1.6. ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ ТЕРМИНОВ НАПРАВЛЕНИЯ «ЦИФРОВАЯ ЭКОНОМИКА»: ГРАФООРИЕНТИРОВАННЫЙ ПОДХОД

Милкова М.А., научный сотрудник,
Центральный экономико-математический институт РАН

Статья посвящена извлечению ключевых терминов из правительственных документов, выпущенных в период 2013-2018 годы и связанных с направлением Цифровая экономика. Изучение правительственных документов представляет интерес с точки зрения анализа одного из источников зарождения терминологии цифровой экономики. В статье приводится краткий обзор основных подходов к извлечению ключевых терминов из текста, а также дается детальное описание одного из графоориентированных методов – алгоритма TextRank. Выбранный алгоритм был протестирован на 13 правительственных документах. Результатом обработки каждого текста явилось построение взвешенного графа семантических связей между ключевыми словами, на основании которого были выделены ключевые термины.

Введение

Извлечение ключевых терминов является важной задачей, лежащей на стыке таких областей знаний, как интеллектуальный анализ текста (Text Mining), информационный поиск (Information Retrieval) и обработка естественного языка (Natural Language Processing). Под ключевыми терминами понимаются ключевые слова или ключевые фразы, которые наилучшим образом характеризует содержание исследуемого текста. Задача автоматического определения ключевых терминов представляет собой необходимый этап обработки текста для решения таких задач, как: создание и развитие терминологических ресурсов, автоматический информационный поиск, аннотирование, классификация и кластеризация документов, суммаризация текста и др.

Особый интерес представляет извлечение терминологии из текстов, относящихся к новым, только формирующимся направлениям. Так, направление цифровой трансформации является в настоящий момент своего рода мейнстримом мирового экономического развития и затрагивает многие промышленные и социальные сферы. Следуя глобальному тренду, Россия также движется в сторону цифровой трансформации экономики, что порождает большое число текстовой информации в данной предметной области (документы, статьи, новости, стенограммы и др.). Готова ли Россия к цифровой трансформации - остается вопросом (Dobrolyubova et.al., 2017), однако направление «Цифровая экономика» было выделено правительством РФ в качестве приоритетного – в 2017 году была разработана и утверждена программа, в ходе которой определены цели, задачи, направления и сроки реализации основных мер государственной политики по созданию необходимых условий для развития в России цифровой экономики. Для управления программой определены пять базовых направлений на период до 2024 года: нормативное регулирование, кадры и образование, формирование исследовательских компетенций и технических заделов, информационная инфраструктура и информационная безопасность.

Таким образом, образовался ряд правительственных документов, представляющих собой неплохую базу для извлечения терминологии по направлению «цифровая экономика». Конечно, ограничение сферы анализа исключительно правительственными документами не претендует на представление полной картины, но является неплохой отправной точкой как для отбора самых ключевых терминов, так и для тестирования методов. Кроме того, извлечение ключевых терминов из правительственных документов интересно с точки зрения анализа одного из источников зарождения терминологии цифровой экономики (так, например, термин «сквозные технологии» был впервые упомянут Президентом РФ В. В. Путиным в послании Федеральному собранию, 01.12.2016).

В данной статье мы приводим краткий обзор основных подходов к извлечению ключевых терминов из текста и фокусируемся на графоориентированных методах, в частности, на алгоритме TextRank (Mihalcea and Tarau, 2004). Выбранный алгоритм был протестирован на 13 правительственных документах, непосредственно связанных с направлением Цифровая экономика. Результатом обработки каждого из документов явилось построение графа семантических связей между ключевыми словами, на основании которого возможно выделение ключевых терминов (то есть ключевых словосочетаний и/или ключевых слов).

Краткий обзор подходов к извлечению ключевых терминов

Анализ литературы в области извлечения ключевых терминов выявил большое число методов и их модификаций, однако общепринятой в научном сообществе классификации данных подходов на текущий момент не существует. Действительно, решение задачи автоматического выделения ключевых терминов ведется одновременно по двум направлениям. С одной стороны, методы различаются по типу математического аппарата распознавания ключевых слов (статистические, методы на основе машинного обучения, структурные), с другой – по типу используемых (или не используемых) лингвистических ресурсов (словари, онтологии и тезаурусы, корпуса текстов). Наиболее простым статистическим методом извлечения ключевых слов является метод ранжирования всех словоформ по частоте. При

подсчете частоты употребления ключевого слова учитываются все его словоформы в тексте. Создаваемые на основе данного подхода алгоритмы являются недостаточно точными, т.к. признак частотности ключевых слов не является превалирующим (Salton, Yang, 1973).

