Методы и алгоритмы извлечения ключевых слов

Ванюшкин А.С., войсковая часть 49911
alexmandr@mail.ru
Гращенко Л.А., Академия ФСО России
graschenko@mail.ru

Аннотация

В работе выполнен систематизированный обзор методов и алгоритмов извлечения ключевых слов из текстов, приводится их классификация и хронология. Показано наличие небольшого числа перспективных подходов к выделению ключевых слов из русскоязычных текстов, несмотря на значительное количество публикаций в данной предметной области.

1 Введение

Объемы циркулирующей в мировых телекоммуникационных сетях и хранящейся на серверах информации демонстрируют динамику взрывного роста. По оценкам компании *Cisco Systems*, с 2010 по 2015 год ежемесячный объем передаваемого в сети Интернет трафика, включающий тексты и веб-данные, возрос с 2,4 до 8,6 экзабайт. А к 2018 году прогнозируется удвоение этого числа. Пропорционально растут показатели рынка текстовой аналитики, емкость которого по данным *International Data Corporation (IDC.com)* в 2015 году составила 2,65 млрд. долл., а прогноз на 2020 год – 5,9 млрд. долл. При этом в настоящее время анализируется менее 1% текстов, а рост рынка происходит в основном за счет анализа данных социальных сетей.

Все указанное обуславливает увеличение состава и сложности программных решений в области обработки текстов на естественных языках, в основе которых лежит ряд базовых алгоритмов, в том числе – алгоритмы выделения или извлечения из текстов ключевых слов (*Keyword Extraction*). В общем представлении ключевыми называются важные слова или фразы, дающие высокоуровневое описание содержания текстового документа, позволяющие выявить его тематику. Выделенный из текста список ключевых слов (КС) может выступать в качестве метаинформации, представляя текстовый документ при решении задач информационного поиска, классификации, кластеризации, аннотирования и реферирования [Manning, 1999]. При разработке автоматизированных систем, реализующих перечисленные задачи, необходимо опираться на эффективные алгоритмы, относительно которых имеется точная и проверенная информация о показателях их функционирования и границах применения. С позиций настоящей статьи, прежде всего, речь идет о русском языке, обладающим рядом особенностей по отношению к английскому, для которого разрабатывалась основная масса доступных открыто алгоритмов. Зачастую, авторы утверждают о высоких показателях работы своих алгоритмов извлечения КС, не приводя данных ни об их применении в тех или иных продуктах, ни об испытаниях на произвольном наборе иноязычных текстов.

В связи с указанными противоречиями, в данной работе предпринята попытка систематизировать имеющиеся сведения о проблемной области извлечения ключевых слов, разработанных методах и реализующих их алгоритмах, а также обосновать выбор направлений перспективных исследований.

2 Проблематика выделения ключевых слов

Анализ предметной области показывает наличие различных подходов к определению понятия «ключевое слово». Помимо общенаучного понимания ключевых слов как определяющих содержание текста и передающих его основной смысл, данный феномен рассматривается такими научными и прикладными дисциплинами как психолингвистика, теория коммуникации, компьютерная и когнитивная лингвистика, информатика [Москвитина, 2009]. Задача извлечения КС, именуемых также различными авторами «словами-концептами», «лексическими доминантами» и «смысловыми вехами», является одной из труднейших задач лингвистики текста [Попуша, 2008].

Обзор доступных публикаций показывает, что приводимые различными авторами с позиций разнообразных подходов определения

КС зачастую однотипны, при этом недостаточно формализованы. Обобщая многочисленные взгляды можно заключить, что *ключевыми словами называют особо важные, общепонятные, ёмкие и показательные для отдельно взятой культуры слова в тексте, набор которых может дать высокоуровневое описание его содержания для читателя, обеспечив компактное представление и хранение его смысла в памяти* [Стожок, 2009; Москвитина, 2009; Гринева, Гринев, 2009; Rose, 2010]. Вследствие этого, КС используются в информационном поиске, упрощая описание того или иного информационного ресурса, снижая объём необходимых для этого данных.

