Разработка интеллктуальной системы анализа патентов химической отрасли для представления данных в структурированном виде

Студент 2-го курса: Кайда Анатолий Сергеевич
Научный руководитель: Глинский Андрей Владимирович

Формулировка проблемы

- Количество ежегодно публикуемых статей и патентов в химии растет экспоненциально
- Процесс работы с в патентной и литературной информацией по-прежнему остается в значительной степени ручным
- Сложность навигации в большом объеме литературы приводит к тому, что важные научные открытия остаются незамеченными в течение длительного времени

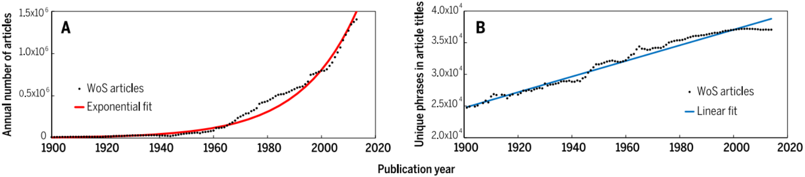


Рис. 1: (A) Годовой выпуск научных статей, индексированных в базе данных WoS. (B) Рост идей, охватываемых статьями, индексированными в WoS. Это было определено путем подсчета уникальных заглавных фраз (концепций) в фиксированном количестве статей.

Формулировка проблемы

1. Ananikov V. Top 20 Influential AI-Based Technologies in Chemistry. Chemistry, 2024.

В последние годы набирает обороты «цифровизация» химии.

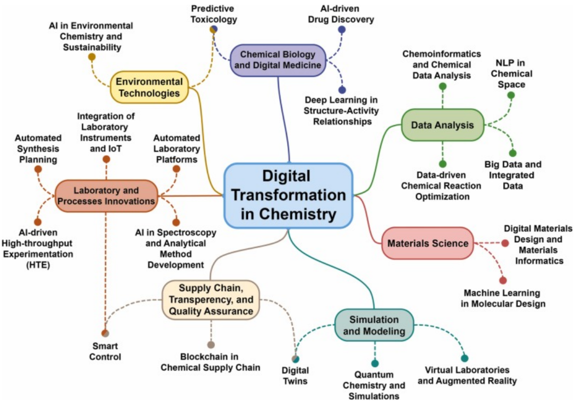
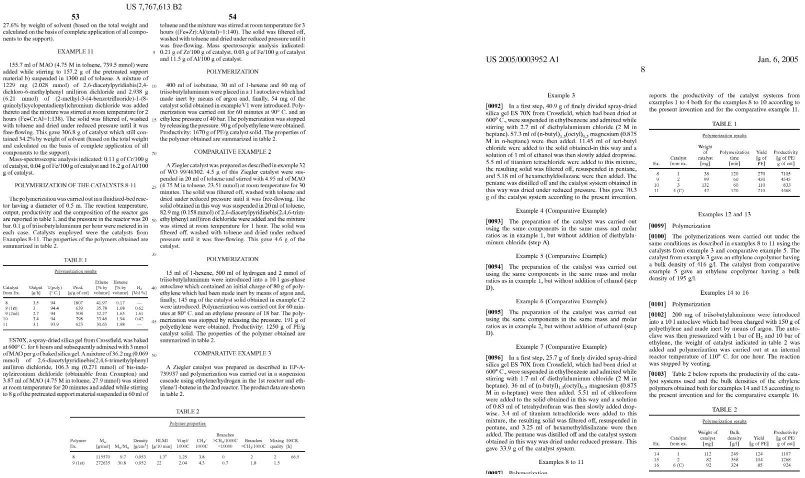

Digital Transformation in Chemistry
Chemical Biology and Digital Medicine
Predictive Toxicology
AI-driven Drug Discovery
Deep Learning in Structure-Activity Relationships
Environmental Technologies
AI in Environmental Chemistry and Sustainability
Data Analysis
Chemoinformatics and Chemical Data Analysis
NLP in Chemical Space
Big Data and Integrated Data
Data-driven Chemical Reaction Optimization
Laboratory and Processes Innovations
Automated Synthesis Planning
Integration of Laboratory Instruments and IoT
Automated Laboratory Platforms
AI-driven High-throughput Experimentation (HTE)
AI in Spectroscopy and Analytical Method Development
Smart Control
Materials Science
Digital Materials Design and Materials Informatics
Machine Learning in Molecular Design
Supply Chain, Transparency, and Quality Assurance
Blockchain in Chemical Supply Chain
Simulation and Modeling
Digital Twins
Quantum Chemistry and Simulations
Virtual Laboratories and Augmented Reality

Рис. 2: Связь технологий на основе ИИ с более широкими темами в зависимости от их применения.

