# РАЗРАБОТКА МЕТОДА ПОИСКА КЛЮЧЕВЫХ СЛОВ ДЛЯ ЗАДАЧИ ВИЗУАЛИЗАЦИИ НАУЧНЫХ СТАТЕЙ

Sandrikova Maria (mashasandrikova@gmail.com)
Zhuravlev Aleksei (aa_Zhur@mail.ru)
Sinevich Valeria (valeriya.sinevich@phystech.edu)
Korbut Daniel (korbut.daniel@gmail.com)
Iglina Alexandra (aiglina@gmail.com)
Fedotov Semen (simon23rus@gmail.com)

Moscow Institute of Physics and Technology (MIPT)
Moscow, Russia

**Keywords:** scientometrics, text visualisation, articles, ABBYY Compreno, natural language processing, keywords extraction.

**Annotation:** Scientometrics is an important field of study that helps to understand main trends in science and investigate its evolution. Often scientometrics operates with different statistics like a number of scholar articles and their citation rate. The results of such a research are usually presented in a text form.

Charts are an another popular way to represent the results of a research in scientometrics. They show the connections between fields of study, research centres and scholar articles. However, this approach has two problems:

- Static: visualisations do not depend on dynamic data and rapidly become obsolete.
- Generalisation: often it is not possible to get visualisation for a particular article such as the graph of references.

Our work is aimed to solve these problems using the smart keyword search in articles and visualisations that are simple for people's perception: interactive charts, maps and graphs.

# Введение

Наукометрия — важная дисциплина, помогающая понять основные тренды в науке и изучить ее эволюцию. Зачастую наукометрия оперирует с различной статистической информацией, например, количеством научных статей и цитируемостью, и результаты исследований представляются в текстовом виде.

Также популярно представление результатов исследований в виде графов, отображающих связи между областями науки, научными центрами или статьями. У этих представлений обычно есть две проблемы:

- Статичность: визуализации не основываются на динамических данных и быстро устаревают.
- Обобщение: обычно нет возможности получить визуализацию по отдельной статье, например, граф цитирования.

Наша работа направлена на решение этих проблем с помощью поиска ключевых слов в научных статьях и отображении информации в легко воспринимаемом формате: в виде карт и интерактивных графиков.

# Обзор темы

Визуальный подход к анализу массива научных работ не новый. Многие ученые пробовали таким образом подступиться к задачам наукометрии.

Например, создатели сайта scientometrics.ifpri.org [4] визуализировали научные работы организации CGIAR (Consultative Group for International Agricultural Research) в виде графа соавторства различных дочерних организаций, карты научных дочерних организаций и частотности ключевых слов. Их подход отличается от нашего статичностью данных, на основе которых построена визуализация, и более обобщенным подходом к авторству: они рассматривают не отдельных ученых, а организации, в которых они работают.

Еще одним примеров служит работа Jason Kessler, который в рамках своего проекта github.com/JasonKessler/scattertext [5] визуализировал понятия наиболее характеризующие выбранную область в виде цветовой карты.

# Подходы к отображению информации

Наш подход основан на четырех основных визуализациях:

- **Карта научных центров** отображает расположение статей по всему миру по заданной теме в выбранный промежуток времени. Чем больше круг города на карте, тем больше в университетах этого города было написано статей. С помощью этой визуализации можно отслеживать географическую историю развития той или иной области науки. Ключевые слова, отображаемые сверху карты, добавляют контекст этому процессу развития.
- **Граф цитируемости статей** показывает, как статьи ссылаются друг на друга. Из статьи идет ребро в другую статью, если в ней упоминается последняя. Такой способ отображения данных помогает глубже погрузиться в контекст научной работы с помощью быстрого доступа к ее источникам, находящимся не на первом уровне, то есть к источникам источников.

- **Граф ученых** позволяет отслеживать кластера ученых, писавших совместные работы. В графе соавторов между двумя авторами проведено ребро, если они участвовали в написании одной статьи. Такая визуализация также позволяет исследовать контекст статьи, но с другой точки зрения. Это может быть полезно для быстрого поиска дополнительных источников исследования среди других работ у соавторов.

- **Анализ трендов** сориентирует в последних тенденциях науки и их влиянии друг на друга. Необходимо задать области исследований, временной интервал и ключевые слова, чтобы график отобразил динамику их упоминания в научных статьях. Подобный инструмент может быть интересен, как аналитикам для исследования трендов в различных областях наук, так и историкам-исследователям для отслеживания того, как тренды сменяли друг друга или как тренды развивались совместно, поддерживая друг друга.

Все описанные выше визуализации можно попробовать на сайте meetarticles.me

# Извлечение смысловых ключевых слов

Поиск «смысловых ключевых слов» — важное преимущество инструмента. Смысловые ключевые слова — это привычные ключевые слова вместе с семейством своих синонимов, которые могут даже не упоминаться в статье. Такие ключевые слова позволяют провести более полный анализ трендов.