Понятия «ключевое слово» и «ключевое словосочетание» (фраза) рассматриваются большинством ученых как синонимы. Но отдельные исследователи не приемлют такой подход, указывая на существенные различия в содержании данных понятий. Ключевые фразы представляют собой сочетание двух или более слов, которые как могут следовать друг за другом в тексте, так и быть разделенными другими языковыми единицами [Turney, 2000]. Действительно, не все входящие в состав ключевых фраз слова при отдельном рассмотрении являются ключевыми. Но также вполне очевидно, что выделением отдельных КС затруднительно выразить основной смысл содержимого. Поэтому на практике востребовано выделение именно ключевых фраз, что близко к задаче создания списка терминов и ссылок на них (*back-of-book indexes*). Указанные вопросы рассматривалась в работах [Астраханцев, 2014; Захаров, Хохлова, 2014; Лукашевич, Логачев, 2010]. Основное отличие - длина списка на выходе алгоритма. Документу обычно соответствуют единицы ключевых фраз, длина же списка терминов, основой которых также являются КС [Стожок, 2009] колеблется от десятков до сотен. Составление перечня ключевых словосочетаний является одной из трудностей в рассматриваемой предметной области.

В результате систематизации данных различных исследователей нами выделен перечень существенных свойств и функций ключевых слов в текстах, значимых в контексте моделирования и алгоритмизации процесса их извлечения. Итак, ключевые слова характеризуются тем, что:

- являются наиболее употребительными (частотными) наименованиями [Стожок, 2009], обозначают признак предмета, состояние или действие [Попуша, 2008];
- представлены значимой лексикой, достаточно обобщены по своей семантике (средней степени абстракции), стилистически нейтральны, не оценочны [Стожок, 2009];
- связаны друг с другом сетью семантических связей, пересечения значений [Москвитина, 2009];
- более половины слов ядра тематического компонента состоит из ключевых слов, а минимальный набор КС приближается к инварианту содержания при их логическом упорядочивании;
- набор КС состоит из 5-15 [Turney, 2000] или 8-10 слов, что соответствует объему оперативной памяти человека [Москвитина, 2009], в тексте содержится 25-30% ключевых слов [Попуша, 2008];
- набор КС определяет контексты слов, обладающих максимальной предсказуемостью.

В процессе восприятия текста ключевые слова выделяют по синтаксической позиции (заголовок или первое предложение), по частотности употребления, лексическим паттернам, необычным сочетаниям, отношениям синонимии, антонимии, морфологической и семантической производности [Попуша, 2008], рис. 1.

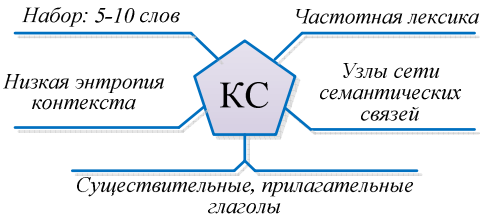

Рис. 1. Основные признаки ключевых слов

Несмотря на большое число специализированных и междисциплинарных работ, посвященных ключевым словам, до настоящего времени не разработана последовательная методика обнаружения ключевых слов человеком. Экспериментально подтверждено, что эта операция выполняется людьми интуитивно, и является личностно и даже гендерно обусловленной [Ноздрина, 2015]. Отсюда вытекает и сложность разработки методов и алгоритмов извлечения КС для вычислительной техники. Отсутствие четких формализованных моделей, чрезвычайно размытые определения с точки зрения компьютерной лингвистики и других инженерных дисциплин за-

трудняют создание и верификацию соответствующего инструментария.

Так как приведенные характеристики ключевых слов проявляются на нескольких уровнях рассмотрения текста – морфологическом, лексическом, синтаксическом и, прежде всего, прагматическом, то их распознавание подразумевает относительную сложность используемых методов и многоэтапность реализующих их алгоритмов. Действительно, библиографический обзор показывает, что в современных алгоритмах извлечения КС можно выделить три последовательных этапа, рис. 2.

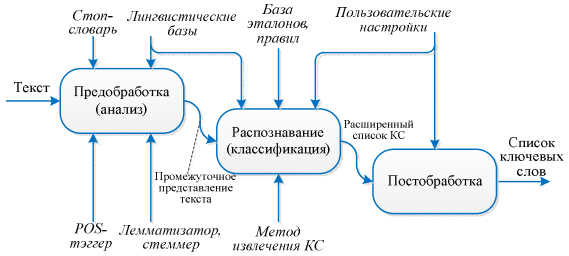

Рис 2. Типовая последовательность этапов извлечения ключевых слов

На первом этапе выполняется предварительная обработка текста, осуществляемая на графемном, морфемном и лексическом уровнях, призванная представить текст в формате, удобном для последующего распознавания. Здесь могут быть реализованы такие вспомогательные процедуры как графематический анализ (токенизация текста, удаление разметки), морфологический разбор, лексическая нормализация (в том числе согласование синонимов), лемматизация (стемминг), частеречевая разметка (*POS-tagging*), удаление стоп-слов (служебной лексики) и т.д. [Надеждин, 2015]. Все эти процедуры требуют использования специфических лингвистических баз и словарей, формирование которых зачастую не является тривиальной задачей [Усталов, 2012; Litvak, Last, Kandel, 2013]. Поэтому данный этап – языкозависимый, что означает различие в содержании предварительной обработки для разных языков. Поэтому большинство имеющихся алгоритмов выделения КС требуют адаптации для русского языка. На данном этапе может осуществляться первичный отбор кандидатов в КС с формированием списка слов или словосочетаний.