Для повышения корректности автоматического извлечения ключевых слов, статистический метод дополняется лингвистическими процедурами (морфологическим, синтаксическим или семантическим анализом). Такие методы могут требовать или не требовать корпусов текстов[1]. Использование корпуса текстов получило достаточно широкое распространение, однако отсутствие таких корпусов для каждой конкретной предметной области в реальной жизни делает применение таких корпусных моделей и методов весьма проблематичным (Шереметьева, Осминин, 2015). Методы на основе машинного обучения рассматривают задачу извлечения ключевых терминов как задачу классификации – вычисление вероятности отнесения слова к ключевому на основе обучающей выборки – корпуса документов с размеченными ключевыми словами. В основе структурных методов лежит представление о тексте как системе семантически и грамматически взаимосвязанных элементов-слов, которые, в свою очередь, характеризуются набором лингвистических признаков (Ванюшкин, Гращенко, 2016). Здесь могут быть выделены два подкласса – графовые и синтаксические (шаблонные) методы.

В данной статье мы не будем останавливаться на детальном обзоре различных подходов, отметим лишь, что с наиболее системным русскоязычным обзором читатель может ознакомиться в работе (Ванюшкин, Гращенко, 2016)). Отметим также, что, несмотря на достаточно большое количество исследований, автоматическое извлечение ключевых терминов (многокомпонентных ключевых слов) представляет собой проблему, которая до сих пор не решена (Sag, et.al., 2002); (Шереметьева, Осминин, 2015). Более того, применение некоторых методов ограничено языками с бедной морфологией. Так, чисто статистические модели извлечения ключевых слов, удовлетворительно работающие, например, на материале английского языка, непригодны для естественных языков с богатой морфологией, в частности, для русского языка, где каждая лексема характеризуется большим количеством словоформ с низкой частотностью в каждом конкретном тексте (Шереметьева, Осминин, 2015).

Графоориентированный подход

Выбор того или иного алгоритма извлечения ключевых терминов обуславливается в первую очередь естественным языком, спецификой исследуемой темы, объемом анализируемого текста. В данной статье был выбран графоориентированный подход к извлечению ключевых терминов из русскоязычных документов. Графовые модели представляют большой интерес для области обработки естественного языка благодаря своей универсальности (не зависят от естественного языка) и эффективности основанных на них алгоритмов. Графовые методы не предполагают использование лингвистических ресурсов для настройки критериев принятия решений при распознавании ключевых терминов. Вместо этого, в работе алгоритмов подразумевается контекстно-независимое выделение ключевых терминов, что является оптимальным решением для гомогенных по функциональному стилю корпусов текстов, например, научных работ или нормативных актов (Ванюшкин, Гращенко, 2016), а также работ, посвященных новым, развивающимся темам, для которых не существует разработанных словарей, тезаурусов и т.п. Подробный обзор и классификацию графовых алгоритмов можно найти в работах (Beliga, et.al., 2015), (Mihalcea and Radev, 2011). В данной статье мы ограничимся описанием общих базовых моментов.

Итак, в основе графовых моделей лежит процедура построения графа, в вершинах которого стоят лексические единицы (слова или предложения), а отношения между ними представлены в виде ребер графа. Отношение между лексическими единицами может быть основано на различных принципах, наиболее распространенными из которых являются:

- Отношение совместной встречаемости – связанные слова встречаются в тексте внутри окна фиксированного размера; связаны все слова внутри предложения, параграфа или документа.
- Семантическое отношение – связанные слова имеют одинаковое значение, синонимы, антонимы, омонимы и др.

В зависимости от наличия или отсутствия направленности ребер граф может быть ориентированным (показывать последовательность появления слов в тексте) или неориентированным (показывать наличие связи). Также ребра графа могут быть как взвешенными, так и невзвешенными в зависимости от отношения между вершинами. Вес ребра может представлять собой расстояние между двумя словами внутри окна (предложения, параграфа, текста) или частоту совместной встречаемости пары слов в тексте.

Для вершин полученного графа вычисляется мера центральности как индикатор определения наиболее значимых вершин внутри графа. Центральность вершины *v* – это мера, отражающая то, насколько вершина *v* участвует в процессе распространения информации между остальными вершинами в графе (Цынгуев, 2015). В области извлечения ключевых слов используются различные меры центральности (Beliga, et.al., 2015); (Цынгуев, 2015), на основе которых производится ранжирование слов текста. Среди обилия графовых алгоритмов (Beliga, et.al., 2015) был выбран алгоритм TextRank,

[1] Корпус текстов - подобранная и обработанная по определенным правилам совокупность текстов, используемых в качестве базы для исследования языка.

предложенный в работе (Mihalcea and Tarau. 2004) и являющийся приложением алгоритма PageRank[2] к задачам обработки естественного языка.

Алгоритм TextRank

В основе TextRank лежит процедура построения взвешенного графа, в вершинах которого стоят лексические единицы (слова или предложения), а ребра взвешены в соответствии с силой связи между ними. В нашей работе мы будем пользоваться алгоритмом TextRank для извлечения ключевых слов, имеющих между собой семантическую связь, тем самым получая ключевые термины. После того, как произведена предобработка текста, производится построение взвешенного неориентированного графа $G = (V, E)$, где *V* – множество уникальных лексических единиц (вершины); *E* – множество связей между ними (ребра).

