УДК 004.65:378.4

НОРМАЛИЗАЦИЯ СЛОВОФОРМ ПРИ АНАЛИЗЕ РЕПОЗИТОРИЯ УНИВЕРСИТЕТА С ИСПОЛЬЗОВАНИЕМ ГРАФОВОЙ БАЗЫ ДАННЫХ

В.А. Прытков
Проректор по учебной работе БГУИР, кандидат технических наук, доцент

И.И. Пилецкий
Доцент кафедры информатики БГУИР, кандидат физико-математических наук, доцент

Н.А. Волорова
Заведующая кафедрой информатики БГУИР, кандидат технических наук, доцент

Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», Республика Беларусь
E-mail: prytkov@bsuir.by

В.А. Прытков
Кандидат технических наук, доцент, проректор по учебной работе БГУИР. Автор 38 научных и методических публикаций, являлся исполнителем, ответственным исполнителем или руководителем 12 научно-исследовательских проектов, член организационного комитета ряда международных научно-практических конференций. Область научных интересов: обработка изображений и текстурный анализ, синтаксические методы обработки информации, анализ слабоструктурированных данных.

И.И. Пилецкий
Кандидат физико-математических наук, доцент, доцент кафедры информатики БГУИР, научный руководитель совместной лаборатории БГУИР — IBA и Академического центра компетенций технологий IBM. Имеет большой опыт в реализации и разработке промышленных решений в ИТ-области, являлся ведущим разработчиком, системным архитектором, руководителем и научным руководителем нескольких десятков крупных проектов, связанных с разработкой программного обеспечения и баз данных корпоративного уровня. Автор более 80 научных публикаций (в том числе монографий и учебных пособий) в области моделирования, технологии разработки программного обеспечения и баз данных. Член редакционной коллегии журнала «Baltic Journal of Modern Computing». Область научных интересов: технологии и аналитические комплексы анализа Big Data, NPL и ML алгоритмы.

Н.А. Волорова
Кандидат технических наук, доцент, заведующая кафедрой информатики БГУИР. Автор более 100 научных и методических публикаций, 19 изобретений, принимала участие в 20 научно-исследовательских работах в качестве исполнителя, ответственного исполнителя или руководителя. Член организационного комитета ряда международных научно-практических конференций. Область научных интересов: имитационные модели сложных систем, системы автоматизации моделирования.

Аннотация. При обработке текстов на естественном языке для широкого класса прикладных задач возникает потребность в построении словаря терминов. Его наличие позволяет определять предметные области текстов, упрощает поиск и систематизацию информации. При этом возникает задача сведения различных словоформ термина к единой форме. В работе рассматривается подход к нормализации словоформ на примере построения системы анализа университетского репозитория БГУИР с использованием графовой базы данных Neo4j. Данный алгоритм опирается на эвристический алгоритм стемминга, однако для удобства отображения информации основа приводится к лемме путем использования статистической информации.

Ключевые слова: стемминг, лемматизация, обработка естественного языка, большие данные, репозиторий, автор, публикация, словоформа, NoSQL, графовые базы данных, Neo4j.

Введение. Алгоритмы обработки текстов на естественном языке на сегодняшний день становятся все более и более востребованными. Связано это, в первую очередь, с необходимостью анализа слабоструктурированной и неструктурированной текстовой информации, объемы которой постоянно и нелинейно возрастают – это и новостные ленты, и блоги, и анонсы технических новинок и мероприятий от различных организаций, и обычные страницы веб-сайтов, и тексты научных и публицистических статей, размещенных онлайн. Для решения прикладных задач эта информация нуждается в определенной систематизации и классификации. Составление соответствующих словарей, деревьев поиска, графов знаний являются нетривиальными задачами. Зачастую исследователи идут эмпирическим путем, определяя исходные классы или вершины дерева либо узлы графа, представляющие собой базовые, наиболее общие понятия и области знаний. Дальнейшая иерархия структур выстраивается на основе анализа текстов, но и здесь широко используются эвристические алгоритмы.

Одной из частных задач при построении таких словарей и деревьев является нормализация (канонизация) – приведение различных форм к единой нормальной (канонической) форме. Применительно к текстам – это сведение различных словоформ термина к единой форме. Как правило, сначала выполняется определенная фильтрация: удаляются малозначимые символы (дефисы, кавычки, апострофы и т. д.), символы преобразуются к одному регистру и пр. Далее, в зависимости от решаемой задачи, используются стемминг либо лемматизация. В случае стемминга для исходного слова определяется его основа, при этом алгоритм стемминга может выполнить усечение словоформы до подстроки, не обязательно совпадающей с морфологической основой. В случае лемматизации словоформы приводятся к нормальной (словарной) форме (лемме). По сравнению со стеммингом лемматизация более сложный процесс, при котором в большей степени используется морфологический анализ слов, при этом удаляются, как правило, только флективные окончания. Алгоритмы стемминга и лемматизации могут ошибочно относить к одной основе (лемме) различные слова (ошибка первого рода, overstemming), так же как и поставить в соответствие разные основы (леммы) морфологическим формам одного слова (ошибка второго рода, understemming).