Существенные различия в содержании основных современных алгоритмов извлечения КС проявляются при реализации второго этапа – собственно распознавания ключевых слов (или фильтрации предварительного списка кандидатов). После установления значений ансамбля признаков в зависимости от выбранного подхода производится их сравнение с эталонами (порогом) и принятие решения о принадлежности того или иного слова-кандидата к множеству КС. В зависимости от базового метода извлечения КС в алгоритме могут использоваться различные лингвистические ресурсы - словари, корпуса, онтологии, поэтому данный блок может быть как языкозависимым, так и языконезависимым.

На заключительном этапе постобработки выходные данные - список КС - представляется в соответствии с пользовательскими или программными настройками в том или ином формате. Здесь может осуществляться усечение списка, его ранжирование и упорядочивание, визуализация методами когнитивной графики т.д. [Воронина, 2010].

Таким образом, ядром любого алгоритма извлечения КС является блок распознавания, основанный на конкретном методе в рамках того или иного подхода, классификация и диахроническое рассмотрение которых описывает содержание предметной области.

3 Технологические аспекты автоматического извлечения ключевых слов

3.1 Классификация методов извлечения ключевых слов

Доступные публикации описывают классификации методов автоматического извлечения КС разной степени полноты и детализации. В самом простом случае исследователи выделяют статистические и основанные на машинном обучении методы [Chen, Lin, 2010]. Схожая классификация приводится в работе отечественных авторов, которые рассматривают статистические и гибридные модели КС, на основе которых рассматриваются конкретные методы [Шереметьева, Осьминин, 2015]. Более развернутая классификация подразумевает выделение четырех страт: не требующих обучения простых статистических методов; лингвистических методов; основанных на машинном обучении методов и их комбинации [Zhang, 2008].

Последний из доступных отечественных обзоров предметной области извлечения ключевых слов приводит классификацию на основе типа системы распознавания, которая подразумевает выделение лингвистических, статистических и гибридных (лингво-

статистических) методов [Виноградова, Иванов, 2015]. Однако и эта, и прочие классификации, на наш взгляд, не отражают весь спектр и специфику существующих решений.

Так как любой алгоритм извлечения КС, по сути, реализует одну или несколько систем распознавания образов, разбивающих входное множество слов на два класса (ключевые и прочие), то предлагается использовать не иерархическую, а фасетную классификацию соответствующих методов и выбрать следующую совокупность признаков, рис. 3:

- наличие элементов обучения и подходы к его реализации;
- тип математического аппарата системы распознавания, обусловленного формой информации представления признаков ключевых слов;
- тип используемых для реализации метода лингвистических ресурсов.

Рис. 3. Классификация методов извлечения ключевых слов

По наличию элементов обучения выделяют необучаемые, обучаемые и самообучаемые методы извлечения КС. Более простые необучаемые методы подразумевают контекстно-независимое выделение КС из отдельного текста на основе априорно составленных моделей и правил. Они подходят для гомогенных по функциональному стилю корпусов текстов, увеличивающихся со временем в объемах, например научных работ или нормативных актов. Обучаемые методы предполагают использование разнообразных лингвистических ресурсов для настройки критериев принятия решений при распознавании ключевых слов. Здесь большое значение имеет корректное выделение КС в выборке, используемой для обучения. Среди методов с обучением можно выделить подкласс самообучаемых, если обучение ведется без учителя или с подкреплением (на основе пассивной адаптации).

По второму признаку классификации (рис. 3), прежде всего, следует выделить статистические и структурные методы извлечения КС. Статистические методы учитывают относительные частоты встречаемости морфологических, лексических, синтаксических единиц и их комбинаций. Это делает создаваемые на их основе алгоритмы довольно простыми, но недостаточно точными, т.к. признак частотности ключевых слов не является превалирующим [Salton, Yang, 1973]. Одним из классических методов в данном классе является расчет для каждого слова меры *TF-IDF* (*Term Frequency-Inverse Document Frequency*) [Jones, 1972], отражающей его важность в тексте, рассматриваемого как элемент коллекции документов.

В основе структурных методов лежит представление о тексте, как системе семантически и грамматически взаимосвязанных элементов-слов, которые в свою очередь характеризуются набором лингвистических признаков. Поэтому многие исследователи называют этот класс методов лингвистическим. Здесь в первом приближении могут быть выделены два подкласса - графовые и синтаксические (шаблонные) методы.