Формулировка проблемы



US 7,767,613 B2

53

27.6% by weight of solvent (based on the total weight and calculated on the basis of complete application of all components to the support).

EXAMPLE 11

155.7 ml of MAO (4.75 M in toluene, 739.5 mmol) were added while stirring to 157.2 g of the pretreated support material b) suspended in 1300 ml of toluene. A mixture of 1229 mg (2.028 mmol) of 2,6-diacetylpyridinbis(2,4-dichloro-6-methylphenyl anil)iron dichloride and 2.938 g (6.21 mmol) of (2-methyl-3-(4-benzotrifluoride)-1-(8-quinolyl)cyclopentadienyl)chromium dichloride was added thereto and the mixture was stirred at room temperature for 2 hours (Fe+Cr:Al=1:138). The solid was filtered of, washed with toluene and dried under reduced pressure until it was free-flowing. This gave 306.8 g of catalyst which still contained 34.2% by weight of solvent (based on the total weight and calculated on the basis of complete application of all components to the support).

Mass-spectroscopic analysis indicated: 0.11 g of Cr/100 g of catalyst, 0.04 g of Fe/100 g of catalyst and 16.2 g of Al/100 g of catalyst.

POLYMERIZATION OF THE CATALYSTS 8-11

The polymerization was carried out in a fluidized-bed reactor having a diameter of 0.5 m. The reaction temperature, output, productivity and the composition of the reactor gas are reported in table 1, and the pressure in the reactor was 20 bar. 0.1 g of triisobutylaluminum per hour were metered in in each case. Catalysts employed were the catalysts from Examples 8-11. The properties of the polymers obtained are summarized in table 2.

TABLE 1

Polymerization results

Catalyst from Ex.	Output [g/h]	T(poly) [° C.]	Prod. [g/g of cat]	Ethene [% by volume]	Hexene [% by volume]	H_2 [Vol %]
8	3.5	94	1807	41.97	0.17	—
9 (1st)	3	94.4	639	35.78	1.68	0.62
9 (2nd)	2.7	94	504	32.27	1.65	1.61
10	3.4	94	798	33.46	1.94	0.42
11	3.1	93.9	623	30.63	1.98	—

ES70X, a spray-dried silica gel from Crossfield, was baked at 600° C. for 6 hours and subsequently admixed with 3 mmol of MAO per g of baked silica gel. A mixture of 36.2 mg (0.069 mmol) of 2,6-diacetylpyridinebis(2,4,6-trimethylphenyl anil)iron dichloride, 106.3 mg (0.271 mmol) of bis-indenylzirconium dichloride (obtainable from Crompton) and 3.87 ml of MAO (4.75 M in toluene, 27.9 mmol) was stirred at room temperature for 20 minutes and added while stirring to 8 g of the pretreated support material suspended in 60 ml of

54

toluene and the mixture was stirred at room temperature for 3 hours ((Fe+Zr):Al(total)=1:140). The solid was filtered off, washed with toluene and dried under reduced pressure until it was free-flowing. Mass spectroscopic analysis indicated: 0.21 g of Zr/100 g of catalyst, 0.03 g of Fe/100 g of catalyst and 11.5 g of Al/100 g of catalyst.

POLYMERIZATION

400 ml of isobutane, 30 ml of 1-hexene and 60 mg of triisobutylaluminum were placed in a 1 l autoclave which had made inert by means of argon and, finally, 54 mg of the catalyst solid obtained in example V1 were introduced. Polymerization was carried out for 60 minutes at 90° C. and an ethylene pressure of 40 bar. The polymerization was stopped by releasing the pressure. 90 g of polyethylene were obtained. Productivity: 1670 g of PE/g catalyst solid. The properties of the polymer obtained are summarized in table 2.

COMPARATIVE EXAMPLE 2

A Ziegler catalyst was prepared as described in example 32 of WO 99/46302. 4.5 g of this Ziegler catalyst were suspended in 20 ml of toluene and stirred with 4.95 ml of MAO (4.75 M in toluene, 23.51 mmol) at room temperature for 30 minutes. The solid was filtered off, washed with toluene and dried under reduced pressure until it was free-flowing. The solid obtained in this way was suspended in 20 ml of toluene, 82.9 mg (0.158 mmol) of 2,6-diacetylpyridinebis(2,4,6-trimethylphenyl anil)iron dichloride were added and the mixture was stirred at room temperature for 1 hour. The solid was filtered off, washed with toluene and dried under reduced pressure until it was free-flowing. This gave 4.6 g of the catalyst.

POLYMERIZATION

15 ml of 1-hexene, 500 ml of hydrogen and 2 mmol of triisobutylaluminum were introduced into a 10 l gas-phase autoclave which contained an initial charge of 80 g of polyethylene which had been made inert by means of argon and, finally, 145 mg of the catalyst solid obtained in example C2 were introduced. Polymerization was carried out for 60 minutes at 80° C. and an ethylene pressure of 18 bar. The polymerization was stopped by releasing the pressure. 191 g of polyethylene were obtained. Productivity: 1250 g of PE/g catalyst solid. The properties of the polymer obtained are summarized in table 2.

COMPARATIVE EXAMPLE 3

A Ziegler catalyst was prepared as described in EP-A-739937 and polymerization was carried out in a suspension cascade using ethylene/hydrogen in the 1st reactor and ethylene/1-butene in the 2nd reactor. The product data are shown in table 2.

TABLE 2

Polymer properties

Polymer Ex.	M_w [g/mol]	M_w/M_n	Density [g/cm^3]	HLMI [g/10 min]	Vinyl/ 1000C	CH_3/ 1000C	Branches >CH_3/1000C <10000	Branches >CH_3/1000C	Mixing quality	ESCR [h]
8	115570	9.7	0.953	1.3^b	1.25	3.8	0	2	2	66.5
9 (1st)	272635	30.8	0.952	22	2.04	4.3	0.7	1.8	1.5	

(a)

US 2005/0003952 A1 8 Jan. 6, 2005

Example 3

[0092] In a first step, 40.9 g of finely divided spray-dried silica gel ES 70X from Crossfield, which had been dried at 600° C., were suspended in ethylbenzene and admixed while stirring with 2.7 ml of diethylaluminum chloride (2 M in heptane). 57.3 ml of (n-butyl)$_{1.5}$(octyl)$_{0.5}$ magnesium (0.875 M in n-heptane) were then added. 11.45 ml of tert-butyl chloride were added to the solid obtained-in this way and a solution of 1 ml of ethanol was then slowly added dropwise. 5.5 ml of titanium tetrachloride were added to this mixture, the resulting solid was filtered off, resuspended in pentane, and 5.18 ml of hexamethyldisilazane were then added. The pentane was distilled off and the catalyst system obtained in this way was dried under reduced pressure. This gave 70.3 g of the catalyst system according to the present invention.

Example 4 (Comparative Example)

[0093] The preparation of the catalyst was carried out using the same components in the same mass and molar ratios as in example 1, but without addition of diethylaluminum chloride (step A).

Example 5 (Comparative Example)

[0094] The preparation of the catalyst was carried out using the same components in the same mass and molar ratios as in example 1, but without addition of ethanol (step D).

Example 6 (Comparative Example)

[0095] The preparation of the catalyst was carried out using the same components in the same mass and molar ratios as in example 2, but without addition of ethanol (step D).

Example 7 (Comparative Example)

[0096] In a first step, 25.7 g of finely divided spray-dried silica gel ES 70X from Crossfield, which had been dried at 600° C., were suspended in ethylbenzene and admixed while stirring with 1.7 ml of diethylaluminum chloride (2 M in heptane). 36 ml of (n-butyl)$_{1.5}$(octyl)$_{0.5}$ magnesium (0.875 M in n-heptane) were then added. 5.51 ml of chloroform were added to the solid obtained in this way and a solution of 0.83 ml of tetrahydrofuran was then slowly added dropwise. 3.4 ml of titanium tetrachloride were added to this mixture, the resulting solid was filtered off, resuspended in pentane, and 3.25 ml of hexamethyldisilazane were then added. The pentane was distilled off and the catalyst system obtained in this way was dried under reduced pressure. This gave 33.9 g of the catalyst system.

Examples 8 to 11

[0097] Polymerization

reports the productivity of the catalyst systems from examples 1 to 4 both for the examples 8 to 10 according to the present invention and for the comparative example 11.

TABLE 1

Polymerization results

Ex.	Catalyst from ex.	Weight of catalyst [mg]	Polymerization time [min]	Yield [g of PE]	Productivity [g of PE/ g of cat]
8	1	38	120	270	7105
9	2	99	60	450	4545
10	3	132	60	110	833
11	4 (C)	47	120	210	4468

Examples 12 and 13

[0099] Polymerization

[0100] The polymerizations were carried out under the same conditions as described in examples 8 to 11 using the catalysts from example 3 and comparative example 5. The catalyst from example 3 gave an ethylene copolymer having a bulk density of 416 g/l. The catalyst from comparative example 5 gave an ethylene copolymer having a bulk density of 195 g/l.

Examples 14 to 16

[0101] Polymerization

[0102] 200 mg of triisobutylaluminum were introduced into a 10 l autoclave which had been charged with 150 g of polyethylene and made inert by means of argon. The autoclave was then pressurized with 1 bar of H_2 and 10 bar of ethylene, the weight of catalyst indicated in table 2 was added and polymerization was carried out at an internal reactor temperature of 110° C. for one hour. The reaction was stopped by venting.

[0103] Table 2 below reports the productivity of the catalyst systems used and the bulk densities of the ethylene polymers obtained both for examples 14 and 15 according to the present invention and for the comparative example 16.

TABLE 2

Polymerization results

Ex.	Catalyst from ex.	Weight of catalyst [mg]	Bulk density [g/l]	Yield [g of PE]	Productivity [g of PE/ g of cat]
14	1	112	249	124	1107
15	2	82	358	104	1268
16	6 (C)	92	324	85	924

(b)

Рис. 3: Примеры входных данных

Цели и задачи

Цель работы: Создать цифрового «ассистента» на основе большой языковой модели для извлечения структурированных данных из патентной документации

Задачи:

- Выбрать домен для проведения исследования и провести релевантный патентный поиск и создать БД документов для дальнейшего извлечения информации
- Провести обзор современных фреймворков и технологий для работы с БЯМ
- Создать агентов на основе БЯМ способных решать следующие задачи:
  - Сегментация и фильтрация текста
  - Классификация текста и выделение информации о синтетических процедурах
  - Запрос к БЯМ
  - Формирование датасета

Обзор литературы

2. Ramos M.C., Collison C.J., White A.D. A Review of Large Language Models and Autonomous Agents in Chemistry: arXiv:2407.01603. arXiv, 2024.

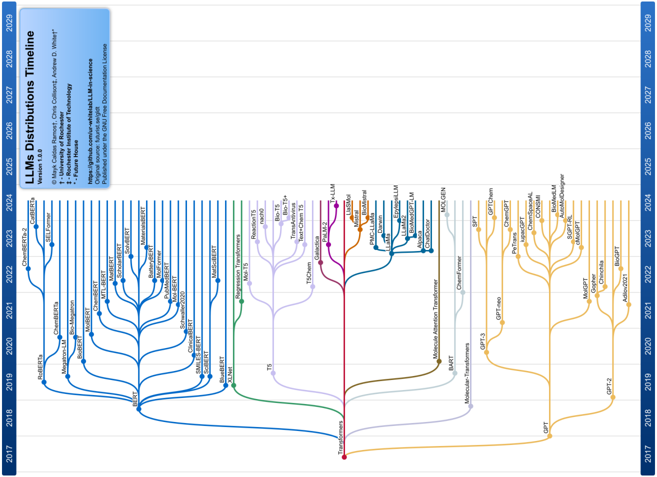


Рис. 4: Иллюстрация хронологической эволюции больших языковых моделей.

Обзор литературы

3. Lála J. et al. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research: arXiv:2312.07559. arXiv, 2023.

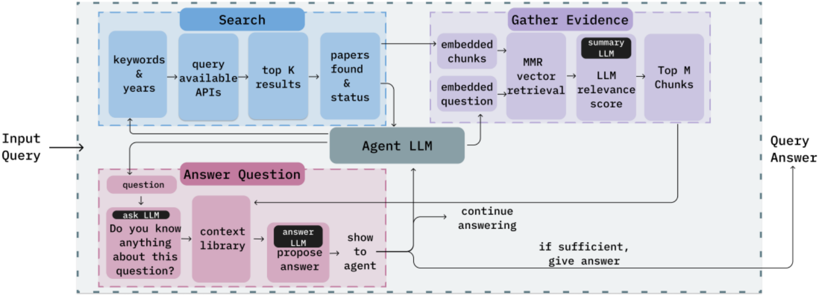


Рис. 5: PaperQA — это агент, который преобразует вопрос в ответ с указанием источников. Агент использует три инструмента: поиск, сбор данных и ответ на вопрос. Инструменты позволяют ему находить и анализировать соответствующие полнотекстовые исследовательские работы, определять конкретные разделы в работе, которые помогают ответить на вопрос, суммировать эти разделы с контекстом вопроса (называемые доказательствами), а затем генерировать ответ на основе доказательств.

