ББК Ч48
УДК 37:004.4

Д. С. Ботов, Ю. Д. Кленин

ПРИМЕНЕНИЕ АЛГОРИТМОВ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ И МОДЕЛЕЙ ВЕКТОРНОГО ПРЕДСТАВЛЕНИЯ СЛОВ В ЗАДАЧЕ АНАЛИЗА ОБРАЗОВАТЕЛЬНОГО КОНТЕНТА

Проведён обзор популярных подходов к извлечению ключевых слов применительно к задаче анализа образовательного контента (рабочих программ дисциплин). В ходе эксперимента по оценке качества работы алгоритмов извлечения ключевых слов, не требующих тренировочной выборки, был предложен подход с использованием модели векторного представления слов word2vec, решающей проблему семантического разрыва при сравнении результата работы алгоритмов с тестовыми наборами, составленными экспертами.

Введение

В условиях увеличения количества доступных вариантов получения высшего профессионального образования с выстраиванием индивидуальных образовательных траекторий наблюдается быстрый прирост объёмов существующего образовательного контента: учебных материалов, образовательных программ, программ курсов, рабочих программ дисциплин, фондов оценочных средств и т. п. Документация по образовательным программам регулярно актуализируется и перерабатывается образовательными организациями в условиях постоянно изменяющихся требований образовательных и профессиональных стандартов, потребностей работодателей.

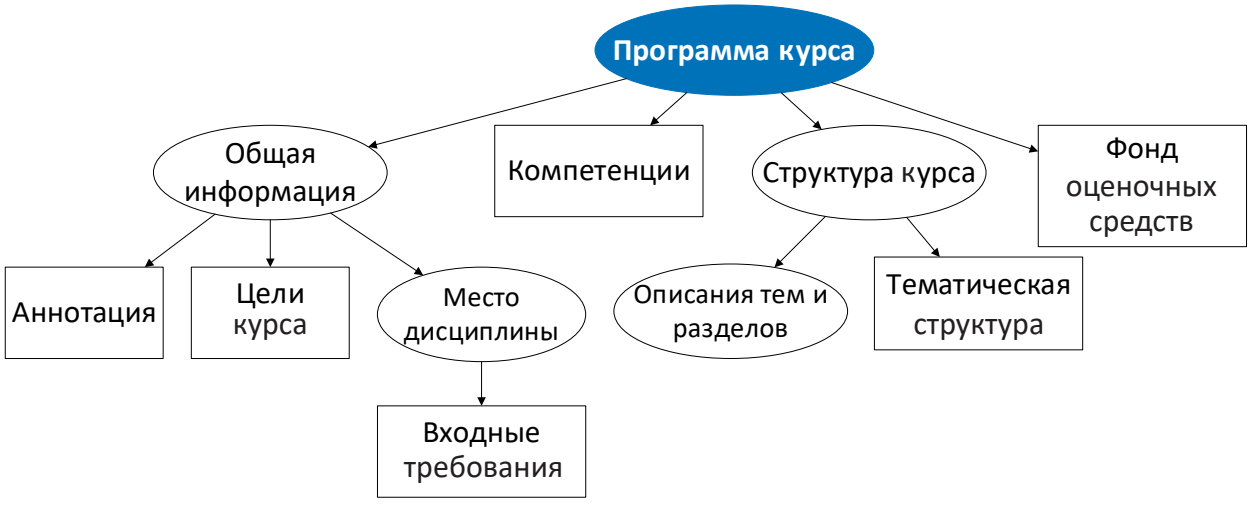
В нашем исследовании мы рассматриваем возможные варианты интеллектуального анализа образовательного контента с целью автоматизации процесса обработки больших массивов слабоструктурированных данных об образовательных программах, что позволило бы значительно уменьшить нагрузку на персонал образовательных организаций, упростить процесс разработки и актуализации образовательного контента и в конечном итоге повысить качество образовательного контента и уровень подготовки выпускников.