Графовые (граф-ориентированные) методы представляют текст множеством слов-вершин (или вершин-словосочетаний) и ребер-отношений между ними. Эти отношения могут выражать для каждой пары слов факты последовательного появления в тексте, наличия в окне заданного размера и семантическую близость. Для вершин полученного графа вычисляются меры центральности и по пороговому критерию отбираются ключевые слова. Различия между данными методами состоят в особенностях учета значимости каждой вершины и вычисления отношений между ними.

В основе синтаксических (шаблонных методов) лежит представление о регулярных синтаксических конструкциях, содержащих на определенных позициях ключевые слова. В чистом виде такие методы слабо применимы к рассматриваемой задаче, но могут использоваться в сочетании с другими.

Нейросетевые методы к задаче извлечения КС стали применяться сравнительно недавно и основаны на свойстве искусственных нейронных сетей к обобщению и выделению скрытых зависимостей между входными и выходными данными. Однако для формирования наборов данных для обучения и функционирования нейросетей требуется выделение структурных и статистических признаков, поэтому на практике методы выделения КС являются гибридными, т.е. сочетающими в себе элементы основных рассмотренных классов.

Наконец, алгоритмы извлечения КС, реализующие означенные методы могут не использовать какие-либо лингвистические ресурсы, или использовать разного рода словари, онтологии и тезаурусы, а также корпуса текстов (без разметки или с разметкой).

Стоит отметить, что приведенная классификация достаточно условна и не претендует на полноту при описании реально существующих разработок в рассматриваемой предметной области.

3.2 Обзор основных алгоритмов извлечения ключевых слов

Подходы к автоматическому извлечению КС менялись по мере развития моделей ключевых слов и теории распознавания образов. Графически представленная динамика исследований позволяет заключить, что фактор совершенствования средств вычислительной техники обусловил всплеск интереса и количества программных решений в рассматриваемой предметной области в последние годы, рис. 4.

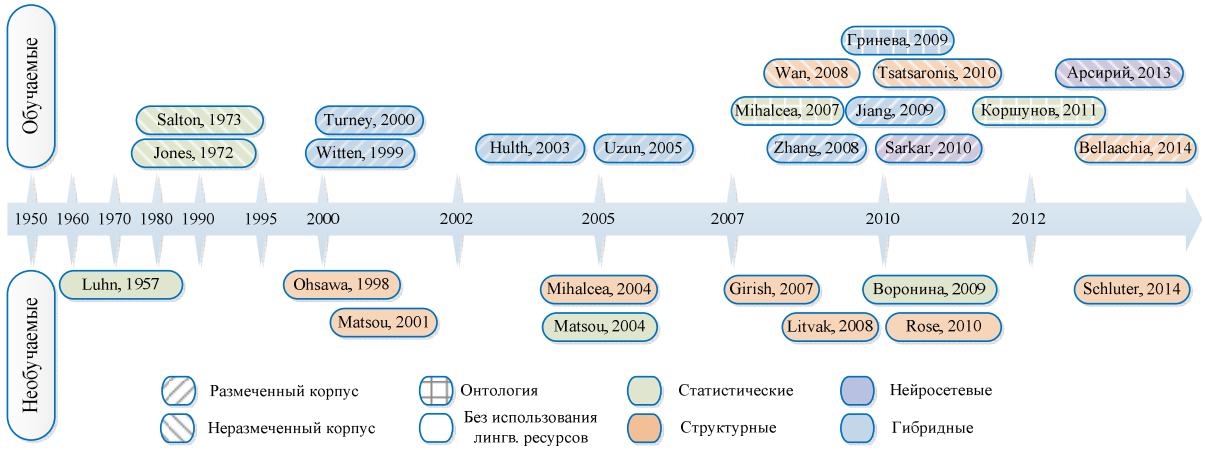

Рис. 4. Динамика исследований в области извлечения ключевых слов

Изначальные представления о выраженности фактора частотности при идентификации ключевых слов привели к тому, что первые алгоритмы реализовывали статистический подход, как например, алгоритм *KWIC* (*Key Words In Context*) [Luhn, 1958]. Впрочем, разработки на основе этих методов продолжаются [Matsuo, Ishizuka, 2004], в том числе отечественными исследователями [Воронина, 2009]. Последние модификации алгоритмов совмещают статистический подход с обучением, где в качестве корпуса используются ресурсы Википедии [Mihalcea, Csomai, 2007; Коршунов, 2011].