В качестве меры связи между словами TextRank использует отношение совместной встречаемости: две вершины соединяются ребром, если их лексические единицы встречаются вместе внутри окна из *N* слов, $N \in [2, 10]$. В работе (Усталов, 2012) предложено определять вес каждого ребра по формуле:

$$w_{ij} = \begin{cases} 1 - \dfrac{d(w_i, w_j) - 1}{(N-1)}, & \text{если } d(w_i, w_j) \epsilon (0, N) \\ 0, & \text{если } d(w_i, w_j) \geq (0, N) \end{cases} \quad (1)$$

где $d(w_i, w_j)$ – расстояние между словом w_i и w_j (модуль разности позиций), *N* – размер окна.

В нашей работе формула (1) была расширена добавлением «штрафа» за совместную встречаемость слов внутри окна, но в разных предложениях:

$$w_{ij} = \begin{cases} 1 - \dfrac{d(w_i, w_j) - 1}{(N-1)\left(2 \cdot d\left(s(w_i), s(w_j)\right) + 1\right)}, & \text{если } d(w_i, w_j) \epsilon (0, N) \\ 0, & \text{если } d(w_i, w_j) \geq (0, N) \end{cases} \quad (2)$$

где $d\left(s(w_i), s(w_j)\right)$ – расстояние между предложениями, в которых находятся слова.

На следующем этапе по итеративной формуле вычисляется TextRank (TR), получаемый случайным блужданием для каждой вершины из V:

$$TR(\mathrm{V_i}) = (1-d) + d \cdot \sum_{V_j \epsilon In(V_i)} \frac{w_{ij}}{\sum_{V_k \epsilon Out(V_j)} w_{jk}} TR(V_j), \quad (3)$$

где W_{ij} – вес соответствующего ребра, $In(V_i)$ – набор вершин, из которых идет связь в V_i (предшественники), $Out(V_i)$ – набор вершин, в которые идет связь из *Vi* (последователи). Данные обозначения введены, так как TextRank, как было упомянуто выше, является приложением алгоритма PageRank для ранжирования веб-страниц, где построенный граф является ориентированным. В случае задачи обработки текста граф является неориентированным и $In(Vi)=Out(Vi)$.

$d \in (0,1)$ – так называемый коэффициент затухания (damping factor) – в контексте веб-серфинга d определяет вероятность того, что на странице пользователю станет скучно и он перейдет на другую случайную страницу. В некоторых работах (Brin and Page,1998); (Mihalcea and Tarau, 2004) коэффициент d предлагается брать равным 0.85.

Начальное значение TR для каждой из вершин предполагается равным 1. Алгоритм повторяется до достижения порогового уровня сходимости.

В соответствии с итоговыми значениями *TR* вершины графа ранжируются, отбираются T «лучших» вершин (с наибольшим значением *TR*). Ключевые фразы получаются путем извлечения из текста последовательностей, состоящих из T-лучших слов.

Применение

В нашей работе алгоритм TextRank был реализован для 13 правительственных документов, выпущенных в период 2013-2018 годы и связанных с направлением Цифровая экономика. Схематично связи между проанализированными документами представлены на (рис. 1). Центральное место в схеме занимает документ «Цифровая экономика», входящие в него блоки представляют документы, на которые ссылается «Цифровая экономика», а исходящие – документы, которые на «Цифровую экономику»

[2] PageRank – алгоритм используемый Google Search для ранжирования веб сайтов в результатах выдачи поисковой системы (Page et al., 1998).

ссылаются сами. Отдельно стоящий блок «Фабрика проектного финансирования» не имеет прямых ссылок на «Цифровую экономику», но направлен на решение поставленных в ней задач. В дальнейшем схема может быть расширена добавлением новых блоков – мероприятий, протоколов заседаний, совещаний и т.п.