В предыдущей работе [1] рассматривались различные форматы представления информации о курсах и образовательных программах в образовательных учреждениях по всему миру и анализировались основные компоненты данных документов. В нашей текущей деятельности мы сфокусировались на проблеме анализа рабочих программ дисциплин (РПД) — официального формата представления данных об образовательных курсах, регулируемого образовательными стандартами, указами и инструкциями Министерства образования и науки РФ.

Анализ структуры и компонентов РПД

После изучения различных программ дисциплин была предложена следующая модель, отражающая структуру программы дисциплины как документа (рисунок).

Более детально рассмотрев программы дисциплин ведущих российских и зарубежных университетов, можно прийти к выводу, что наиболее важная информация, максимально точно отражающая суть описываемого в РПД курса, представлена следующими элементами данной модели: общая



Модель структуры документа рабочей программы дисциплины

информация, компетенции (результаты обучения) и структура курса.

Раздел общей информации содержит укрупнённое описание того, чему курс посвящён, кому его следует проходить, какие общие задачи и результаты преследуются в рамках изучения и как данный курс взаимодействует с предыдущими и последующими курсами в рамках общей образовательной программы.

В разделе компетенций приводится перечень компетенций образовательной программы, частью которой является данный курс, перечень результатов освоения дисциплины в виде конкретизированных знаний, умений и навыков, а также связи между компетенциями образовательного стандарта и конкретными результатами обучения.

Структура курса, как правило, представлена в виде списка разделов и тем в порядке их изложения на лекциях (иногда приводятся также перечни тем рассматриваемых на практиках и в рамках самостоятельной работы студентов), а также более детального описания содержания отдельных тем. Это содержание в подавляющем большинстве случаев имеет вид перечня концептов области знания, изучаемых в рамках каждой отдельной темы.

Поскольку формат РПД достаточно жёстко задаётся нормативными документами и сильно формализован, его структура, в том числе и отдельные формулировки, зачастую остаются неизменными от документа к документу. В текущем исследовании можно исходить из предположения, что каждый такой документ содержит два принципиальных подмножества лексики: первая — общая для всех документов данного типа, содержащая организационную терминологию и терминологию образовательного процесса, а вторая — относящаяся к конкретной области знания, изучаемой в рамках дисциплины.

Общая терминология образовательного процесса имеет, на наш взгляд, низкое значение при сопоставлении семантики различных курсов в сравнении с терминологией, описывающей, что именно будет пройдено в рамках курса с точки зрения изучаемой области знания дисциплины. Эта область знания, как правило, содержит ряд концептов и процессов, которые изучаются каждым конкретным курсом и представлены в рамках РПД конкретными терминами. Более того, в виду специфики этой терминологии можно считать, что эти термины будут употребляться в виде ключевых слов данного документа.

Исходя из данных соображений в данной работе фокус направлен на процесс извлечения ключевых слов из указанных выше разделов РПД.

Обзор алгоритмов извлечения ключевых слов

Существует множество разнообразных методов, связанных с получением набора ключевых слов для документа или коллекции документов. Эти методы и подходы можно условно разбить на несколько категорий: подходы на основе знаний предметной области, подходы на основе лингвистических знаний, подходы на основе простых статистик и подходы на основе машинного обучения и комбинации таковых [2; 3].

Методы, использующие знания конкретной предметной области или даже конкретной анализируемой коллекции, тесно связаны с понятием назначения ключевых слов (keyword assignment), обозначающим процесс выбора ключевых слов из внешней таксономии, онтологии или иной модели предметной области. Например, этот подход может быть применён путём замены потенциальных кандидатов в тексте на дескрипторы из словарей [4; 5]. Хотя такой подход и даёт возможность получить ключевые слова в виде заранее продуманных, качественных терминов, он требует наличия достаточно крупного источника, привязанного к конкретной области знания. В силу обширности областей знания, покрываемых рабочими программами дисциплин, мы не рассматриваем подходы данной группы в рамках настоящей работы.