Среди алгоритмов, реализующих чисто структурные методы, описаны преимущественно граф-ориентированные, первые разработки которых появились на рубеже веков [Ohsawa, Benson, Yachida, 1998; Matsuo, 2001]. В короткие сроки появилось большое число улучшенных модификаций таких алгоритмов, особенности которых рассмотрим ниже.

Языконезависимый граф-ориентированный алгоритм *TextRank* был предложен в 2004 году [Mihalcea, Tarau, 2004] на основе известного алгоритма ранжирования веб-страниц PageRank. Значимость вершины в графе рассчитывается через значимости смежных вершин. Ребром графа здесь может служить любое отношение между лексическими единицами. Для задачи выделения КС это отношение смежности или совместного появления, задаваемое расстоянием между словами. Две вершины смежны, если соответствующие им лексические единицы появляются внутри окна ширины $2 \le N \le 10$. Перед добавлением вершин в граф может использоваться фильтрация лексики, например по принадлежно-

сти к частям речи. После построения графа выполняется подсчет значимости узлов, их ранжирование, и первые 5 - 20 сохраняются для дальнейшей обработки как потенциальные КС. Последовательности смежных КС образуют ключевые словосочетания, остальные признаются ключевыми словами.

В алгоритме *Rake* [Rose & etc., 2010] сначала формируется список потенциальных КС с помощью заданного словаря разделителей фраз, а затем строится граф, вершины которого - отдельные слова. Особенность такого графа состоит в том, что вершины графа могут быть представлены одинаковыми словами. Значимость для слова определяется набором показателей: частота появления вершины, степень вершины, отношение степени к частоте. Значимость потенциальной ключевой фразы рассчитывается как сумма значимостей каждого входящего в него слова. В качестве КС для данного текста отбирается первая треть упорядоченного по убыванию значимости списка вершин.

В алгоритме *DegExt* [Litvak, Last, Kandel, 2013] сперва удаляются стоп-слова, а затем строится граф, в котором дуги между вершинами проводятся только для соседствующих в любом предложении слов, не разделенных знаками пунктуации. Вершины с наибольшими степенями соответствуют кандидатам в КС. Для извлечения ключевых фраз необходимо выделить из списка последовательности смежных слов (заданной длины), встречающиеся в одном предложении. Для каждой фразы вычисляется значимость как средняя степень составляющих ее слов. По сравнению с *TextRank*, данный алгоритм вычислительно менее сложен. Авторы рекомендуют использовать его для выделения большого числа КС (около 15).

В работе [Schluter, 2014] для графа, построение которого аналогично алгоритму *DegExt*, рассматривается семь показателей центральности вершин, которые распределены по трем категориям: центральность по степени (*Degree-like Centrality*), по близости (*Closeness-like Centrality*) и по посредничеству (*Betweenness-like Centrality*). Показано, что лучшие результаты извлечения КС были достигнуты при использовании нормированной по длине центральности по посредничеству (*Length-scaled Betweenness Centrality*).

Другие графовые алгоритмы отличаются от вышеописанных использованием дополнительной информации о позициях слов, их длине, разметке и форматировании текста форматирование текста и т.д. [Воронина, 2009; Matsuo, 2001; Tsatsaronis, Varlamis, Science, 2010; Wan, Xiao, 2008].

Широкое применение машинного обучения к задаче извлечения КС началось с 2000-х годов, и количество таких алгоритмов уже превысило количество необучаемых алгоритмов. Прежде всего, это алгоритмы с обучением с учителем на основе гибридных структурно-статистических методов.

В работе [Turney, 2000] описан обучаемый алгоритм *GenEx*, совмещающий генетический алгоритм с экстрактором КС. Основными признаками для извлечения КС выбраны частоты слов и первые позиции их вхождения в тексте.

Отечественная разработка [Гринева, Гринев, 2009] совмещает две техники: рассчитанную на основе Википедии меру семантической близости и алгоритм обнаружения сообществ в сетях Гирвана-Ньюмана. Первоначально из текста извлекаются все возможные *n*-граммы и для каждой из них находятся статьи из Википедии. Многозначность разрешается путем поиска статей с наибольшей семантической близостью с контекстом. Затем строится семантический граф, где вес ребра равен численному значению семантической близости слов. Далее граф разбивается на сообщества, которые ранжируются (на основе плотности и информативности). Ключевыми терминами считаются термины из сообществ (от одного до трех) с наивысшими рангами. Тестирование алгоритма авторами осуществлялось на англоязычных текстах.