В нашем решении задача извлечения таких слов делится на 2 части: выделение ключевых слов в статье и их объединение в группы смысловых ключевых слов. Исследование было начато с первой задачи.

В первую очередь было проведено исследование наиболее популярных open-source решений для выделения ключевых слов. Были найдены такие готовые решения как Google KEA, надстройка над ним Maui, написанный пользователем github lvsh, класс KeywordFinder и TextRank. Ниже приведено более подробное описание каждого из перечисленных методов.

## KEA

Алгоритм поиска ключевых слов, основанный на обучении с учителем. Следующий шаг одинаков и для обучения, и для предсказаний:

- **Поиск кандидатов в ключевые слова** происходит в несколько шагов. Сначала весь текст разбивается на токены, знаки пунктуации обрабатываются как границы возможных фраз. Символы, которые не удалось токенизировать, удаляются. Возможные ключевые фразы ищутся среди последовательностей слов, обычно не длиннее трех, ни первое, ни последнее слово в которых не является стоп-словом. Список стоп-слов включает в себя 425 слов из разных частей речи и предоставлен авторами алгоритма. Также ключевые слова не могут быть именами собственными. Далее каждое слово в ключевых фразах приводится к нижнему регистру и обрезается до корня. Последнее делается с помощью известных алгоритмов поиска корня в слове, в данном случае используется стеммер Портера. Например, фраза «cutting elimination» превращается в «cut elim».

- **Обучение.** После того, как найдены возможные ключевые фразы для всех документов корпуса, алгоритм приступает к стадии вычисления признаков. В KEA используются простые признаки, такие как TF-IDF и первое появление слова.

  Эти признаки из непрерывных переводятся в дискретные методом построения гистограммы распределения для последующего обучения. На полученных признаках обучается наивный байесовский классификатор.

- **Предсказание**. Для результатов предсказания производится пост-обработка, которая для каждого документа выбирает несколько фраз с самой высокой вероятностью.

## Maui

Maui – это расширение над KEA, написанное теми же авторами. Отличается он улучшенным выделением признаков. Используются контрольные словари и уже индексированные корпусы текстов, такие как Wikipedia, Medical Vocabulary, Energy Physics Thesaurus и другие. Также в Maui используются другие признаки, связанные с размером документа, длиной фразы и семантическими связями между словами.

## lvsh/Keywordfinder

Данный алгоритм практически полностью повторяет KEA, но для извлечения признаков автор использует библиотеку features для python, а в качестве классификатора – логистическую регрессию из стандартного пакета sklearn.

## TextRank

TextRank — приложение алгоритма PageRank к задачам обработки естественного языка. Решение состоит из следующих шагов:

1. Строится взвешенный неориентированный граф $(V; E)$ на основе исходного текста. Ребро в графе соответствует наличию семантической информации между двумя словами.
2. Приближённо вычисляется PageRank для построенного графа.
3. Вершины с наибольшим значением веса TextRank считаются ключевыми словами.

Вычисление значение TextRank производится по итеративной формуле:

$$TR(t_i) = (1 - d) + d \cdot \sum_{(t_j, t_i) \in E)} \frac{w_{ji}}{\sum_{(t_k, t_j) \in E} w_{kj}} \cdot TR(t_j)$$

Здесь d – фактор затухания (параметр алгоритма), $w_{ij}$ – вес ребра $(t_i, t_j)$.

Более подробное описание метода можно прочитать в авторской статье [1].

Для тестирования метода использовалась реализация алгоритма с открытым кодом под названием Summa.

## Подход, основанный на совместной встречаемости слов

Кроме open-source методов был произведен поиск методов, не воплощенных в открытом исходном коде, однако имеющих достаточно подробные описания в виде статей. После изучения этих методов было принято решение перейти к реализации наиболее перспективных из них. На данный момент реализован и хорошо протестирован лишь один из таких методов: статистический метод Y. Matsuo и M. Ishizuka [3]. Ниже приведено подробное описание реализации метода:

- Производится препроцессинг, состоящий из 2-х фаз: удаление стоп-слов и стемминг при помощи алгоритма Портера (в python использовался модуль stopwords из nltk.corpus, а также SnowballStemmer из модуля nltk.stem.snowball).