Лингвистические подходы осуществляют извлечение ключевых слов исходя из лингвистических знаний о языке, используемом в документе. Лингвистический подход часто применяется для извлечения кандидатов в ключевые слова путём отбора только тех, что соответствуют паттернам частей речи [6; 7]. Используемые для этого Part-of-Speech-теги также применяются в качестве свойств кандидатов в методах, основанных на машинном обучении [6–8]. Отдельные исследования также уделяют внимание использованию последовательностей суффиксов в качестве более мелкой меры в сравнении с POS-тегом [9]. Исходя из лингвистических соображений часто производится выделение кандидатов, например, как последовательностей слов, разделённых пунктуацией и стоп-словами [10].

Значительная доля методов использует простые статистические меры при определении ключевых слов. Эти меры зачастую применяются как сами по себе, так и в качестве свойств для методов машинного обучения. В эту группу входит также и самая популярная и известная методика определения ключевых слов — взвешивание по схеме TF-IDF [9]. TF-IDF зарекомендовал себя как простой, быстрый, но при этом достаточно сильный базис в задаче извлечения ключевых слов [11]. Другими такими методами являются частоты слов, совместное употребление или расположение в тексте в рамках выбранного окна [12], дистанции между словами, средние значения и дисперсии этих дистанций, частоты слов в конкретных разделах текста и т. д.

Подходы к извлечению ключевых слов на основе машинного обучения следует далее разделить на подходы «с учителем» и «без учителя» [2; 3]. Подходы «с учителем», как правило, рассматривают задачу извлечения ключевых слов как задачу бинарной классификации кандидатов в ключевые фразы на основе

выбранных характеристик, в то время как подходы «без учителя» на основе данных характеристик производят взвешивание кандидатов априори и определяют ключевые слова как наиболее весомые среди кандидатов.

В качестве признаков классификации, алгоритмы «с учителем», как правило, используют те или иные варианты рассмотренных выше лингвистических и статистических показателей [6; 7; 9].

Алгоритмы «без учителя» также используют рассмотренные выше показатели при выборе кандидатов и определении их веса [8; 10], однако в силу отсутствия корпусов, предоставляющих возможность обучения, данные алгоритмы ищут способы извлечения дополнительной информации непосредственно из самого текста.

Одним из популярных подходов в алгоритмах «без учителя» стал переход от векторного представления текста, в том числе модели «мешка слов» (bag of words), на графовое представление. Графовая модель решает ряд проблем векторных моделей, связанных с потерей структуры текста и чрезмерной независимостью отдельных слов [13]. В задаче извлечения ключевых слов графовая модель позволяет воспользоваться различными методами определения весов каждой вершины на основе связей с другими вершинами. Например, популярный алгоритм TextRank [14], базирующийся на алгоритме ранжирования веб-страниц PageRank [15], а также его многочисленные варианты (например, SingleRank и ExpandRank [16]) строят граф на основе отношения совместного употребления слов в документе, устанавливая связи между словами, в случае если они находятся в таком отношении, а затем слова и фразы «рекомендуют» друг друга, повышая таким образом свой вес и вес связанных слов. Исследования также показали, что простые меры определения веса, такие как степень (degree) и сила (strength), показывают более высокие результаты, чем сложные меры [17].

Другой подход к извлечению дополнительной информации предлагает перед оценкой весов кандидатов произвести их кластеризацию, основываясь на внешних мерах близости слов, что позволило бы сгруппировать термины, связанные с различными аспектами и темами документа, и заменить их одним термином-«экземпляром» [18].

Вследствие разнообразия и обширности тематик курсов, а также отсутствия крупных корпусов на текущем этапе работы предлагается оценить качество извлечения ключевых слов с помощью алгоритмов «без учителя».

Анализируемые алгоритмы извлечения ключевых слов

В рамках данного исследования были выбраны следующие алгоритмы извлечения ключевых слов: TF-IDF, RAKE, TextRank, графовые алгоритмы на основе степени (degree) и силы (strength).