В других гибридных алгоритмах на основе обучения применяются байесовская классификация [Witten & etc., 1999; Hulth, 2003; Uzun, 2005]; метод условных случайных полей (CRF) [Zhang, 2008] и метод опорных векторов [Jiang, Hu, Li, 2009].

Стоит отметить, что в последнее время появились решения с самообучением. Это, прежде всего, нейронные сети [Sarkar, Nasipuri, Ghose, 2010] и самоорганизующиеся карты Кохонена [Арсирий и др., 2013].

В общем, несмотря на продолжающееся совершенствование классических статистических и структурных алгоритмов извлечения КС, акцент разработчиков сместился в область гибридных решений с обучением на основе текстовых корпусов.

3.3 Сравнительная характеристика основных алгоритмов извлечения КС

В данном параграфе приведем собранные из доступных источников следующие характеристики существующих алгоритмов извлечения КС: точность (*Precision*), полноту (*Recall*) и F-меру (*F-measure*), табл. 1. В настоящее время точность алгоритмов не может достичь 100% по двум принципиальным причинам. Во-первых, выделенные вручную КС не всегда присутствуют в тексте, вследствие чего возникает задача не извлечения, а генерации КС. Во-вторых, результат зависит от количества извлеченных КС, что во многих алгоритмах задается независимо от размера текста. Отсюда возникает необходимость в разработке объективного и универсального критерия оценки качества работы алгоритмов извлечения КС.

При составлении таблицы 1 обнаружились существенные расхождения авторских результатов и результатов независимых исследователей. Отсюда следует, что для объективного сравнения алгоритмов необходимо проводить тестирование на одном и том же тестовом корпусе.

Табл.1. Сравнительная характеристика основных алгоритмов выделения ключевых слов

Алгоритм	**P, %**	**R, %**	**F, %**	**Кто тестировал**	**Языконезависимость**	**Примечания**
TF-IDF (1972)	23,2	28,1	25,4	Wan X.	+	
KeyGraph (1998)	36,0	36,0	36,0	Matsuo Y.	н/д	
Witten I. (1999)	19,0	40,0	25,7	Sakar K.	н/д	Количество ключевых фраз – 10.
GenEx (2000)	11-29	-	-	автор	+	Количество КС от 4 до 16 на текстах различных тематик.
Hulth A. (2003)	25,2	51,7	33,9	автор	н/д	
Matsuo Y. (2004)	42,0	46,0	43,9	Matsuo Y.	н/д	
TextRank (2004)	31,2	43,1	36,2	автор	+	Количество ключевых слов – 1/3 от количества слов после фильтра (500 abstracts from the Inspec database).
Wikify (2007)	94,3	70,5	80,7	автор	+	Для работы необходима Википедия
DegExt (2008)	75,0	15,0	24,0	автор	+	База данных документов DUC 2002. 30 ключевых слов.
Zhang C. (2008)	66,3	41,9	51,2	автор	н/д	
Гринева М. (2009)	52,0	73,0	60,0	автор	+	Для работы необходима Википедия.
ExpandRank (2008)	28,6	35,2	31,6	автор	н/д	Количество ключевых фраз – 10
Rake (2010)	33,7	41,5	37,2	автор	+	Количество ключевых слов – 1/3 от количества слов после фильтра.
Sakar K. (2010)	22,0	46,0	29,7	автор	н/д	Количество ключевых фраз – 10
SemanticRank (2010)	42,1	53,2	47,0	автор	н/д	Количество ключевых фраз – 20
Коршунов А. (2011)	40,0	68,6	50,5	автор	н/д	Для работы необходима Википедия

3.4 Программные средства

Помимо алгоритмов, опубликованных в научной периодике существует множество коммерческих решений в области автоматической обработки текстов, в которых реализован функционал извлечения КС. Для большинства из них алгоритм работы не опубликован. Независимое тестирование продуктов Lexalitycs[1], Alchemy[2], Extractor[3] показало худшие, чем у авторов, результаты [Mihalcea, Tarau, 2004]. В табл. 2 приведено сравнение данных продуктов.

Табл.2. Сравнение программных продуктов

Продукт	РЯ	API	Интерфейс	Окраска слов
Lexalytics	-	+	+	+
Alchemy	+	+	+	+
Extractor	-	+	-	-

[1] http://www.lexalytics.com
[2] http://www.alchemyapi.com
[3] http://extractor.com/

Как видно, несмотря на хорошие интерфейсные решения и наличие встроенного API для сторонних разработчиков, поддержка русского языка заявлена не во всех продуктах, и объективных данных о результативности в отношении русского языка нет.