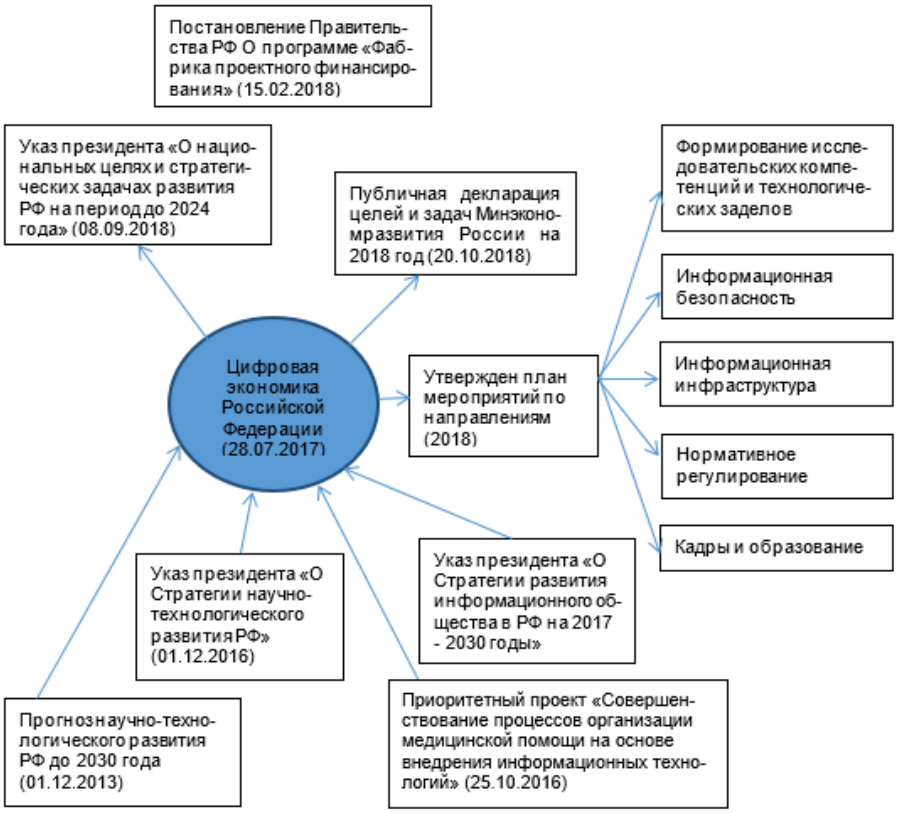

Рисунок 1. Схематичное отображение связей между правительственными документами по программе Цифровая экономика

Алгоритм TextRank был реализован на языке программирования Python с использованием готовых библиотек, а также с помощью адаптации кода некоторых библиотек с учетом особенностей русского языка и специфики решаемой задачи.

- Предобработка текста

Тексты анализировались в формате plain text. На первом этапе была проведена предобработка текста, которая включала следующие шаги:

1. Токенизация текста – разбивка на отдельно значимые единицы, в нашем случае – слова. Начальная фильтрация – из списка полученных токенов были убраны знаки пунктуации (все, кроме точек).
2. Лемматизация текста – токены приведены к нормальной (словарной) форме. Так, например, словоформы «цифровые», «цифровых», «цифровыми» преобразовываются к лемме «цифровой». Морфологическая обработка осуществлялась с помощью библиотеки Mystem, разработанной компанией Yandex[3].
3. Фильтрация – исключение стоп-слов. Фильтрация производилась в два этапа. Первичный список стоп-слов формировался на основе библиотеки NLTK (Bird, et.al., 2009) и состоял из предлогов, союзов, междометий и т.п.

На втором этапе формировался дополнительный список стоп-слов, в который входили существительные, прилагательные и глаголы- кандидаты в ключевые слова, полученные после завершения

[3] Отметим, что в изначальном алгоритме (Mihalcea and Tarau, 2004), используется процедура стемминга - нормализация словоформы к ее квази-основе. Так, например, вышеприведенные словоформы будут усечены до формы «цифров». Очевидно, что в силу особенностей русского языка приведение итоговых ключевых слов в виде усеченных квази-основ значительно усложняет восприятие полученных результатов, в силу чего и было принято решение о замене стемминга лемматизацией.

работы алгоритма. Список дополнительных стоп-слов подбирался для каждого документа индивидуально и включал слова, отражающие специфику текста, но не имеющие смысловой нагрузки. Например, к дополнительным стоп-словам были отнесены: «гражданин», «акт», «необходимо», «срок» и др.

- Построение графа

Полученные слова-леммы использовались в качестве вершин графа, который был построен с помощью библиотеки NetworkX. В некоторых работах предлагается ограничить число вершин в графе, например, включая только имена существительные и имена прилагательные (Mihalcea and Tarau, 2004); (Усталов, 2012). Однако в нашей работе были оставлены и глаголы.

Далее был реализован алгоритм установления связей между вершинами. Размер окна для поиска совместной встречаемости слов выбирался экспериментально и варьировался от 2 до 4. Значения N >4 не рассматривались, поскольку увеличение размера окна приводит к заметной деградации точности извлечения терминов (Усталов, 2012). Также в процессе работы алгоритма была учтена разбивка на предложения: величина связи между двумя словами, находящимися внутри окна N, но в разных предложениях устанавливалась меньшей, чем между словами в одном предложении, как это показано в (2).

Из вычисленного TR по формуле (3) составлялось множество кандидатов в ключевые слова. Список вершин упорядочивался по убыванию значения TR, после чего отбирались первые T вершин. В некоторых работах (Mihalcea and Tarau, 2004); (Усталов, 2012) предложено выбирать T=1/3|V|, однако в силу большей размерности графа для рассматриваемых нами документов данный подход неприменим. Значение T устанавливалось экспериментально и варьировалось от 15 до 21 в зависимости от длины исходного текста. В случае если в список из T выбранных слов попадали слова, относительно которых принималось решение о внесении их в список дополнительных стоп-слов, алгоритм пересчитывался.

Полученные результаты были представлены в виде графа семантических связей с использованием библиотеки Matplotlib. Размер узла графа пропорционален значению TR слова, а сила связи между словами выражена в интенсивности цвета ребра между ними. Приведем граф для документа «Прогноз научно-технологического развития Российской Федерации на период до 2030 года».