TF-IDF-взвешивание основывается на оценке веса термина исходя из его частоты встречаемости внутри документа (Term Frequency) и величины, обратной частоте встречаемости термина в других документах коллекции (Inverted Document Frequency):

$$TFIDF\,(word_i) = freq\,(word_i) \cdot idf\,(word_i).$$

Несмотря на кажущуюся простоту данного подхода, исследования показали его эффективность в извлечении ключевых слов, даже в сравнении с другими, более сложными методами.

Взяв всю коллекцию документов, мы взвешиваем каждый термин отдельно для каждого документа, а затем упорядочиваем полученные наборы терминов по убыванию, получая списки слов по направлению убывания их значимости. Поскольку, даже произведя изначальную фильтрацию стоп-слов, мы всё ещё получаем список всех возможно значимых слов в документе, имеет смысл выбрать границу значимости, по которой производится отбрасывание малозначащих для коллекции слов. Практически мы приняли решение опираться только на верхние 5 % слов в каждом документе.

RAKE производит отбор кандидатов в ключевые слова и фразы путём разбиения документа вначале на предложения, а затем на отдельные составляющие предложений исходя из знаков пунктуации. Из полученных выражений фразы-кандидаты выбираются путём разбиения этих выражений на более мелкие сочетания с учётом списка стоп-слов, вероятность нахождения которых в ключевых словах крайне мала. Это делает RAKE более зависимым от конкретного языка анализируемых текстов.

Полученные кандидаты затем оцениваются исходя из весов входящих в них слов, которые определяются, в свою очередь, на основе частот их употребления в документах коллекции и на основе их степеней (degree):

$$Score(Keyword) = \sum_{\substack{word_i \\ \in Keyword}} \frac{\deg(word_i)}{freq(word_i)}.$$

По аналогии с предыдущим подходом мы упорядочиваем взвешенные слова-кандидаты по убыванию их веса и берём верхние 5 % получившегося списка в качестве ключевых слов.

TextRank — это графовый алгоритм, опирающийся на создание графовой модели документа — взвешенного графа, содержащего потенциальных кандидатов в ключевые слова. По аналогии с оригинальным PageRank узлы графа взвешиваются исходя из того, насколько авторитетны связанные с ними узлы. Данный процесс связан с алгоритмами блуждания по графу, постоянно изменяющими веса вершин на основе весов их соседей до тех пор, пока эти плавающие веса не сойдутся.

Другие рассматриваемые нами графовые алгоритмы основываются на более простых мерах веса в графах — степени и силе вершин. Данные алгоритмы также требуют составления взвешенного графа документа, однако веса кандидатов в вершинах рассчитываются исходя из степени вершин — числа соседей — и силы вершин — суммы весов входящих в вершину рёбер. Сами веса определяются с учётом количества совместных употреблений кандидатов в вершинах.

Для графовых подходов мы производим автоматическое склеивание кандидатов исходя из их совместных употреблений в тексте документа. Затем, как и в случае с другими подходами, мы сортируем полученный список кандидатов и отбираем верхние 5 % в качестве ключевых слов.

Эксперимент

Для сравнения нескольких выбранных нами алгоритмов был составлен тестовый корпус, состоящий из выбранных элементов рабочих программ дисциплин по 12 различным дисциплинам трёх областей знаний: информационные технологии, математика и экономика. Для документов данного корпуса экспертами — преподавателями вузов были выбраны наборы ключевых слов, определяющих содержание данных дисциплин. Данные наборы были использованы в качестве контрольной выборки.

Для каждого документа определяются элементы, которые наиболее важны для анализа этих документов: вводных частей, описания требований к результатам освоения и описания структуры курса как перечня входящих в него разделов и тем.