Алгоритмы стемминга и лемматизации могут основываться на стохастическим или статистическом подходе, используют как поиск на основе имеющегося словаря (таблицы) соответствий, так и усечение аффиксов, как правило, окончаний. Примером одной из ранних работ в этой области является [1]. Более поздним классическим алгоритмом, используемым и на сегодняшний день, считается алгоритм Портера [2]. Способы оценки алгоритмов стемминга и сравнительный анализ нескольких классических алгоритмов приводится в работе [3], также сравнению алгоритмов стемминга посвящена работа [4]. Для русского языка, помимо алгоритмов на основе алгоритма Портера, широко применяются алгоритмы Stemka Андрея Коваленко и используемый Яндексом MyStem Ильи Сегаловича [5]. Сравнение лексического и алгоритмического подхода к стеммингу для ряда европейских языков приведено в [6, 7]. В работе [8] исследуется стемминг для восточноевропейских языков в применении к поисковым алгоритмам. Качество поиска с использованием общедоступных стеммеров для русского языка обсуждается и в [9], а в [10] рассматриваются вопросы лемматизации.

При построении словарей в аналитических системах в качестве термина могут выступать не только отдельные слова, но и терминологические словосочетания. В этом случае нормализация не сводятся к нормализации отдельных слов в словосочетании, т.к. слова в

словосочетании должны быть согласованы и поэтому не всегда будут соответствовать словарной форме.

В данной работе рассматривается алгоритм нормализации, использованный при анализе публикаций, размещенных в репозитории Белорусского государственного университета информатики и радиоэлектроники. Данный алгоритм опирается на эвристический алгоритм стемминга, однако для удобства отображения информации основа приводится к словарной форме, которая определяется с использованием статистической информации в отличие от классического подхода с опорой на морфологический анализ.

Модель и алгоритм нормализации. В основу построения системы положена достаточно простая модель данных, исследованная нами ранее [11]. Эта модель представлена на рис. 1. Модель очень проста, она включает всего три вида сущностей (вершин): публикации, авторы, темы (ключевые слова), и два вида связей: между публикацией и автором (связь авторства), между публикацией и ключевым словом (тематика публикации).

На основе этой первичной модели можно построить расширенную модель (рис. 2), которая будет включать связи соавторства, экспертизы (автора можно считать экспертом в некоторой предметной области, если она описывается в его публикации), а также близости ключевых слов (если они употребляются совместно, т. е. данные предметные области рассматриваются в одной публикации). Это вторичные связи, т. к. могут быть вычислены на основе базовой модели, они изображены на рисунке штриховыми линиями. Эти связи дополнительно могут включать свойства, определяющие силу связи, например, количество совместных работ для связи соавторства, или количество публикаций данного автора по данной предметной области для связи экспертизы.

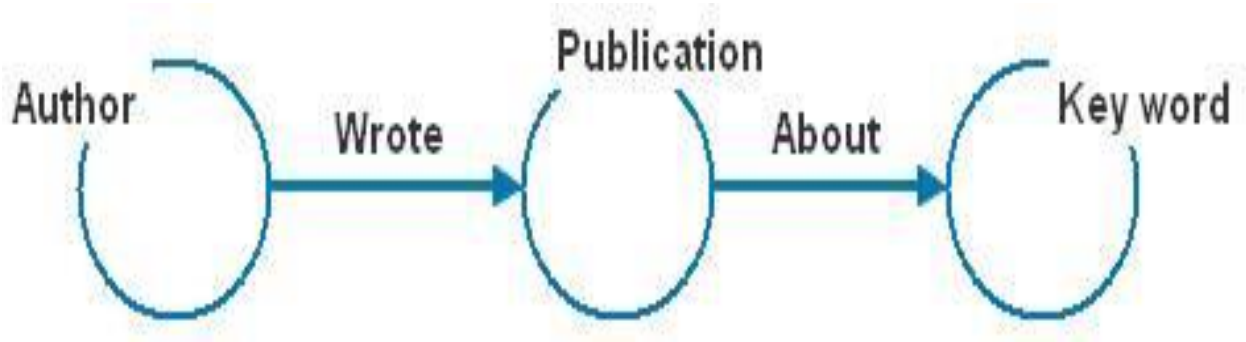

Рисунок 1. Базовая модель

Однако эта модель никак не позволяет учитывать в качестве одного и того же термина или терминологического словосочетания различные его словоформы. Так, например, для словосочетания «онлайн режим» в качестве ключевых слов в публикации может употребляться и «онлайн режимы», а в названии, аннотации и в самом тексте могут встречаться словоформы «онлайн режиме», «онлайн режимами» и т. д., а также различные варианты написания одной и той же словоформы, например, «онлайн-режим», «Онлайн Режим» и др.

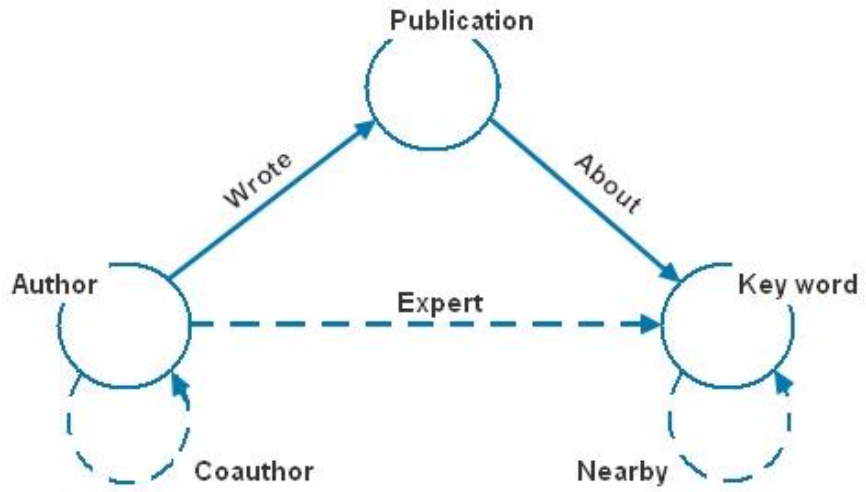

Рисунок 2. Расширенная модель

Учитывая относительно небольшой размер исследуемого репозитория (несколько десятков тысяч публикаций), нами использовался эвристический алгоритм фильтрации и стемминга, который включает в себя:

−преобразование символов к нижнему регистру;

−преобразование, за малым исключением, всех символов, не являющихся буквами либо цифрами, в пробелы;

−удаление предлогов, союзов и частиц малой длины;

−удаление подряд идущих пробелов после всех замен;

−если язык первого или первых двух символов слова отличается от остальных, при этом эти символы имеют одинаковое написание на русском и английском языке, то они заменяются на действительный язык слова;

−для слов на русском языке отсекаются наиболее употребимые окончания, а также постфиксы -ся и -сь.

Если первые четыре элемента достаточно тривиальны, то два оставшихся требуют более детального пояснения. В русскоязычных публикациях довольно часто возникает необходимость употребления отдельных англоязычных терминов, в силу чего автор может переключить раскладку клавиатуры, но забыть сразу после написания такого термина вернуть ее обратно, особенно если первые символы следующего слова по написанию похожи на символы русской раскладки. В итоге в тексте появляются слова, которые корректно читаются, поскольку начертание символов выполнено без ошибок, но некорректно распознаются алгоритмами, поскольку содержат символы другого алфавита. Во многих алгоритмах для избегания этого все символы, имеющие одинаковое начертание в русском и английском алфавите, заменяются на символы одного языка. Однако у такого подхода имеются и серьезные ограничения: он нарушает порядок сортировки терминов, а также не позволяет корректно сравнивать термины с наборами из других словарей, так как в них такой подход может и не использоваться. Поэтому нами был предложен вариант с заменой только первого (или первых двух) символов, если они отличаются от языка оставшихся символов слова. Ограничением такого подхода является его ориентированность на добросовестность пользователей: он не учитывает, что при работе различных программ для рерайтинга такие символы заменяются в произвольных местах слова.

Наиболее употребимыми окончаниями в русском языке являются: -а, -я, -и, -ы, -о, -е, -у, -ю, -ь, -м, -ая, -яя, -ой, -ей, -её, -ёй, -ый, -ий, -ам, -ям, -им, -ими, -ами, -ями, -ах, -ях, -их, -ию, -ия, -ов, -ев, -ёв, -ом, -ому, -его, -ему, -ем, -еми, -емя, -ех, -ёх, -ею, -ём, -ое, -ою, -оё, -ее, -ие, -ые, -ми, -мя, -ал, -ала, -али, -ял, -яла, -яли, -ол, -ола, -оли, -ел, -ела, -ели, -ул, -ула, -

ули, -ет, -ёт, -еть, -ете, -ёте, -ат, -ать, -ят, -ять, -оть, -ут, -уть, -ют, -ыть, -ит, -ить, -ите, -ть, -ти, -ишь, -ешь, -ёшь, -ого, -ум, -умя, -ух, -ую, -шь. Однако обнаружение соответствующей последовательности символов в конце слова не гарантирует, что после ее отсечения останется основа слова. Так, в неизменяемых словах, например, наречиях, может быть отсечена часть основы: «далеко» усечется до «далек». Это, однако, типичная проблема для алгоритмов стемминга. Вторая проблема — в результате такого отсечения может остаться всего один-два символа, например, «сеть» превратится в «с». Эта проблема также связана с морфологией — в данном примере слово «сеть» является существительным, но окончание -еть характерно для глаголов. Наш подход, несмотря на то, что морфологический анализ не выполняется, частично устраняет данную проблему. Она сильнее всего проявляется в коротких словах, поэтому наш алгоритм усекает слова только до определенной длины. Наш вариант уменьшает ошибку первого рода, но при этом может увеличить ошибку второго рода. В данном случае мы считаем, что объединить различные по семантике слова хуже, чем оставить некоторые словоформы не объединенными.

Таким образом, для каждого узла, соответствующего термину или словосочетанию, помимо основного свойства «Value» определяется дополнительное свойство «Normalized Value». Теперь можно включить в модель графа сущность, представляющую нормализованные значения терминологических словосочетаний, с соответствующей связью. Аналогично можно поступить и для узлов авторов, с учетом того, что нет необходимости в усечении окончаний. Данная модель представлена на рис.3. Также стоит обратить внимание, что вторичные связи соавторства, экспертизы и близости ключевых слов будут теперь соответствовать узлам нормализованных значений имен авторов и терминологических словосочетаний.

Фактически теперь все запросы пользователей по анализу информации будут проводиться по этим нормализованным узлам. При пополнении графа будут добавляться вершины не нормализованных имен авторов и терминологических словосочетаний. В случае, если соответствующее новому узлу нормализованное значение еще не представлено в графе, оно добавляется в новую вершину, если же оно уже имеется, то соответствующая связь типа «нормализовано» устанавливается к ней.