4 Выводы

Анализ существующих методов и алгоритмов извлечения КС показывает, что на фоне роста доступных вычислительных ресурсов в настоящее время усилия исследователей направлены на развитие обучаемых гибридных технологий. При этом вычислительно более простые граф-ориентированные алгоритмы обладают рядом дополнительных преимуществ, таких как независимость от языка и размеченных корпусов (онтологий). Работы по данной тематике в большинстве англоязычные, поэтому применимость большинства рассмотренных решений к русскому языку не установлена. Отдельного обсуждения и исследования заслуживают реализуемые на этапе предобработки процедуры нормализации текста и построения стоп-словаря.

В России значимые результаты получены коллективами исследователей из Института системного программирования РАН и Воронежского государственного университета. При этом по-прежнему актуальными остаются вопросы адаптации известных разработок к русскоязычным текстам, создания решений с апробированными показателями функционирования для широкого класса задач (от пользовательских приложений до высокопроизводительных).

Следующим этапом в развитии теории и практики извлечения КС станет попытка решить задачу генерации КС, не находящихся в конкретном тексте.

С учетом приведенных положений, дальнейшая исследовательская работа будет направлена на реализацию программного стенда, с помощью которого планируется провести независимое объективное тестирование передовых алгоритмов извлечения КС на специально подготовленном корпусе русскоязычных текстов.

Список литературы

Арсирий Е.А. *Построение контекстной карты на основе SOM для выделения ключевых слов веб-документов образовательных интернет-ресурсов* / Е.А. Арсирий, А.А. Чугунов, Ю.Н. Ларченко // Труды Одесского политехнического университета. - 2013. - № 1 (40). - С. 49-54.

Астраханцев Н.А. *Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии*. Труды Института системного программирования РАН, 2014. Т. 26. № 4. С. 7–20.

Баканова, Н.Б. *Обзор программных средств автоматизированного поиска и анализа ключевых слов документов* / Н.Б. Баканова // Проблемы современной науки. - 2013. - № 7-3. - С. 40-45.

Виноградова Н.В., Иванова В.К. *Современные методы автоматизированного извлечения ключевых слов из текста*. URL: http://cdokp.tstu.tver.ru/site.services/download.aspx?act=1&dbid=marcmain&did=110935. (дата обращения 05.03.2016).

Воронина, И.Е. *Функциональный подход к выделению ключевых слов: методика и реализация* / И.Е. Воронина и др. // Вестник Воронежского государственного университета. - 2009. - № 1. - С. 68-72.

Воронина, И.Е. *Алгоритмы определения семантической близости ключевых слов по их окружению в тексте* / И.Е. Воронина, А.А. Кретов, И.В. Попова // Вестник Воронежского государственного университета. - 2010. - № 1. - С. 148-153.

Гринева М., Гринев М. *Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов*. Труды Института системного программирования РАН. 2009. Т. 16. С. 155–165.

Захаров В.П., Хохлова М.В. *Автоматическое выявление терминологических словосочетаний. Структурная и прикладная лингвистика*. 2014. № 10. С. 182–200.

Коршунов А.В. *Извлечение ключевых терминов из сообщений микроблогов с помощью Википедии*. Труды Института системного программирования РАН. 2011. Т. 20. С. 269–282.

Кретов, А.А. *Маркемы и ключевые слова в научных текстах* / А.А. Кретов // Мир лингвистики и коммуникации: электронный научный журнал. - 2012. - Т. 1, № 27. - С. 1-13.

Лукашевич, Н.В. *Комбинирование признаков для автоматического извлечения терминов* / Н.В. Лукашевич, Ю.М. Логачев // Вычислительные методы и программирование. - 2010. - Т. 11. - С. 108–116.

Москвитина, Т.Н. *Ключевые слова и их функции в научном тексте* / Т.Н.Москвитина // Вестник ЧГПУ. - 2009. - № 11. - С. 270-283.

Надеждин, Е.Н. *Задача выявления цепочки ключевых слов и предложений при семантическом анализе текста* / Е.Н. Надеждин // Научный альманах. - 2015. - № 9 (11). - С. 773-778.

Ноздрина, Т.Г. *Особенности восстановления текстов – оригиналов на основе ключевых слов* / Т.Г. Ноздрина // Современные проблемы науки и образования. - 2015. - № 1-2. - С. 167.

Папуша, И.С. *Сложное синтаксическое целое: ключевые слова или гермы* / И.С. Папуша // Вестник Ассоциации ВУЗов туризма и сервиса. - 2008. - № 3. - С. 48-54.

Стожок, Е.В. *Ключевые слова как элементы терминосистем* / Е.В. Стожок // Вестник Бурятского государственного университета. - 2009. - № 11. - С. 101-104.