- Выбираются 30% наиболее часто встречающихся слов текста, преобразованных алгоритмом стемминга после 1-го шага.
- Затем для выбранных слов составляется матрица совместной встречаемости: в ячейке $a_{i,j}$ хранится количество предложений в исходном тексте, в которых $i$-ое и $j$-ое слово встретились вместе.
- Данные слова кластеризуются по 2-м статистическим критериям: дивергенция Дженсона-Шеннона и взаимная информация. В качестве признакового описания слова берётся строка из матрицы совместной встречаемости. Два слова попадают в один кластер тогда и только тогда, когда значение дивергенции Дженсона-Шеннона больше $0.95 \cdot log(2.0)$ и значение взаимной информации больше $log(2.0)$
- Вычисляются 2 характеристики для каждого кластера: $n_c$ – количество слов текста, которые встречаются в предложениях, содержащих хотя бы одно слово кластера $c$, а также $p_c = n_c/N_{tota}$l, где $N_{total}$ – общее число слов текста после стемминга.
- После стемминга для каждого слова считаем статистику по следующей формуле:

$$\chi'^2(w) = \sum_c \frac{(freq(w,c) - n_w p_c)^2}{n_w p_c} - \max_c \frac{(freq(w,c) - n_w p_c)^2}{n_w p_c},$$

  где $freq(w,c)$ – количество предложений, в которых встречаются и терм $w$, и один из термов кластера $c$.
- На финальном шаге выбирается необходимое количество слов с наибольшими значениями статистики $\chi'^2$, которые и будут ключевыми словами текста.

# Результаты

Обучение и тестирование производилось на датасете THESES80 [6], в котором собраны различные научные работы с размеченными ключевыми словами (суммарно 86 статей). Из текстов оставлены были только абстракты. Выборка разделена на тестовую и обучающую случайным образом в соотношении 1:3.

В качестве целевых метрик рассматриваются precision и recall для класса ключевых слов. В контексте задачи precision показывает, какой процент слов был правильно классифицирован. Recall показывает какая доля ключевых слов была найдена алгоритмом. Измерялась также среднее геометрическое этих 2 метрик: $F1$-мера. Значения метрик измерялись отдельно для каждой статьи из тестового множества и были усреднены по всем статьям.

Стоит также отметить, что некоторые методы выделяют исключительно ключевые слова, а некоторые способны выделять ключевые фразы. Разметка датасета также иногда содержит ключевые фразы. Для более честного сравнения мы производили отдельные тесты, в которых ключевые фразы разбивались на отдельные ключевые слова. Ниже приводится таблица с результатами.

| Метод | Фразы? | Mean Precision | Mean Recall | Mean F1 |
|---|---|---|---|---|
| KEA | да | 0.2172 | 0.1023 | 0.0755 |
| Maui | да | 0.0781 | 0.1058 | 0.0899 |
| KEA | нет | 0.0620 | 0.2059 | 0.2026 |
| Maui | нет | 0.3000 | 0.2200 | 0.2500 |
| KeywordFinder | нет | 0.1517 | 0.1879 | 0.1611 |
| TextRank | нет | 0.1005 | 0.1460 | 0.1097 |
| Co-occurrence Statistics | нет | 0.1001 | 0.1080 | 0.1031 |

Отметим, что качество даже у лучших методов довольно низкое, что можно связать с несколькими причинами:

- Небольшой размер выборки, как обучающей, так и тестовой: нет репрезентативности, оценка качества сильно смещена.
- Используемые в методах признаки слабо представительны и недостаточны для решения задачи выделения ключевых слов.
- Общая сложность задачи: разметка ключевых слов производится субъективными асессорами, модель оценки которых тяжело предсказуема методами машинного обучения.

# Планы по развитию проекта

Исследование показывает, что наиболее популярные методы для выделения ключевых слов не дают достаточного качества. Планируется продолжить исследование методов, описанных в различных статьях, но не имеющих публично-доступной реализации.

В качестве методов машинного обучения вместо рассмотренных здесь наивного байеса и логистической регрессии будут проверены другие модели, такие как XGBoost, Conditional Random Fields и другие. Также немалое внимание планируется уделить применению технологий глубинного обучения для решения задач поиска ключевых слов. В планы экспериментов входят классические нейронные сети и другие архитектуры: автоэнкодеры и рекуррентные нейронные сети.

Переход от обычных ключевых слов к смысловым также запланирован на будущее с помощью выделения семантического класса слова. Для решения этой задачи мы планируем применить технологию ABBYY Compreno.

# Заключение

В заключении хотелось бы отметить, что описанный в нашей статье подход к отображению данных с помощью графов, карт и графиков и анализ трендов на основе смысловых ключевых слов можно использовать для визуализации не только баз научных статей, но и для любых других текстовых документов, например, патентов или статей в СМИ.

# Список литературы


[1] R. Mihalcea, P. Tarau *TextRank: Bringing Order into Texts.* 2004

[2] lena Medelyan *Human-competitive automatic topic indexing.* 2009

[3] Y. Matsuo, M. Ishizuka *Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information.* 2003

[4] Koo, Jawoo, Glenn Hyman, Silvia-Elena Castaño, and Grant McKenzie. *CGIAR Scientometric Trends.* International Food Policy Research Institute, scientometrics.ifpri.org. Accessed on 14 April 2017

[5] Jason Kessler. *Scattertext.* github.com/JasonKessler/scattertext. Accessed on 14 April 2017

[6] THESES80 scientific artiles dataset github.com/qcri/maui-indexer/blob/master/data/models/theses80. Accessed on 25 May 2017