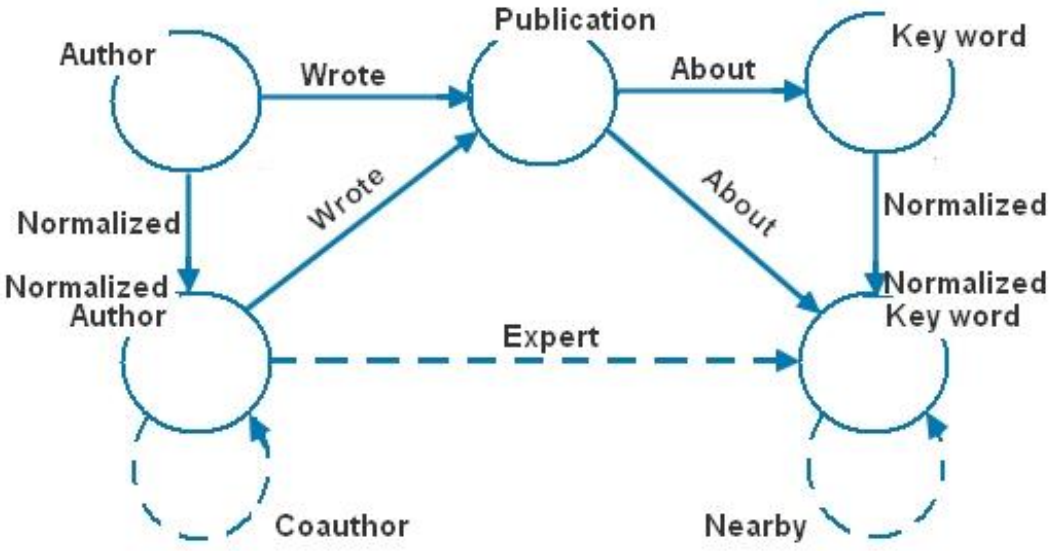

Рисунок 3. Модель с нормализованными сущностями

Однако нормализованные значения, полученные путем подобного стемминга, не удобны для визуального восприятия пользователем. Возникает очевидная мысль: почему бы вместо анализа морфем не воспользоваться уже собранной в базе статистикой? Выдвинем гипотезу: наиболее часто используемая (т. е. фактически общепризнанная) форма написания термина (словосочетания) является грамматически и синтаксически корректной. Опираясь на эту гипотезу, алгоритм выполняет следующее: для всех вершин, имеющих одинаковое нормализованное представление, в качестве леммы выбирается наиболее часто используемое исходное значение. Так, например, если для нормы «баз данн» исходная форма «база данных» использовалась в публикациях чаще, чем «базы данных», то в качестве леммы будет выбрана именно она. Эта статистика элементарно рассчитывается по количеству связей вида «About» для терминов и «Wrote» для авторов. Полученную лемму удобно хранить в виде отдельного свойства для соответствующих узлов «Normalized Author» и «Normalized Key word».

При добавлении новых вершин к уже имеющемуся графу нет необходимости постоянно переопределять леммы – только для случая, когда узла с соответствующим нормализованным значением в графе еще не было. Для остальных узлов достаточно определять их при накоплении некоторой «критической массы» новых узлов, например, 5 или 10% от общего количества.

Особенностью модели является то, что она позволяет изменять в дальнейшем используемый алгоритм нормализации. В этом случае из графа необходимо удалить все вершины с нормализованными значениями с соответствующими им связями, и выполнить переопределение основ, дополнить граф новыми вершинами для обнаруженных основ, и пересчитать для них леммы. Цена за такую гибкость — практически удвоение количества вершин для авторов и терминов, так же, как и соответствующих связей, которые в любом случае сохраняют первичную, «сырую» информацию.

Помимо рассмотренных выше ограничений, модель имеет еще одно: при расчете статистик для определения леммы возможно использовать только те терминологические словосочетания, которые указаны в качестве ключевых слов, а не определенных на основе парсинга названия, аннотации или текста статьи, поскольку в этом случае наиболее употребимой может оказаться абсолютно любая словоформа, в т.ч. в произвольном падеже и времени.

Кроме того, модель не учитывает синонимы, так же, как и представление термина на разных языках, а также перестановки слов внутри словосочетаний. Не учитывается и возможное дублирование одного и того же материала. Устранение таких ограничений требует семантического анализа и в общем случае является самостоятельной и далеко не тривиальной задачей.

Созданию и развитию аналитических систем, включающих автоматический сбор информации из веб-источников, а также анализ и обнаружение взаимосвязей между авторами и публикациями, с последующим построением соответствующего графа отношений, в настоящее время уделяется все более пристальное внимание. Примером работ по этой проблематике являются [12, 13]. В работе [14] описывается модель, адаптированная для использования ряда NLP-алгоритмов. Различным аспектам вопроса добавления ключевых слов (терминов) в граф посвящены работы [15-18].

Эксперимент. Репозиторий БГУИР расположен по следующему адресу https://libeldoc.bsuir.by/, причем разметки страниц публикаций содержат теги метаописания для поисковых систем, которые включают авторов (отдельный тег для каждого), ключевые слова (отдельный тег для каждого), название, полное библиографическое описание, дату публикации, ссылку на саму страницу, ссылку на страницу с публикацией в виде pdf файла, аннотацию, издателя и др. Такая структура существенно упрощает парсинг.