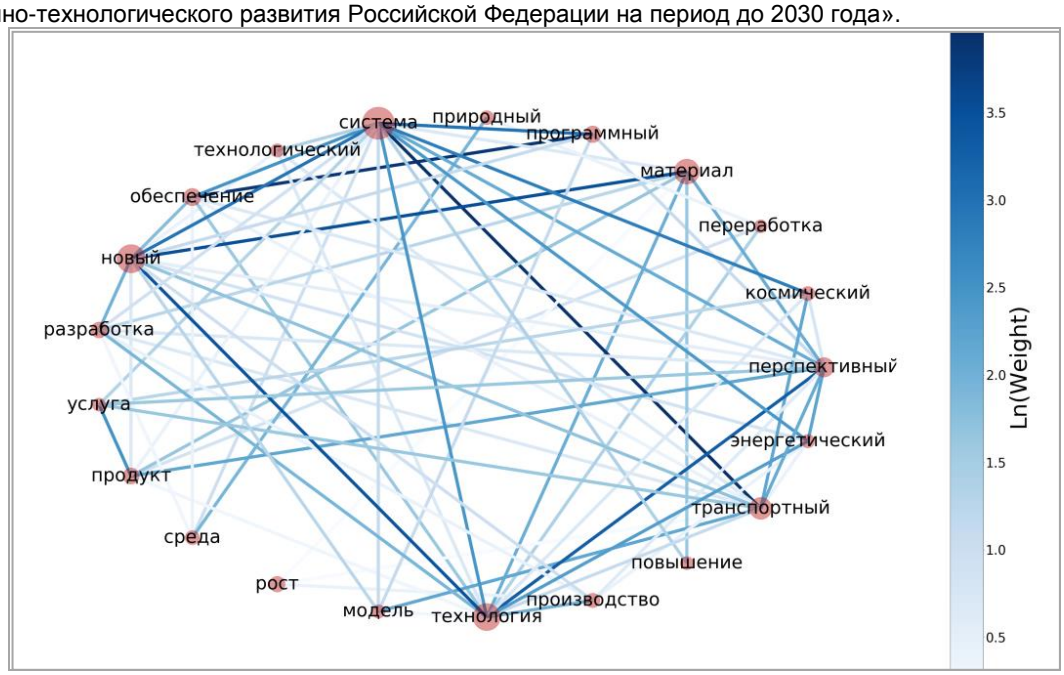

Рисунок 2. Граф семантических связей между ключевыми словами в документе «Прогноз научно-технологического развития Российской Федерации на период до 2030 года»

Как видно из графа, для данного документа могут быть выделены следующие ключевые термины: Программное обеспечение, транспортная система, программная система, новые материалы и технологии, новые разработки, перспективные технологии, космическая система, разработка новых технологий, транспортная модель, природная среда, энергетическая система, перспективные транспортные и космические системы, космические услуги, перспективные технологии переработки, продукты переработки. Ниже приведен граф для Указа президента «О Стратегии развития информационного общества в Российской Федерации на 2017 – 2030 годы».

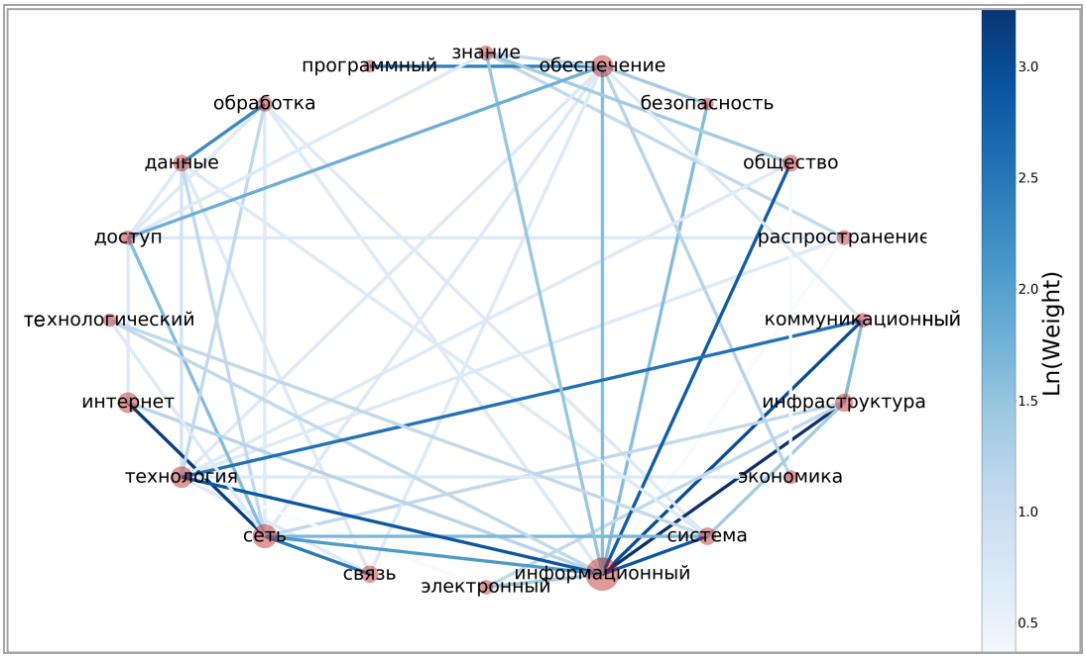

Рисунок 3. Граф семантических связей между ключевыми словами в документе «Указ президента о Стратегии развития информационного общества в Российской Федерации на 2017 – 2030 годы»