Из выбранных фрагментов текста исключаются стоп-слова — междометия, словосочетания, часто употребляемые союзы и предлоги. Оставшийся текст затем очищается от нетекстовых символов, разбивается на слова, которые затем проходят лемматизацию.

Затем для каждого из полученных документов применяются выбранные алгоритмы. Следует отметить, что на практике было принято решение удалить из кандидатов для всех алгоритмов кроме TF-IDF те слова, которые имели слишком низкий вес, избавляясь таким образом от общей лексики, встречающейся во всех документах, а также сокращая общее количество генерируемых ключевых слов приблизительно вдвое. На практике эксперимент проводится с 90 %-м порогом — оставляются только верхние 10 % слов, отранжированных по их TF-IDF-весу.

Таким образом, для каждого документа в коллекции мы получаем взвешенный набор его лемматизированных ключевых слов для всех пяти алгоритмов.

Оценка качества алгоритмов

Для оценки качества в задачах выбора ключевых слов принято использовать метрики точности (precision) и полноты (recall). Традиционно попаданием или достоверным положительным результатом считается наличие ключевого слова из результатов работы оцениваемого алгоритма в наборе ключевых слов, составленных экспертами. Иногда также рассматриваются попадания ключевого слова в аннотацию, составленную экспертом.

На практике в данном эксперименте можно прийти к выводу, что экспертная оценка образовательного контента, как правило, приводит к набору укрупнённых концептов того, что упоминается в документе, в то время как автоматическое извлечение, по меньшей мере в выбранных нами алгоритмах, опирается на конкретные упоминания в тексте документа, не укрупняя их до более общих понятий, а также зачастую, состоит из менее корректно построенных фраз, чем те, что были выбраны экспертом.

Это наблюдение позволяет отметить, что традиционное сопоставление результатов работы алгоритмов и оценки человеком в данном эксперименте трудновыполнимо из-за необходимости сопоставления каждой экспертной фразы с каждой сгенерированной фразой в попытке проследить перекрёстные упоминания терминов. Однако, даже проведя такой анализ, можно столкнуться с проблемой семантического разрыва — выбранные экспертами ключевые фразы могут содержать крайне малое число общих слов с автоматически сгенерированными фразами или не иметь их вовсе.

Примером такой ситуации могут являться следующие ключевые фразы для документа, посвящённого программе курса по базам данных:

– совместное использование данных;
– коллективный доступ;
– разграничение доступа.

Первая из этих ключевых фраз была предложена экспертом, две другие выбраны соответственно алгоритмами TF-IDF и TextRank. И хотя среди сгенерированных фраз не содержится ни одного слова, встречающегося во фразе, выбранной экспертом, они так или иначе имеют один и тот же смысл.

Если проигнорировать данный фактор, а также существенно снизить планку сложности, считая успешным наличие хотя бы одного общего слова между экспертной и сгенерированной фразами, то выбранные алгоритмы показывают схожие результаты с другими работами, оценивающими их качество. Грубо оценённые таким образом алгоритмы показывают достаточно близкие средние значения точности и полноты: $P \approx 0{,}35$ и $R \approx 0{,}4$. Чуть более высокую полноту и точность достигает TF-IDF: $P \approx 0{,}38$ при $R \approx 0{,}43$. RAKE показывает более высокую точность, но теряет в полноте: $P \approx 0{,}5$ при $R \approx 0{,}33$.

Тем не менее такая оценка качества работы алгоритма не вполне корректна, в связи с чем был применён альтернативный подход с использованием векторных представлений слов. Пользуясь моделью языка, мы можем оценить качество как близость между двумя наборами ключевых слов — экспертным

и сгенерированным. Модель Word2vec [19], обученная на корпусе русскоязычной Wikipedia, даёт достаточно большой охват терминологии, чтобы приближённо считать эту векторную модель моделью языка. Таким образом, можно вычислить вектора всех ключевых слов в обоих наборах, объединить их в общий для набора вектор и сравнить их через косинусную меру близости. Этот подход был предложен авторами genism framework [20].