Модель хорошо описывается графом, соответственно, наиболее подходящим решением для работы с такими данными будут являться графовые базы данных [19, 20]. Из доступных решений единственный наиболее подходящий вариант – Neo4j, которая не только является свободно распространяемой, но и достаточно хорошо распространена. Так, по данным портала DB-Engines [21], на январь 2019 года Neo4j занимает 22 место по популярности из 343 СУБД, включая реляционные, и является лучшей среди графовых СУБД. Кроме того, Neo4j использует графовую модель со свойствами и метками, что прекрасно сочетается с требованиями к структуре данных, предъявляемыми описанным алгоритмом. Хорошими примерами работ, рассматривающих прикладные аспекты использования Neo4j совместно с NLP-библиотеками, являются [22, 23].

В результате построения графа на основе данных репозитория по состоянию на январь 2019 года нами был получен граф, содержащий 20092 публикации, 16239 авторов (при общем количестве словоформ 16550), 30885 тематик работ (при общем количестве словоформ 34553), 47080 связей авторства, 80411 связей между публикациями и тематиками. Пример случайной выборки из двух публикаций, с соотносящимися с ними авторами и тематиками, представлен на рис. 4. По сравнению с январем 2018 года количество публикаций увеличилось на 4677 (30,3%), авторов (включая словоформы) — на 3417 (26,0%), терминов (тематик)? включая словоформы — на 6612 (23,7%), связей авторства — на 11615 (32,8%), связей между публикациями и тематиками — на 17512 (27,8%). Всего полученный граф содержал более 118 тысяч узлов и более 650 тысяч связей.

