имеет большое количество различных функций упрощающих процесс разработки программ и др. Еще одной технологией, которая применяется для разработки, стала Microsoft Azure [2], которая облегчит хранение временных баз данных.

Разработанная нами система учета археологических находок [3] сокращает временные затраты на процесс описания и обработку археологических находок, повышает степень достоверности обрабатываемой информации, исключает появление ошибок.

Библиографический список

1. Голощапов А.Л. Microsoft Visual Studio 2010. – СПб.: БХВ-Петербург. – 2011. – 544 с.

2. Таллоч Митч и команда Windows Azure. Знакомство с Windows Azure. Для ИТ-специалистов. – М.: ЭКОМ Паблишерз. – 2014. – 154 с.

3. Камышникова Н.Н., Шевченко А.С., Гумеров К.А., Грибенников А.В., Костенко В.В., Шалда С.В. Разработка информационной системы учета археологических находок // Современные научные исследования и разработки. – 2016. – №6(6). – С. 273–279.

УДК 004.032.26

Система автоматической кластеризации текстов с применением искусственных нейронных сетей

А.В. Шицелов, В.В. Бурлуцкий, В.В. Славский
ЮГУ, г. Ханты-Мансийск

Введение. В настоящее время обнаружение материалов в сети интернет, которые противоречат Российскому законодательству, является актуальной задачей. Для того чтобы быстро и эффективно находить такие материалы среди большого потока данных необходимо использовать специальные системы, которые способны автоматически определять принадлежность текста или его части к определённой категории. Сложность данной задачи определяется текстом, содержавшим большое количество специфических терминов и жаргонного сленга, а также орфографических ошибок. В статье описана система кластеризации текста с применением искусственных нейронных сетей. Искусственные нейронные сети в настоящий момент являются эффективным инструментом в задачах прогнозирования [6], распознавания образов [3], идентификации образов [4].

Модель кластеризации текста. Для любой обработки текста его следует преобразовать в числовой вид, так как компьютер умеет хо-

рошо оперировать числам, а не словами, к тому же числа занимают намного меньше памяти, нежели слово. Есть несколько способов преобразования слова в число. В разрабатываемой системе использовался способ, основанный на сопоставление каждому слову числового вектора, предварительно подобранный, так что бы максимально отражать семантику слова. Данный метод называется Word2Vec. Данный алгоритм построен на основе нейронных сетей глубокого доверия [2] и процесс подбора вектора называется обучением сети. При обучении нейронной сети глубокого доверия использовались подходы, описанные в [1, 5, 7]. После обучения сети word2vec каждому слову в словаре сопоставлен многомерный вектор, который очень сильно отражает семантику слова. Пример такой зависимости представлен на рисунке 1.

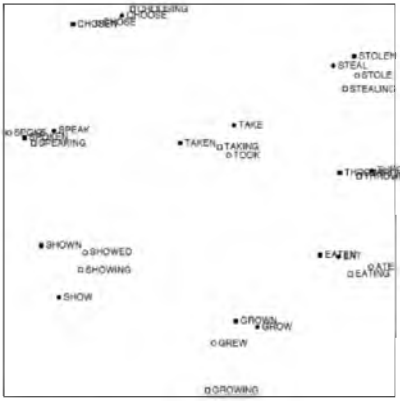

Рисунок 1 – Распределение слов в пространстве

Стоит так же упомянуть что перед подачей слов на вход word2vec, текст предварительно разбивается по словам, а также чистится от служебных слов и цифр. После обучения word2vec на входе мы получает очень большой словарь слов. К сожалению, в него попадает очень много похожих слов. Что и вызывает разрастание его в размере. Для дальнейшей работы необходимо уменьшить размер словаря, объединив похожие слова в одно общее слово, тем самым сильно уменьшив количество слов. Для реализации такой задачи было решено использовать нейронную сеть, а именно самоорганизующеюся сеть Кохонена. Идея обобщения заключается в следующем – похожие слова сгруппированы и при этом разные слова наоборот находятся далеко друг от друга (в разных группах). К тому же если учесть особенности работы word2vec, то можно понять, что к похожим словам будут отнесены все слова с ошибками в оригинальном слове, его синонимы, и его употребления в жаргонных сленгах, и близкие по значению термины, а

также слова из других языков. Суть обобщения всех этих слов в том, чтобы найти центр кластера слов и определять слова на некотором расстоянии от него как принадлежащие этому кластеру.

В теории кластеризации и классификации текста можно выделить два подхода:

1. Выделение общих черт текста через понимания самого текста и затем их кластеризация.

2. Выделение темы текста через нахождения слов среди текстов с одной темой.

Для реализации системы был выбран второй подход. В качестве алгоритма кластеризации был выбран LDA (Латентное распределение Дирихле). Главной причиной для выбора данного алгоритма стало то, что данный метод кластеризирует, а не классифицирует данные. По-другому говоря методы, при которых данные классифицируются предполагают, что у каждого текста из обучающей выборке есть заранее известная тема и на основе этой пары текст-тема происходит обучение. Кластеризация же не использует заранее известную тему. Алгоритмы кластеризации относятся к алгоритмам обучения без учителя и способны сами выделять набор тем среди текстов. В предоставленном наборе текстов для обучающей выборки отсутствовало разбиение текстов по темам это и стало поводом для проведения кластерного анализа за место классификации. LDA – иерархическая байесовская модель, состоящая из 2 уровней:

1. На первом уровне – смесь, компоненты которой соответствуют темам.