Введём следующие обозначения:

$$\overrightarrow{L2_w} = \frac{\overrightarrow{V_w}}{|\overrightarrow{V_w}|},$$

где $L2$ — норма word2vec вектора слова w;

$$\overrightarrow{m_{ws}} = \frac{\sum_{w_i \in ws} \overrightarrow{L2_{w_i}}}{|ws|},$$

где $L2$ — норма word2vec векторов всех слов в наборе ws.

Близость между наборами ws_1 и ws_2 — это косинус между средними слов в наборах:

$$sim_{wordsets}(ws_1, ws_2) = \cos(\overrightarrow{m_{ws_1}}, \overrightarrow{m_{ws_2}}).$$

Таким образом, близость между документами d_i и d_j может быть вычислена как близость между наборами входящих в них ключевых слов:

$$sim(d_i, d_j) = sim_{wordsets}(Keywords(d_i), Keywords(d_j)).$$

В результате такого подхода к оценке получаются следующие усреднённые по всему корпусу близости наборов ключевых слов:

Алгоритм	$sim_{wordsets}$
TF-IDF	0,96974
TextRank	0,96608
Graph (Strength)	**0,97381**
Graph (Degree)	0,96904
RAKE	0,96478

Полученные данные показывают сравнительную близость данных алгоритмов, которая также отражается в близости самих сгенерированных наборов ключевых слов. Стоит заметить, что использование модели word2vec для оценки близости может потенциально вносить дополнительный элемент погрешности, однако он значительно меньше ошибок, получаемых за счёт допущений в альтернативных подходах к оценке качества.

Заключение

В рамках данного исследования мы рассмотрели укрупнённую структуру рабочих программ дисциплин и выявили наиболее информативные их элементы. Был проведён обзор существующих популярных методов извлечения ключевых слов и осуществлён эксперимент с оценкой качества методов, не требующих тренировочной выборки.

Для оценки качества работы данных методов в рамках рассматриваемой задачи был подготовлен тестовый корпус различных рабочих программ дисциплин и реализован ряд алгоритмов извлечения ключевых слов, результаты работы которых затем были сравнены с наборами слов, выбранными экспертами. Результаты показали близость качества работы выбранных алгоритмов.

Следующими основными шагами исследования являются:

– использование полученных данных для улучшения качества сравнения документов;

– оценка эффективности данных алгоритмов для других видов образовательного контента;

– интеграция структурных мер оценки близости контента с мерами на базе ключевых слов;

– применение методов вероятностного тематического моделирования к задачам анализа образовательного контента.

Библиографический список

1. Ботов, Д. С. Подход к определению меры семантической близости результатов обучения / Д. С. Ботов, Ю. Д. Кленин // Информационные технологии и системы : тр. Пятой Междунар. науч. конф., Банное, Россия, 24–28 февр. 2016 г. Челябинск : Изд-во Челяб. гос. ун-та, 2016. С. 76–82.

2. Sifatullah, S. Keyword and Keyphrase Extraction Techniques: A Literature Review / S. Sifatullah, S. Aditi // Intern. J. of Computer Applications. 2015. Vol. 109, № 2. P. 18–23.

3. Beliga, S. Keyword extraction: a review of methods and approaches / S. Beliga / University of Rijeka, Department of Informatics. Rijeka, 2014.

4. Pouliquen, B. Automatic annotation of multilingual text collections with a conceptual thesaurus / B. Pouliquen, R. Steinberger, C. Ignat // Ontologies and Information Extraction : workshop at EUROLAN'2003: The Semantic Web and Language Technology — Its Potential and Practicalities. Bucharest, 28 July — 8 August 2003. arXiv preprint cs/0609059 (2006.9.12).

5. Medelyan, O. Thesaurus based automatic keyphrase indexing / O. Medelyan, I. H. Witten // Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries. New York : ACM Press. 2006. P. 296–297.