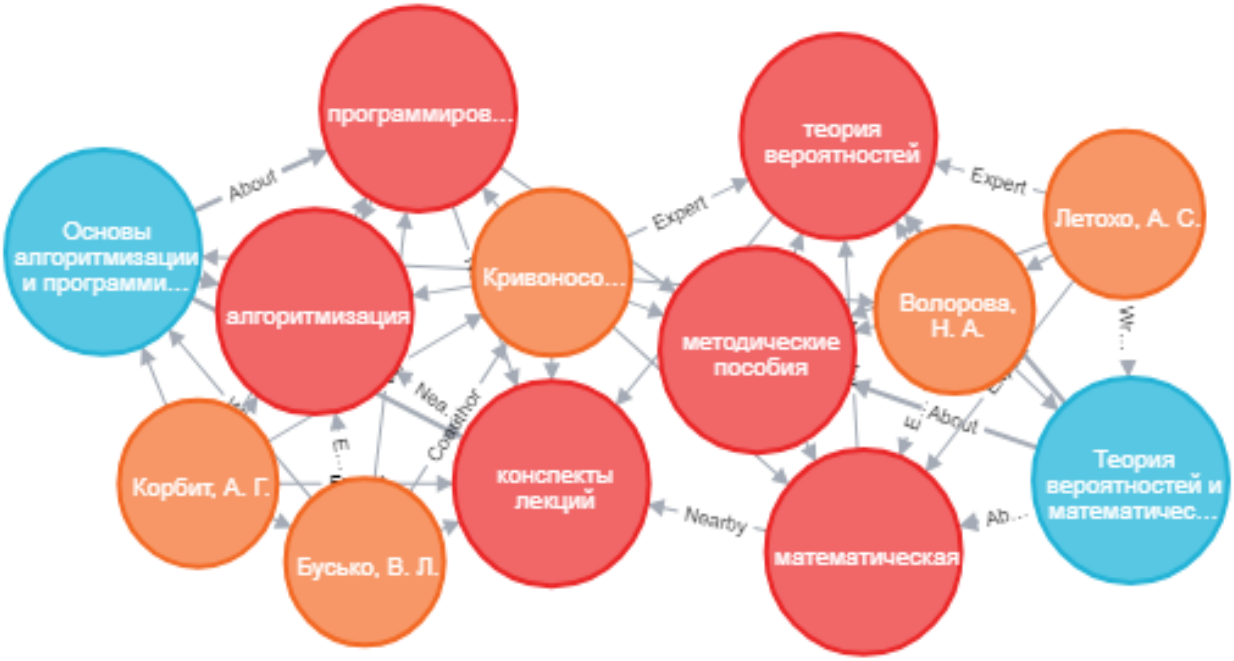

Рисунок 4. Пример случайной выборки

Ниже представлен пример результатов стемминга для отдельных авторов и терминов. В каждой строке представлены: полученная в результате обработки рассмотренным алгоритмом основа, количество записей исходного графа с данной словоформой, и сама соответствующая словоформа. Для удобства восприятия данные разделяются двоеточиями, также пробельные символы отображаются в виде центральной точки:

```
бурак·и·а:·12:·Бурак,·И.·А.
бурак·и·а:·1:·Бурак,·И.А
бурак·и·а:·1:·Бурак,·И.А.
молочко·а·п:·14:·Молочко,·А.·П.
молочко·а·п:·1:·Молочко,·А.П.
молочко·а·п:·1:·Молочко,·А.·П,
петровский·а·а:·56:·Петровский,·А.·А.
петровский·а·а:·2:·Петровский,·А.··А.
петровский·а·а:·1:·Петровский,·А.·А..
майсеня·л·и:·26:·Майсеня,·Л.·И.
майсеня·л·и:·1:·Майсеня,·Л.·И..
майсеня·л·и:·1:·Майсеня,·л.·и.
ташлыкова·бушкевич·и·и:·25:·Ташлыкова-Бушкевич,·И.·И.
ташлыкова·бушкевич·и·и:·1:·Ташлыкова·-·Бушкевич,·И.·И.
ташлыкова·бушкевич·и·и:·1:·Ташлыкова·-·Бушкевич,·И.·И.

информационн·технолог:·1:·Информационные·технологии
информационн·технолог:·1:·информационные·технологии
информационн·технолог:·3:·информационная··технология
информационн·технолог:·30:·информационные··технологии
информационн·технолог:·185:·информационные·технологии
информационн·технолог:·1:·информационные·технологий
информационн·технолог:·1:·информационная·технология
web·приложен:·6:·Web-приложения
web·приложен:·1:·WEB··приложение
web·приложен:·2:·web·-·приложение
web·приложен:·14:·web-приложение
web·приложен:·1:·web·приложения
web·приложен:·22:·web-приложения
скрыт·марковск·модел:·1:·скрытая··марковская··модель
скрыт·марковск·модел:·1:·скрытые··марковские··модели
скрыт·марковск·модел:·1:·скрытая·Марковская·модель
скрыт·марковск·модел:·1:·скрытые·Марковские·модели
скрыт·марковск·модел:·1:·скрытые·марковские·модели
скрыт·марковск·модел:·1:·Скрытые·Марковские·модели
радиотехническ·сист:·1:·радиотехнический·системы
радиотехническ·сист:·1:·радиотехническая·системы
радиотехническ·сист:·6:·радиотехнические·системы
радиотехническ·сист:·1:·радиотехническая··система
радиотехническ·сист:·2:·радиотехнические··системы
радиотехническ·сист:·5:·радиотехническая·система
big·data:·1:·Big··Data
big·data:·1:·BIG·DATA
big·data:·1:·big··data
big·data:·5:·big·data
big·data:·1:·Big·data
big·data:·28:·Big·Data
```

На рисунках 5 и 6 представлены примеры выборок авторов и терминов соответственно, в которых различным словоформам сопоставлена общеупотребимая лемма:

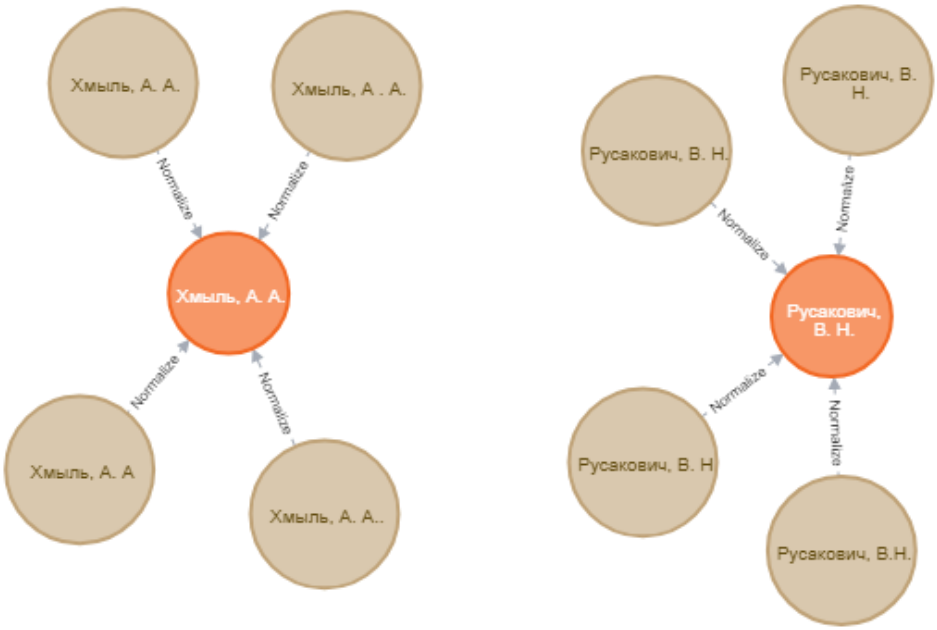

Рисунок 5. Пример выборки авторов и соответствующие им леммы

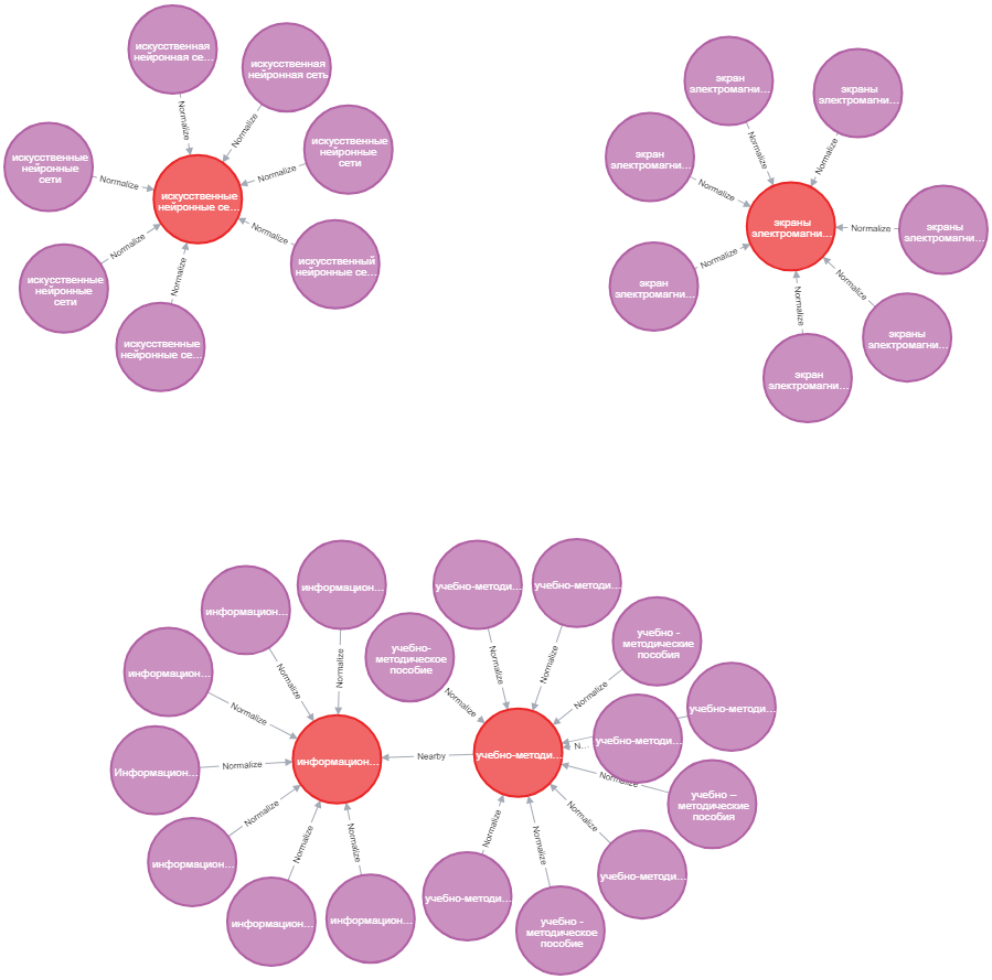

Рисунок 6. Пример выборки терминов и соответствующие им леммы

Так, у авторов широко распространены лишние (или, напротив, отсутствующие) пробелы и точки, используются запятые вместо точек, строчные символы инициалов, тире вместо дефиса (символы имеют разные коды). Легко заметить, что подавляющее большинство словоформ имеют корректные написания, а ошибочные написания связаны, судя по их характеру, с ошибками оператора при вводе данных. Окончания для авторов не отсекаются.

У терминов также присутствуют подобные различия в написании, помимо этого, добавляются собственно разные словоформы как таковые (особенно часто используется как единственное, так множественное число), а также явные синтаксические ошибки, когда слова в словосочетании не согласованы по форме. Иногда практически равноупотребимы несколько словоформ. С прикладной точки зрения не принципиально, какая из них будет использована в качестве леммы.

Видно, что как в случае с авторами, так и в случае с терминами, алгоритм в качестве леммы выбирает грамматически корректную форму, эмпирически подтверждая выдвинутую ранее гипотезу.

Таким образом, предлагаемый алгоритм на имеющемся наборе данных из репозитория БГУИР по отношению к случаю, когда нормализация не используется, снижает ошибку второго рода (отнесение к разным основам вариантов написания либо словоформ одного и того же слова) для авторов в среднем на 1,9%, для терминов – на 11,9%. С высокой вероятностью можно ожидать близких результатов и для иных наборов данных аналогичного типа. Ошибка первого рода (отнесение к одной основе словоформ разных слов) не вычислялась, так как это требует практически ручного анализа полученного результата.

Заключение. Разработан простой алгоритм нормализации, адаптированный к графовой модели представления данных. Алгоритм последовательно включает в себя стадию фильтрации и стемминга, и стадию лемматизации.

Фильтрация и стемминг выполняются эвристическим алгоритмом, отличающимся от типовых тем, что замена символов, имеющих одинаковое написание в различных языках, производится только для первых символов слова, что позволяет сохранить порядок сортировки, а также тем, что усечение производится до определенного порога, что позволяет уменьшить ошибку второго рода для коротких слов.

Лемматизация полученных основ выполняется на базе имеющейся в построенном графе статистики употребления тех или иных словоформ, что позволяет отображать результаты анализа в удобном для пользователя виде.

Модель позволяет не переопределять лемму каждый раз при внесении в граф новой вершины – достаточно делать это с некоторой периодичностью. Кроме того, модель позволяет в дальнейшем изменять используемый алгоритм нормализации, однако ценой за такую гибкость является практически двукратное увеличение числа вершин и связей графа.

На наборе данных из репозитория БГУИР алгоритм снижает ошибку второго рода по отношению к случаю, когда нормализация не используется, для авторов в среднем на 1,9%, для терминов – на 11,9%.

Основными ограничениями модели являются:

1. алгоритм стемминга не учитывает особенности работы программ для рерайтинга;
2. стемминг не гарантирует усечения именно до морфологической основы;
3. даже ограниченно не используется морфологический анализ;
4. алгоритм лемматизации допускает использовать только терминологические словосочетания из числа указанных в ключевых словах;
5. не учитываются синонимы, омонимы, представление терминов на различных языках, перестановки слов внутри словосочетаний.

Устранение некоторых из этих ограничений возможно при наличии словаря, который может строиться по мере роста графа, однако в общем случае они требуют морфологического либо семантического анализа, что является самостоятельной и далеко не тривиальной задачей.

Литература

[1]. Julie Beth Lovins. (1968) Development of a Stemming Algorithm. In: Mechanical Translation and Computational Linguistics, vol.11, nos.1 and 2

[2]. Porter, Martin F. (1980). An Algorithm for Suffix Stripping. In: Program, 14 (3), pp 130-137.

[3]. Chris D. Paice. (1994) An evaluation method for stemming algorithms. In: Proceeding of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. — P. 42—50. — ISBN 0-387-19889-X.

