М., Клименко С.В., Кузнецов К.И.* Система PullEnti – извлечение информации из текстов естественного языка и автоматизированное построение информационных систем // Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности – SCVRT2015-16: сб. тр. Междунар. конф. (Пущино, 21–24 нояб. 2016 г.): в 2 т. Протвино, 2016. Т. 2. С. 28–35.

3. *Золотарев О.В., Шарнин М.М., Клименко С.В., Мацкевич А.Г.* Исследование методов автоматического формирования ассоциативно-иерархического портрета предметной области // Вестник Российского нового университета. Серия «Сложные системы: модели, анализ и управление». 2018. № 1. С. 91–96.

4. *Микова Н., Соколова А.* Мониторинг глобальных технологических трендов: теоретические основы и лучшие практики // ФОРСАЙТ. 2014. Т. 8. № 4.

5. Ali Ghodsi, Lec 13: Word2Vec Skip-Gram. URL: https://www.youtube.com/watch?v=GMCwS7tS5ZM/

6. *Jurafsky D., Martin J.H.* Speech and Language Processing (3rd ed. draft, 2018). URL: // http://web.stanford.edu/~jurafsky/slp3/

7. *Makri A.* Pakistan and Egypt had highest rises in research output in 2018. URL: https://www.nature.com/articles/d41586-018-07841-9

8. National Science Board – Science & Engineering Indicators 2018. URL: https://www.nsf.gov/statistics/2018/nsb20181/

9. models.word2vec – Word2vec embeddings. URL: https://radimrehurek.com/gensim/models/word2vec.html#gensim.models.word2vec.Word2Vec/

10. Scimago Journal & Country Rank. URL: https://www.scimagojr.com/countryrank.php?-year=2017

11. Word2Vec: как работать с векторными представлениями слов. URL: https://neurohive.io/ru/osnovy-data-science/word2vec-vektornye-predstavlenija-slov-dlja-mashinnogo-obuchenija/

12. Word2Vec Tutorial – The Skip-Gram Model. URL: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Literatura

1. *Zolotarev O.V., Sharnin M.M., Eromasova A., Tezadova F.M.* Sovremennye podkhody k obrabotke mnogoyazychnykh tekstov, osnovannye na metodakh distributivnoy semantiki // Sbornik trudov mezhdunarodnoy nauchnoy konferentsii po fiziko-tekhnicheskoy informatike – CPT2018 (Pushchino, 28–31 maya 2018 g.). Protvino, 2018. S. 43–47.
2. *Zolotarev O.V., Sharnin M.M., Klimenko S.V., Kuznetsov K.I.* Sistema PullEnti – izvlechenie informatsii iz tekstov estestvennogo yazyka i avtomatizirovannoe postroenie informatsionnykh sistem // Situatsionnye tsentry i informatsionno-analiticheskie sistemy klassa 4i dlya zadach monitoringa i bezopasnosti – SCVRT2015-16: sb. tr. Mezhdunar. konf. (Pushchino, 21–24 noyab. 2016 g.): v 2 t. Protvino, 2016. T. 2. S. 28–35.
3. *Zolotarev O.V., Sharnin M.M., Klimenko S.V., Matskevich A.G.* Issledovanie metodov avtomaticheskogo formirovaniya assotsiativno-ierarkhicheskogo portreta predmetnoy oblasti // Vestnik Rossiyskogo novogo universiteta. Seriya "Slozhnye sistemy: modeli, analiz i upravlenie". 2018. № 1. S. 91–96.
4. *Mikova N., Sokolova A.* Monitoring global'nykh tekhnologicheskikh trendov: teoreticheskie osnovy i luchshie praktiki // FORSAYT. 2014. T. 8. № 4.
5. Ali Ghodsi, Lec 13: Word2Vec Skip-Gram. URL: https://www.youtube.com/watch?v=GMCwS7tS5ZM/
6. *Jurafsky D., Martin J.H.* Speech and Language Processing (3rd ed. draft, 2018). URL: // http://web.stanford.edu/~jurafsky/slp3/
7. *Makri A.* Pakistan and Egypt had highest rises in research output in 2018. URL: https://www.nature.com/articles/d41586-018-07841-9
8. National Science Board – Science & Engineering Indicators 2018. URL: https://www.nsf.gov/statistics/2018/nsb20181/
9. models.word2vec – Word2vec embeddings. URL: https://radimrehurek.com/gensim/models/word2vec.html#gensim.models.word2vec.Word2Vec/
10. Scimago Journal & Country Rank. URL: https://www.scimagojr.com/countryrank.php?year=2017
11. Word2Vec: kak rabotat' s vektornymi predstavleniyami slov. URL: https://neurohive.io/ru/osnovy-data-science/word2vec-vektornye-predstavlenija-slov-dlja-mashinnogo-obuchenija/
12. Word2Vec Tutorial – The Skip-Gram Model. URL: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

DOI: 10.25586/RNU.V9187.19.02.P.088 УДК 004.738; 614.842.4

А.А. Шуваев

ПЕРСПЕКТИВЫ РАЗВИТИЯ СИСТЕМ ОХРАННОЙ И ПОЖАРНОЙ СИГНАЛИЗАЦИИ

Рассмотрены проблемы современных проводных и беспроводных систем сигнализации, способы их устранения и предполагаемые перспективы развития. Обсуждаются надежность систем пожарной сигнализации и систем оповещения и управления эвакуацией, а также преимущества цифровизации и необходимость стандартизации. Приведены примеры перспективных систем сигнализации.

Ключевые слова: системы передачи тревожных извещений, охранно-пожарная сигнализация, оповещение о пожаре, LPWAN, IPv6, IoT.