6. Hulth, A. Improved automatic keyword extraction given more linguistic knowledge / A. Hulth // Proceedings of the 2003 conference on Empirical methods in natural language processing / Association for Computational Linguistics. Morristown (USA), 2003. P. 216–223.

7. Hong, B. An Extended Keyword Extraction Method / B. Hong, D. Zhen // International Conference on Applied Physics and Industrial Engineering, Physics Procedia. 2012. Vol. 24, part B. P. 1120–1127.

8. Pudota, N. A New Domain Independent Keyphrase Extraction System / N. Pudota, A. Dattolo, A. Baruzzo, C. Tasso // CCIS. 2010. Vol. 91. P. 67–78.

9. Nguyen, T. D. Keyphrase extraction in scientific publications / T. D. Nguyen, M. Y. Kan [et al.] // ICADL. LNCS. 2007. Vol. 4822. P. 317–326.

10. Rose, S. Automatic keyword extraction from individual documents / S. Rose, D. Engel, N. Cramer, W. Cowley // Text Mining: Applications and Theory. Hoboken : John Wiley & Sons, 2010. P. 1–20.

11. Hasan, K. Conundrums in unsupervised keyphrase extraction: Making sense of the state-of-the-art / K. Hasan, V. Ng // Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Beijing, Chinaб August 23–27, 2010. Stroudsburg (USA), 2010. P. 365–373.

12. Lahiri, S. Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks / S. Lahiri, S. R. Choudhury, C. Caragea // arXiv preprint arXiv:1401.6571 (2014)

13. Sonawane, S. S. Graph based Representation and Analysis of Text Document: A Survey of Techniques / S. S. Sonawane, P. A. Kulkarni // Int. Jour. Of Computer Applications. 2014. Vol. 96, № 19. P. 1–8.

14. Mihalcea, R. TextRank: Bringing order into texts / R. Mihalcea, P. Tarau // Empirical Methods in Natural Language : Processing. Barcelona, 2004. P. 404–411.

15. Brin S. The anatomy of a large-scale hypertextual Web search engine / S. Brin, L. Page // Computer Networks and ISDN Systems. 1998. № 30. P. 1–7.

16. Wan, X. Single document keyphrase extraction using neighborhood knowledge / X. Wan, J. Xiao // Proceedings of the 23rd AAAI Conference on Artificial Intelligence. Chicago (Illinois), 2008. P. 855–860.

17. Lahiri, S. Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks / S. Lahiri, S. R. Choudhury, C. Caragea // arXiv preprint arXiv:1401.6571 (2014).

18. Liu, Zh. Clustering to find exemplar terms for keyphrase extraction / Zh. Liu, Li Peng, Yabin Zheng, Maosong Sun // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. USA, 2009. P. 257–266.

19. Mikolov, T. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean // Advances in neural information processing systems 26. 2013. P. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

20. Řehůřek, R. Software Framework for Topic Modelling with Large Corpora / R. Řehůřek, P. Sojka // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Malta, 2010. P. 45–50.

Сведения об авторах

Ботов Дмитрий Сергеевич — старший преподаватель кафедры информационных технологий и экономической информатики Института информационных технологий, Челябинский государственный университет, Челябинск. dmbotov@gmail.com

Кленин Юлий Дмитриевич — магистрант Института информационных технологий, Челябинский государственный университет, Челябинск. jklen@yandex.ru

---

D. S. Botov, J. D. Klenin

APPLYING KEYWORD EXTRACTION ALGORITHMS AND VECTOR REPRESENTATION OF WORDS IN THE PROBLEM OF EDUCATIONAL CONTENT MINING

The paper presents an overview of keyword extraction algorithms in applied to educational data mining (course description analysis task specifically) and provides a general comparison of algorithms. In the experiment for evaluating of keyword extraction algorithms, it proposed an approach using a model of the vector representation of words (word2vec), solves the problem of the semantic gap by comparing the results of the algorithms with the test set composed by experts.