Таким образом, для рассматриваемого документа могут быть выделены следующие ключевые термины: сеть интернет, информационное общество, информационные технологии, информационные и коммуникационные технологии, информационные и коммуникационные инфраструктуры, обработка данных, сети связи, обеспечение доступа, распространение доступа, распространение знаний, информационная безопасность, программное обеспечение и др. Далее представлен граф для документа «Цифровая экономика Российской Федерации».

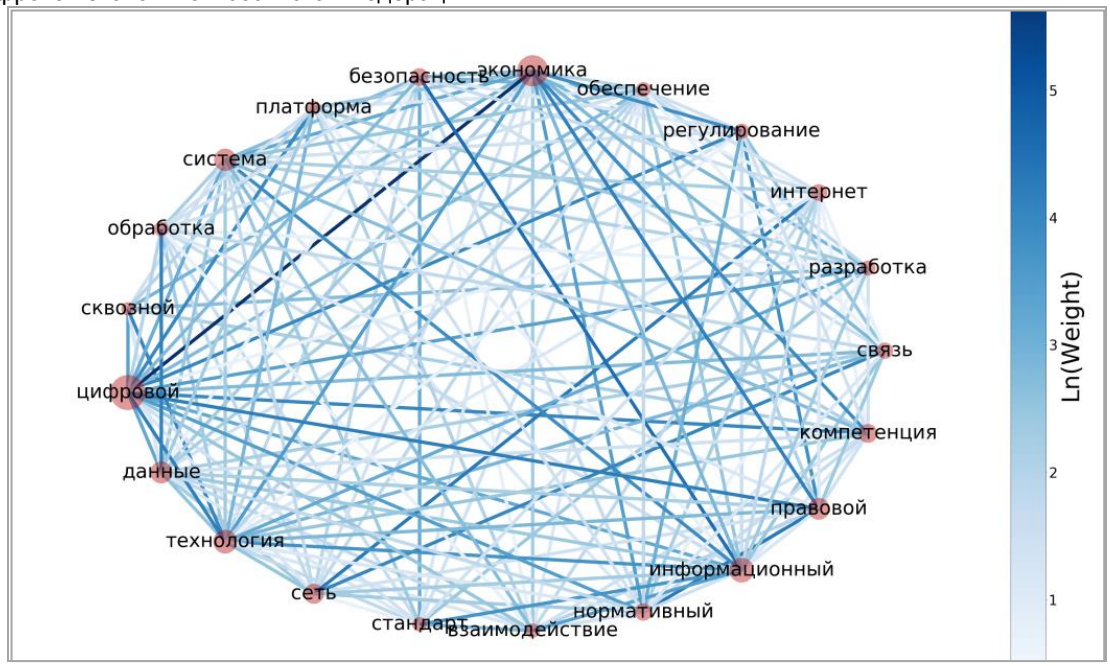

Рисунок 4. Граф семантических связей между ключевыми словами в программе «Цифровая экономика Российской Федерации»

Как видно из построенного графа, для данного документа можно выделить следующие ключевые термины: цифровая экономика, регулирование цифровой экономики, цифровая платформа, стандарт информационной безопасности, цифровые компетенции, сквозные технологии, сквозные цифровые

технологии, информационная безопасность, информационные технологии, сеть Интернет, сети связи, обработка данных, отечественные разработки, нормативный, правовой.

Аналогичным образом были проанализированы и остальные документы направления Цифровая экономика, после чего был сформирован итоговый список выделенных ключевых терминов (см. Приложение 1). Помимо выделения ключевых терминов как характеристики представленных документов, интересна динамика изменения состава связей. Так, например, если до утверждения программы «Цифровая экономика» термин «технология» связывался с такими терминами, как «новый», «перспективный», «коммуникационный», «информационный», «инновационный», то после – большинство связей отходит на второй план или исчезает, уступая место связям с «цифровой», «сквозной», «квантовый».

Сам термин «цифровой» тесно связан (помимо связи с «технология») с такими словами, как: «экономика», «платформа», «компетенция», «сквозной», «отечественный», «информационный», «безопасность», «РИД» (результаты интеллектуальной деятельности), «инфраструктура», «данные», «разработка», «внедрение», «взаимодействие», «нормативный», «правовой», «образовательный», «профессиональный», «кадры», и др.