2. На втором уровне – мультиномиальная переменная с априорным распределением Дирихле, которое задаёт распределение тем в документе.

Пример разбиения текста на темы представлен на рисунке 2.

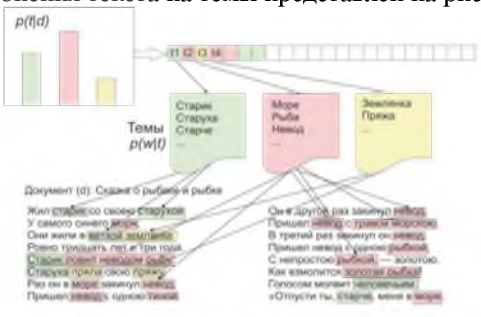

Рисунок 2 – Схема работы алгоритма LDA

В итоге получаем 4 этапа работы с текстом:

1. Обработка текста – разбитие текста на слова и откидывание не нужных слов.

2. Определение вектора слова – используя алгоритм word2vec, вычислить вектор для каждого слова в тексте.

3. Нахождение обобщающего слова – используя самоорганизующеюся нейронную сеть Кохонена для каждого слова, получаем другое слово, являющееся общим для всего кластера слов к которому относится исходное слово.

4. Определение категории, к которым принадлежит текст с использованием алгоритма LDA.

Библиографический список

1. Татьянкин В.М. Подход к формированию архитектуры нейронной сети для распознавания образов // Вестник Югорского государственного университета. – Ханты-Мансийск, 2016. – №2 (41). – С. 61–64.

2. Татьянкин В.М., Дюбко И.С. Нейронные сети глубокого доверия в сравнение с многослойным персептроном // Вестник Югорского государственного университета. – Ханты-Мансийск, 2015. – №2 (37). – С. 87–89.

3. Татьянкин В.М., Дюбко И.С. Обучающая выборка в задаче распознавания образов при использовании нейронной сети // Вестник Югорского государственного университета. – Ханты-Мансийск, 2015. – №2 (37). – С. 94–98.

4. Татьянкин В.М. Способ идентификации образов // Вестник Югорского государственного университета. – Ханты-Мансийск, 2015. – №2 (37). – С. 79–81.

5. Татьянкин В.М. Алгоритм формирования оптимальной архитектуры многослойной нейронной сети // Новое слово в науке: перспективы развития: материалы II Междунар. науч.-практ. конф. (Чебоксары, 30 дек. 2014 г.) / редкол.: О.Н. Широков [и др.]. – Чебоксары: ЦНС «Интерактив плюс», 2014. – С. 187–188.

6. Татьянкин В.М. Использование многослойных нейронных сетей в прогнозирование временных рядов // Приоритетные направления развития науки и образования: материалы III Междунар. науч.-практ. конф. (Чебоксары, 4 дек. 2014 г.) / редкол.: О.Н. Широков [и др.]. – Чебоксары: ЦНС «Интерактив плюс», 2014. – С. 195–197.

7. Татьянкин В.М. Модифицированный алгоритм обратного распространения ошибки // Приоритетные направления развития науки и образования: материалы III Междунар. науч.-практ. конф. (Чебоксары,

04 дек. 2014 г.) / редкол.: О.Н. Широков [и др.]. – Чебоксары: ЦНС «Интерактив плюс», 2014. – №3 (3). – С. 197–198.

УДК 004.5

Информационно-аналитический модуль геоинформационной системы доступности объектов социальной инфраструктуры

А.Р. Шугуров, С.П. Семенов
ЮГУ, г. Ханты-Мансийск

Во многих развитых странах мира значительное внимание уделяется проблеме перемещения маломобильных граждан. Один из аспектов такой проблемы – недостаточность информации об объектах социальной инфраструктуры (далее ОСИ). Возможным решением является разработка геоинформационной системы доступности мест социальной инфраструктуры. Система Geowheel [3, 4] является примером методики проектирования и реализации информационного ресурса на базе ГИС-технологий, отражающего комплексную оценку современного фонда городской застройки с точки зрения доступности для маломобильных граждан.

Однако, наряду с этим, существует потребность в динамическом представлении информационно-аналитических данных об ОСИ для эффективного принятия управленческих решений органами местного самоуправления.

Разработка информационно-аналитического модуля (ИАМ) геоинформационной системы доступности мест социальной инфраструктуры Geowheel позволит обеспечить доступ органов местного самоуправления к информационно-аналитическим данным для детальной проработки проблемы физической недоступности ОСИ и поиска путей ее эффективного решения.

В данной статье рассматривается проектирование и реализация информационно-аналитического модуля геоинформационной системы Geowheel.

Модуль разрабатывается с целью предоставления органам власти различного уровня систематизированной и структурированной оперативной информации по объектам социальной инфраструктуры в виде интерактивного веб-приложения для комплексного анализа состояния фонда городской застройки и принятия управленческих решений.

Процесс разработки модуля можно разделить на несколько основных этапов: