

КОНВЕЙЕР ОБРАБОТКИ ДАННЫХ НАУЧНЫХ ПУБЛИКАЦИЙ ДЛЯ ПОСТРОЕНИЯ ГРАФА СЕМАНТИЧЕСКОГО СХОДСТВА

Ушаков С.Н.

Томский политехнический университет, ИШИТР, гр. А1-39, e-mail: wowman9721@gmail.com

Аннотация

В работе предлагается метод построения конвейера для обработки данных научных публикаций с целью возможности их аннотирования методами машинного обучения и построения семантических графов сходства.

Ключевые слова: граф, конвейер, парсинг, TF-IDF, API

Введение

В современном информационном обществе объем научной литературы неуклонно растет, при этом ученые и специалисты сталкиваются с задачей извлечения ценной информации из этого огромного объема данных. Однако, доступность информации не всегда означает ее усвоение и использование в полной мере [1]. Именно поэтому разработка методов для эффективного извлечения знаний из научных публикаций и оценка их сходства становится важной и актуальной задачей. Анализ научных статей и выявление их сходства имеет ряд важных применений в научном и инженерном сообществе. Это может включать в себя автоматическое создание обзоров литературы, выявление тенденций в исследованиях, выявление потенциальных коллабораций между учеными и многое другое [2]. Кроме того, с ростом объема научной информации становится все более сложно ориентироваться в множестве публикаций и определять наиболее релевантные исследования по конкретной теме.

Целью данной работы является разработка алгоритма для конвейера обработки данных научных публикаций с целью возможности дальнейшего их аннотирования и построения семантических графов сходства.

Описание построения конвейера обработки данных

Конвейер обработки данных научных публикаций включает в себя следующие шаги:

1. Поиск научных публикаций по ключевым словам.
2. Парсинг найденных публикаций и извлечение содержимого аннотаций
3. Оценка сходства публикаций по аннотациям и по всему содержанию

На рис. 1 изображена схема предложенного конвейера для обработки данных научных публикаций.

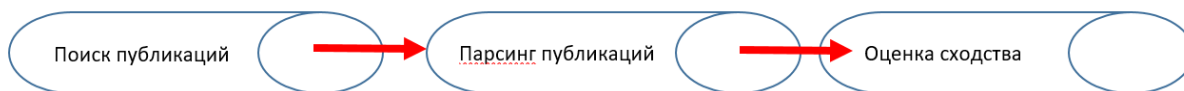


Рис. 1. Схематичное представление конвейера обработки данных научных публикаций

Поиск научных публикаций осуществляется по ключевым словам из специального датасета по синтезу материалов. Для реализации поиска научных публикаций использовалась площадка Google Scholar, которая имеет API в качестве Python-библиотеки под названием scholarly. После выполнения поиска scholarly возвращает результаты в виде объектов Python, которые можно анализировать и обрабатывать для извлечения необходимой информации. Таким образом, каждый результат найденных с помощью scholarly публикаций включает в себя: название статьи, имена авторов, год публикации, количество цитирований и URL-адрес, по которому можно обратиться к данной публикации. Результаты автоматического поиска с помощью библиотеки scholarly представлены на рис. 2.

Title: Synthesis of single-phase high-entropy carbide powders
 Author: L Feng, WG Fahrenholtz, GE Hilmas, Y Zhou
 Publication Year: N/A
 Citation Count: 188
 URL: <https://www.sciencedirect.com/science/article/pii/S135964621830678X>

Title: High-entropy carbide ceramics: A perspective review
 Author: Z Wang, ZT Li, SJ Zhao, ZG Wu
 Publication Year: N/A
 Citation Count: 31
 URL: <https://link.springer.com/article/10.1007/s42864-021-00085-7>

Title: High-entropy carbide: A novel class of multicomponent ceramics
 Author: J Zhou, J Zhang, F Zhang, B Niu, L Lei, W Wang
 Publication Year: N/A
 Citation Count: 321
 URL: <https://www.sciencedirect.com/science/article/pii/S0272884218321576>

Title: High entropy carbide ceramics from different starting materials
 Author: XF Wei, JX Liu, F Li, Y Qin, YC Liang
 Publication Year: N/A
 Citation Count: 168
 URL: <https://www.sciencedirect.com/science/article/pii/S0955221919302304>

Title: Phase stability and mechanical properties of novel high entropy transition metal carbides
 ...
 Publication Year: N/A
 Citation Count: 24
 URL: <https://www.sciencedirect.com/science/article/pii/S0955221921003538>

Рис. 2. Результат автоматического поиска научных публикаций

Процесс парсинга осуществляется путем разбиения публикации на заголовки и извлечения метаданных раздела «Abstract». Результат парсинга представлен на рис. 3.

Abstract
 Chemical vapor deposition (CVD) is a vacuum deposition method used to produce high-quality and high-performance solid materials. The process is often used in the semiconductor industry to produce thin films. Microfabrication processes widely use CVD to deposit materials in various forms, including monocrystalline, polycrystalline, amorphous, and epitaxial. These materials include silicon (dioxide, carbide, nitride, oxynitride), carbon (fiber, nanofibers, nanotubes, diamond, and graphene), fluorocarbons, filaments, tungsten, titanium nitride, and various high-k dielectrics. Chemical deposition takes advantage of the chemical reaction where the product self-assembles and deposits on a suitable substrate. Chemical deposition is commonly used for generating thin nanostructured blend films of crystalline inorganic materials, such as ZnS, CuSe, InS, CdS, etc. Depending on the deposition conditions, several terms have been used, such as chemical bath deposition, CVD, and ECD. Depending on the material and the deposition conditions, different surface morphologies have been obtained, from nanopins to nanotubes to nanorods. Reagents in sedimentary reactions are usually water-soluble ionic compounds. When these compounds are dissolved in water, they separate from each other to form anion and cation ions. If a cation of one compound forms an insoluble compound with an anion of another compound, precipitation occurs. Applications of this method include the removal of heavy metals and anions from wastewater, reducing water hardness, and metal recovery. Synthesis processes occur by chemical deposition based on deposition reactions (substitution), co-precipitation, oxidation-reduction, thermolysis, hydrolysis, polymerization, and condensation. The control of various variables in a synthetic system plays an important role in controlling particle size and morphology. The products of sedimentary processes under various synthetic conditions range from coarse crystals to nanostructured colloidal particles. Co-precipitation chemical methods allow the synthesis of metal nanoparticles, metal oxides, as well as many metal semiconductor compounds. Also, a wide range of properties and characteristics can be achieved by changing the synthesis conditions. The basis of these methods is the preparation of products from soluble precursors using different systems such as electrochemical equipment, microwave radiation, ultrasound, and high-energy beams.

Рис. 3. Результат парсинга

Результаты оценки сходства

Для оценки сходства публикаций были использованы метрики TF-IDF, расстояние Джакара и косинусное сходство [3-4]. Оценка осуществлялась по сходству Abstract и по всему содержанию публикаций. В таблице 1 представлены результаты оценки сходства по аннотациям.

Таблица 1

Результаты сходства по Abstract

TF-IDF	Косинусное расстояние	Расстояние Джакара
0.51	0.31	0.32

По результатам 1 можно сделать вывод, что сходство между аннотациями документов, оцененное с использованием TF-IDF, наибольшее (0.51), что может указывать на значительное пересечение важных терминов или фраз в этих документах с учетом их веса и значимости в контексте всей коллекции аннотаций документов. Косинусное расстояние ближе к 1 (0.31), что указывает то, что направления векторов TF-IDF для этих документов достаточно близки друг к другу, что может быть

связано с более схожими темами или контекстом этих аннотаций. Расстояние Жаккара показывает, что в текстах аннотаций есть определенное пересечение по используемым терминам, но в большинстве своем они уникальны, что может говорить о различных экспериментах, описанных в данных публикациях. В таблице 2 представлены результаты сходства по полному содержанию научных публикаций.

Таблица 2

Результаты сходства по полному содержанию

TF-IDF	Косинусное расстояние	Расстояние Джакара
0.44	0.29	0.28

Результаты таблицы 2 указывают на то, что внутри текстов статей существует больше уникальных слов или концепций, что снижает степень сходства на уровне TF-IDF. Значение косинусного сходства также ниже, чем при анализе аннотаций, что указывает на более высокое сходство в содержании абстрактов, чем в полном тексте публикаций. Сходство Жаккара также ниже, чем при анализе абстрактов, но ближе к косинусному расстоянию. Это может указывать на более высокое пересечение используемых терминов внутри текстов статьи по сравнению с анализом только абстрактов.

Заключение

В результате работы был предложен конвейер обработки данных научных публикаций для возможности дальнейшего аннотирования с целью проектирования различных баз знаний и построения семантических графов сходства. С помощью данной методики было проанализировано около 300 статей по различным тематикам одного направления. В качестве направления для анализа был выбран синтез материалов.

В результате значительное отличия результатов сходства можно сделать вывод, что анализируемые публикации достаточно близки по тематикам исследований, но значительно различаются по предлагаемым методикам и используемым технологиям.

Список использованных источников

1. Jones, K.S. A statistical interpretation of term specificity and its application in retrieval / K.S. Jones // Journal of Documentation. – 2021. – Vol. 60(5). – P. 493-502.
2. Воронцов К.В. Многокритериальные и многомодальные вероятностные тематические модели коллекций текстовых документов / К.В. Воронцов, А.А. Потапенко, А.И. Фрей, М.А. Апишев, Н.В. Дойков, А.В. Шапулин, Н.А. Чиркова // 10-я Междунар. конф. ИОИ-2022: Тезисы докладов. – 2022. – С. 198.
3. Шеннон К. Работы по теории информации и кибернетики / К. Шеннон; пер. с англ. – М.: Иностранная литература, –2020. – С. 669–686. (Shannon C.E. Prediction and entropy of printed English. BSTJ –1951; – V. 30(1) – P. 50-64)
4. Царьков С.В. Автоматическое выделение ключевых фраз для построения словаря терминов в тематических моделях коллекций текстовых документов / С.В. Царьков // Естественные и технические науки. – 2021. – № 6. – С. 456-464. – ISSN 1684-2626