Заключение

Алгоритм TextRank показывает адекватные результаты на обработке текстов сравнительно большого размера – максимальное число вершин графа достигалось для документа «Прогноз научно-технологического развития Российской Федерации на период до 2030 года» и равнялось 2359 (число уникальных слов документа после удаления стоп-слов). В ходе работы алгоритма были выявлены определенные сложности в сборке словосочетаний, что отмечено также и в работе (Усталов, 2012): «склеивание» слов в фразы происходит в полуручном режиме (необходимо участие эксперта), так как автоматическая сборка, во-первых, способна вывести только наборы слов-лемм, во-вторых, недостаточность ограничений на этапе сборки словосочетаний приводит к появлению заметного количества бессмысленных строк, собранных из вершин с большим весом. Основное решение проблемы состоит в учете вложенности и частоты встречаемости терминов в тексте. Несмотря на большое число методов извлечения ключевых слов, до настоящего времени не разработана последовательная методика обнаружения ключевых слов человеком (Ванюшкин, Гращенко, 2016). Экспериментально подтверждено, что эта операция выполняется людьми интуитивно и является личностно обусловленной (Мурзин, Штерн, 1991). Так, в случае извлечения ключевых слов из небольшого числа документов, результаты будут чувствительны к количеству слов, которое – было принято решение – отсечь от общего числа слов-кандидатов в ключевые слова.

Задача извлечения ключевых терминов из текста является основополагающей при изучении отраслевой терминологии. Построенные графы семантических связей наглядно показывают, как меняются в близких по времени создания правительственных документах лексические связи одних и тех же слов. Употребляя одни и те же слова в разных словосочетаниях и контекстах, авторы текстов уходят от четких границ терминов. Тем самым размывается их содержание, исчезает терминологический статус. Таким образом, терминология цифровой экономики на данный момент имеет характер несложившейся, незавершенной совокупности, что является следствием нерешенности проблем государственного и законодательного плана. С другой стороны, с точки зрения терминоведения, очевидно, что термины цифровой экономики обладают признаками «концептуализованной научной картины мира» (Ордокова, 2004) – отражают фундаментальность и глобальность как происходящих на данный момент, так и грядущих изменений всех сфер человеческой деятельности.

Литература

1. Beliga, S., Martincic-Ipsic, S., and Meštrović, A. An Overview of Graph-Based Keyword Extraction Methods and Approaches. Journal of Information and Organizational Sciences, 2015,39(1).
1. Bird, S., Klein, E., and Loper, E., Natural Language Processing with Python. O'Reilly Media, 2009. http://www.nltk.org/book_1ed/
2. Brin, S. and Page, L,The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998, 30(1–7).
3. Dobrolyubova E., Alexandrov O., Yefremov A., Is Russia Ready for Digital Transformation?. In: Alexandrov D., Boukhanovsky A., Chugunov A., Kabanov Y., Koltsova O. (eds) Digital Transformation and Global Society. DTGS 2017. Communications in Computer and Information Science, 2017, vol 745. Springer, Cham
4. Mihalcea, R. and Tarau, P. TextRank: Bringing Order into Texts. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004
5. Mihalcea, R. and Radev, D., Graph-based Natural Language Processing and Information Retrieval. Cambridge University Press., 2011
6. Page, S., Brin, S., Motwani, R, and Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford: Stanford University, 1998
7. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D., Multiword Expressions: A Pain in the Neck for NLP. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, CICLing '02, 2002, 1-15, London, UK.

8. Salton G., Yang C., On the Specification of Term Values in Automatic Indexing. Journal of Documentation, 1973, 29 (4), 351-372.
9. Ванюшкин А.С., Гращенко Л.А. Методы и алгоритмы извлечения ключевых слов, Новые информационные технологии в автоматизированных системах, 2016
10. Мурзин Л.Н., Штерн А.С. Текст и его восприятие. Свердловск: Издательство Уральского университета, 1991.
11. Ордокова, Ф.М., Принципы формирования отраслевой терминологии (На материале терминов экономики сельского хозяйства). Диссертация на соискание степени кандидата филологических наук, 2004
12. Усталов Д.А., Извлечение терминов из русскоязычных текстов при помощи графовых моделей. Теория графов и приложения: материалы конференции, 2012
13. Цынгуев Б.Т., Математические модели ранжирования вершин в графах коммуникационных сетей. Диссертация на соискание степени кандидата технических наук, Забайкальский государственный университет, 2015
14. Шереметьева С.О., Осминин П.Г., Методы и модели автоматического извлечения ключевых слов. Вестник Южно-Уральского государственного университета, 2015, 12 (1), 76–81.

Милкова Мария Александровна

Ключевые слова

Цифровая экономика, Программа «Цифровая экономика», графоориентированные методы, TextRank, семантические связи, интеллектуальный анализ текста

Milkova Maria, Extracting key terms from documents of Digital Economy direction: a graph-based approach

Keywords

Digital economy, Russian Digital Economy Program, graph-based approach, TextRank, semantic links, text mining

Abstract

The paper presents key terms extraction from the government documents issued in the period of 2013-2018 and linked to the Digital economy direction. One of the key interests of the analysis of government documents is to study them as primary source of digital economy terminology. The paper provides a brief review of the main approaches to key terms extraction and gives detailed description of one of the graph-based methods – a TextRank algorithm. The TextRank algorithm was tested on 13 government documents. The results of documents analysis are presented as weighted graphs of semantic links between keywords. Based on these words the lists of key terms are created for each document.