[4]. Anjali Ganesh Jivani. (2011) A Comparative Study of Stemming Algorithms. In: International Journal of Computer Technology and Applications (IJCTA). - Vol 2 (6). - p. 1930-1938. - ISSN:2229-6093

[5]. Ilya Segalovich (2003) A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03, Las Vegas, Nevada, USA

[6]. Tomlinson S. (2004) Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer at CLEF 2003. In: Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003. Lecture Notes in Computer Science, vol 3237. Springer, Berlin, Heidelberg

[7]. Tomlinson S. (2005) Finnish, Portuguese and Russian Retrieval with Hummingbird SearchServer at CLEF 2004. In: Multilingual Information Access for Text, Speech and Images. CLEF 2004. Lecture Notes in Computer Science, vol 3491. Springer, Berlin, Heidelberg

[8]. Dolamic L., Savoy J. (2008) Stemming Approaches for East European Languages. In: Advances in Multilingual and Multimodal Information Retrieval. CLEF 2007. Lecture Notes in Computer Science, vol 5152. Springer, Berlin, Heidelberg

[9]. Губин, М. В. Влияние морфологического анализа на качество информационного поиска / М. В. Губин, А. Б. Морозов // Труды Восьмой Всероссийской научной конференции (RCDL'2006). - Ярославский государственный университет им. П. Г. Демидова, 2006. - с. 95 — 100.

[10]. Jongejan B., Dalianis H. (2009) Automatic Training of Lemmatization Rules that Handle Morphological Changes in pre-, in- and Suffixes Alike. In: Proceeding of the ACL-2009, Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, p. 145—153.

[11]. Прытков, В. А. Анализ репозитория университета с использованием графовой базы данных / В. А. Прытков, И. И. Пилецкий, Н. А. Волорова // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference. – Minsk, BSUIR, 2018. – P. 177 – 183.

[12]. Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. (2008) ArnetMiner: Extraction and Mining of Academic Social Networks. In: Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008), pp.990-998, Las Vegas, Nevada, USA

[13]. Пилецкий, И. И. Аналитический комплекс анализа данных из открытых интернет источников / И. И. Пилецкий, В. А. Прытков, Н. А. Волорова // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference. – Minsk, BSUIR, 2018. – P. 193 – 199.

[14]. Article recommendation system on a citation network using Personalized Pagerank and Neo4j. (September 9, 2018). https://tbgraph.wordpress.com/2018/09/09/article-recommendation-system-on-a-citation-network-using-personalized-pagerank-and-neo4j/

[15]. Alessandro Negro, Vlasta Kůs, Miro Marchi, Christophe Willemsen. (3 October 2017) Efficient unsupervised keywords extraction using graphs. https://graphaware.com/neo4j/2017/10/03/efficient-unsupervised-topic-extraction-nlp-neo4j.html

[16]. Slobodan Beliga, Ana Meštrović, Sanda Martincic-Ipsic. (2015) An Overview of Graph-Based Keyword Extraction Methods and Approaches. In: Journal of Information and Organizational Sciences 39(1), p. 1-20.

[17]. Omar Alqaryouti, Hassan Khwileh, Tarek Ahmed Farouk, Khaled Shaalan. (2018) Graph-Based Keyword Extraction. In book: Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence 740, Springer International Publishing

[18]. Shi, W., Zheng, W., Yu, J.X. et al. Keyphrase Extraction Using Knowledge Graphs. In: Data Science and Engineering (2017) 2: 275. https://doi.org/10.1007/s41019-017-0055-z

[19]. Фаулер, М., Садаладж, П. Дж. NoSQL: новая методология разработки нереляционных баз данных. : Пер. с англ. - М.: ООО "И.Д. Вильяме", 2013. - 192 с.

[20]. Робинсон Я., Вебер Д., Эифрем Э. Графовые базы данных: новые возможности для работы со связанными данными / пер. с англ. Р. Н. Рагимова; науч. Ред. А. Н. Кисилев. – 2-е изд. – М.: ДМК Пресс, 2016. – 256 с. - ISBN 978-5-97060-201-0

[21]. DB-Engines - Knowledge Base of Relational and NoSQL Database Management Systems. https://db-engines.com/en/

[22]. Christophe Willemsen. (24 July 2017) Reverse Engineering Book Stories with Neo4j and GraphAware NLP. https://graphaware.com/neo4j/2017/07/24/reverse-engineering-book-stories-nlp.html

[23]. https://github.com/graphaware/neo4j-nlp#pipelines-and-components

WORD FORMS NORMALIZATION AT ANALYSIS OF THE UNIVERSITY REPOSITORY USING A GRAPH DATABASE

V.A. PRYTKOV
PhD, Associate Professor, Vice-rector for education BSUIR

I.I. PILETSKI
PhD, Associate Professor, Informatics Department of the BSUIR

N.A. VOLARAVA
PhD, Associate Professor, Head of the Informatics Department of the BSUIR

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: prytkov@bsuir.by

Abstract. When texts in natural language are processed, for a wide class of applied problems the need arises to construct a glossary of terms. Its presence allows you to define the subject areas of texts as well as simplifies the search and systematization of information. This raises the problem of reduction various word forms of a term into a single form. The article discusses the approach to the word forms normalization on the example of construction an analysis system for the BSUIR university repository using the Neo4j graph database. This algorithm is based on a heuristic stemming algorithm but for the comfortable display of information the base of the word transforms to the lemma by using statistical information.

Keywords: stemming, lemmatisation, NLP, big data, repository, author, publication, word form, NoSQL, graph databases, Neo4j.