Усталов, Д.В. *Извлечение терминов из русскоязычных текстов при помощи графовых моделей*. URL: http://koost.eveel.ru/science/CSEDays2012.pdf. (дата обращения 05.03.2016).

Шереметьева, С.О. *Методы и модели автоматического извлечения ключевых слов* / С.О. Шереметьева, П.Г. Осминин // Вестник Южно-Уральского государственного университета. - 2015. - Т. 12, № 1. - С. 76–81.

Matsuo Y. *Extracting Keywords from Documents Small World*. Discov. Sci. Springer Berlin Heidelb. 2001. pp. 271–281.

Chen P.I., Lin S.J. *Automatic keyword prediction using Google similarity distance*. Expert Syst. Appl. 2010. Vol. 37. Iss. 3. pp. 1928–1938.

Hulth A. *Improved Automatic Keyword Extraction Given More Linguistic Knowledge* . EMNLP'03 Proc. 2003 Conf. Empir. Methods Nat. Lang. Process. 2003. № 2000. pp. 216–223.

Jiang X., Hu Y., Li H. *A Ranking Approach to Keyphrase Extraction*. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston, MA, USA. 2009. pp. 756–757.

Jones K.S. *A statistical interpretation of term specificity and its application in retrieval*. J. Doc. 1972. Vol. 28. pp. 11–21.

Litvak M., Last M., Kandel A. *DegExt: A language-independent keyphrase extractor*. J. Ambient Intell. Humaniz. Comput. 2013. Vol. 4. pp. 377–387.

Luhn H.P. *The Automatic Creation of Literature Abstracts*. IBM J. Res. Dev. 1958. pp. 159–165.

Manning C.D. *Foundations of statistical natural language processing*. Cambridge: The MIT Press, 1999.

Matsuo Y., Ishizuka M. *Keyword Extraction From a Single Document Using Word Co-Occurrence Statistical Information*. Int. J. Artif. Intell. Tools. 2004. Vol. 13. pp. 157–169.

Mihalcea R., Csomai A. *Wikify ! Linking Documents to Encyclopedic Knowledge*. Proceedings of the sixteenth ACM Conference on information and knowledge management. New York, NY, USA. 2007. pp. 233–242.

Mihalcea R., Tarau P. *TextRank: Bringing order into texts*. Proc. EMNLP. 2004. Vol. 4. PP. 404–411.

Ohsawa Y., Benson N.E., Yachida M. *KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor*. Proc. IEEE Int. Forum Res. Technol. Adv. Digit. Libr. -ADL'98. 1998. pp. 12–18.

Rose S., Engel D., Cramer N., Cowley W. *Automatic Keyword Extraction from Individual Documents*. Text Min. Appl. Theory. 2010. pp. 1–20.

Salton G., Yang C. *On the Specification of Term Values in Automatic Indexing*. Values Autom. Indexing. J. Doc. 1973. Vol. 29. Iss. 4. pp. 351–375.

Sarkar K., Nasipuri M., Ghose S. *A New Approach to Keyphrase Extraction Using Neural Networks*. Int. J. Comput. Sci. Issues. 2010. Vol. 7. Iss. 2. pp. 16–25.

Schluter N. *Centrality Measures for Non-Contextual Graph-Based Unsupervised Single Document Keyword Extraction*. 21 Traitement Automatique des Langues Naturelles. 2014. pp. 455–460.

Tsatsaronis G., Varlamis I., Science I. *SemanticRank : Ranking Keywords and Sentences Using Semantic Graphs*. COLING'10 Proceedings of the 23rd International Conference on Computational Linguistics. 2010. pp. 1074–1082.

Turney P.D. *Learning Algorithms for Keyphrase Extraction*. Inf. Retr. Boston. 2000. Vol. 2. Iss. 4. pp. 303–336.

Uzun Y. Keyword Extraction Using Naive Bayes URL: http://www.cs.bilkent.edu.tr/~guvenir/courses/CS550/Workshop/Yasin_Uzun.pdf (дата обращения 05.03.2016).

Wan X., Xiao J. *Single Document Keyphrase Extraction Using Neighborhood Knowledge*. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence. 2008. pp. 855–860.

Witten I.H., Paynter G.W., Frank E., Gutwin C., Craig G. Nevill-Manning. *KEA: Practical Automatic Keyphrase Extraction.* Proc. 4th ACM Conf. Digit. Libr. 1999. pp. 254–255.

Zhang C., Wang H., Liu Y., Wu D., Liao Y., Wang B. *Automatic Keyword Extraction from Documents Using Conditional Random Fields*. J. Comput. Inf. Syst. 2008. Vol. 4. pp. 1169–1180.