Приложение 1. Ключевые фразы по анализируемым правительственным документам

1. ***Прогноз научно-технологического развития Российской Федерации на период до 2030 года*** – программное обеспечение, транспортная система, программная система, новые материалы и технологии, новые разработки, перспективные технологии, космическая система, разработка новых технологий, транспортная модель, природная среда, энергетическая система, перспективные транспортные и космические системы, космические услуги, перспективные технологии переработки, продукты переработки, продукты и услуги.
2. ***Приоритетный проект «Совершенствование процессов организации медицинской помощи на основе внедрения информационных технологий*** – медицинская информационная система, система здравоохранения, электронный сервис, медицинская помощь, личный кабинет пациента «Мое здоровье» на ЕПГУ, электронная медицинская карта, информационная система.
3. ***О стратегии научно-технологического развития Российской Федерации*** – большой вызов, научно-технологический, инновационный научно-технологический, исследования и разработки, инновационный продукт, научно-технологический и инновационный, поддержка научно-технологического, поддержка исследований и разработок, экономика и общество, переход, наука и общество, наука и технологии, международный научно-технологический, обеспечение.
4. ***О стратегии развития информационного общества в Российской Федерации на 2017-2030 годы*** – сеть интернет, информационное общество, информационные технологии, информационные и коммуникационные инфраструктуры, обработка данных, сети связи, обеспечение доступа, распределение знаний, информационная безопасность, программное обеспечение, обеспечение безопасности (безопасного), информационная система, общество знаний, доступ к знаниям, доступ к сети.
5. ***Цифровая экономика в Российской Федерации*** – цифровая экономика, регулирование цифровой экономики, цифровая платформа, стандарт информационной безопасности, цифровые компетенции, сквозные технологии, сквозные цифровые технологии, информационная безопасность, информационные технологии, сеть интернет, сеть связи, обработка данных,

отечественные разработки, нормативный правовой стандарт безопасности разработки, правовое регулирование, компетенции цифровой экономики.

6. ***План мероприятий по направлению «Формирование исследовательских компетенций и технологических заделов программы «Цифровая экономика Российской Федерации»*** – цифровая экономика, цифровая платформа, цифровые платформы для исследований и разработок, сквозные технологии, разработка (в области) цифровых технологий, научно-технологические разработки, поддержка по исследованиям и разработкам, программная платформа, отбор сквозных технологий РИД (в разрезе) цифровых технологий, решения на базе сквозных технологий, цифровая экономика на базе, технологии на базе, квантовые технологии, разработка квантового.
7. ***План мероприятий по направлению*** "***Информационная инфраструктура» программы «Цифровая экономика Российской Федерации» -*** цифровая экономика, сеть интернет, сети связи, обработка данных, информационная инфраструктура, информационные технологии, требования к цифровым технологиям, информационная система, сеть 5G/IMT-2020, требования обеспечения информационной (безопасности) сети 5G/IMT-2020, Роскосмос.
8. ***План мероприятий по направлению «Информационная безопасность» программы «Цифровая экономика в Российской Федерации»*** – цифровая экономика, информационная безопасность, сеть интернет, стандарты безопасного информационного взаимодействия, стандарты информационной, стандарты безопасности, разработки, разработки (требований) к безопасности сетей, информационное взаимодействие в цифровой экономике, нормативный правовой, программное, отечественное программное, проведен анализ, Минкомсвязь.
9. ***План мероприятий по направлению «Нормативное регулирование» программы «Цифровая экономика в Российской Федерации»*** – цифровая экономика, нормативное регулирование, национальный стандарт, ЕАЭС (Евразийский экономический союз), исследование интернета, электронный, Минюст, Минфин, Минэкономразвития.
10. ***План мероприятий по направлению «Кадры и образование» программы «Цифровая экономика в Российской Федерации»*** – цифровая экономика, кадры и образование, компетенции цифровой экономики, персональная траектория, ключевые компетенции, образовательная организация, профессиональное образование, компетенции и профили, апробация модели, Минобрнауки.
11. ***Постановление Правительства РФ от 15 февраля 2018г. №158 «О программе «Фабрика проектного финансирования»*** – фабрика проектного финансирования, кредиты и займы, синдицированный кредит, транш синдицированного кредита, транш (предоставляемый), обязательства по облигациям, финансирование инвестиционных, Внешэкономбанк.
12. ***О национальных целях и стратегических задачах развития Российской Федерации на период до 2024 года*** – цифровая экономика, цифровые технологии, внедрение цифровых технологий, разработка и внедрение, система поддержки, внедрение системы, медицинский, строительство, разработка и внедрение национального.
13. ***Публичная декларация целей и задач Минэкономразвития России на 2018 год*** – цифровая экономика, социально-экономический, оценка эффективности, повышение эффективности, в рамках поддержки, правовой и нормативный, инвестиционный, экономическая система, электронный.

DOI: 10.34706/DE-2018-04-06