Обзор литературы

4. Zheng Z. et al. ChatGPT Chemistry Assistant for Text Mining and Prediction of MOF Synthesis.

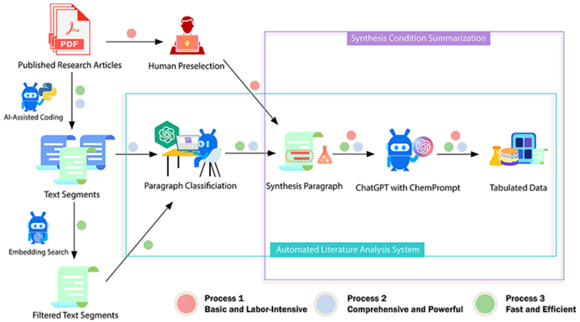


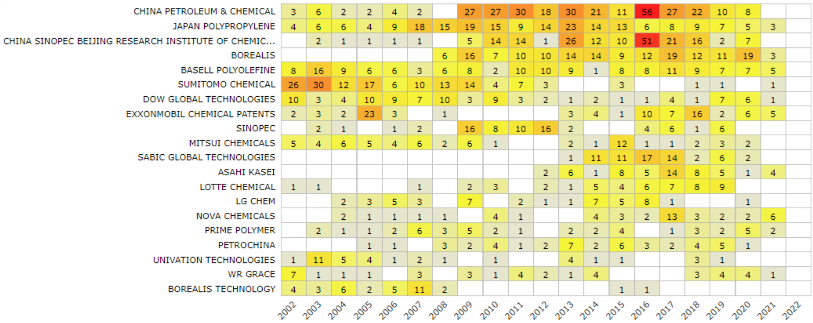
Рис. 6: ChatGPT Chemistry Assistant ChatGPT и ChemPrompt для эффективного анализа текста и обобщения условий синтеза MOF из разнообразного набора опубликованных исследовательских статей.

Текущие результаты и планы

Catalyst type:	PE: ZN on SiO2
KEY WORDS:	T/A/C/D: (((((Ziegler-Natta catalyst+) or (silica?supported Ziegler-Natta catalyst+)) and (titanium +chloride)) and gas-phase) and (PE or polyethylene))
RESTRICTION:	ALL
number of documents before relevance is determined	There were about 2491 documents (ORBIT)
DATE	Priority date from 01/01/2002
Date of research	13.03.2022
2491	patented inventions
0,5	owned by top 10 players

Таблица 1: результаты поиска по ключевым словам

Текущие результаты и планы



	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
CHINA PETROLEUM & CHEMICAL	3	6	2	2	4	2		27	27	30	18	30	21	11	56	27	22	10	8		
JAPAN POLYPROPYLENE	4	6	6	4	9	18	15	19	15	9	14	23	14	13	6	8	9	7	5	3	
CHINA SINOPEC BEIJING RESEARCH INSTITUTE OF CHEMIC...		2	1	1	1	1		5	14	14	1	26	12	10	51	21	16	2	7		
BOREALIS							6	16	7	10	10	14	14	9	12	19	12	11	19	3	
BASELL POLYOLEFINE	8	16	9	6	6	3	6	8	2	10	10	9	1	8	8	11	9	7	7	5	
SUMITOMO CHEMICAL	26	30	12	17	6	10	13	14	4	7	3			3			1	1		1	
DOW GLOBAL TECHNOLOGIES	10	3	4	10	9	7	10	3	9	3	2	1	2	1	1	4	1	7	6	1	
EXXONMOBIL CHEMICAL PATENTS	2	3	2	23	3		1					3	4	1	10	7	16	2	6	5	
SINOPEC		2	1		1	2		16	8	10	16	2			4	6	1	6			
MITSUI CHEMICALS	5	4	6	5	4	6	2	6	1			2	1	12	1	1	2	3	2		
SABIC GLOBAL TECHNOLOGIES												1	11	11	17	14	2	6	2		
ASAHI KASEI											2	6	1	8	5	14	8	5	1	4	
LOTTE CHEMICAL	1	1				1		2	3		2	1	5	4	6	7	8	9			
LG CHEM			2	3	5	3		7		2	1	1	7	5	8	1			1		
NOVA CHEMICALS			2	1	1	1	1		4	1			4	3	2	13	3	2	2	6	
PRIME POLYMER		2	1	1	2	6	3	5	2	1		2	2	4		1	3	2	5	2	
PETROCHINA				1	1		3	2	4	1	2	7	2	6	3	2	4	5	1		
UNIVATION TECHNOLOGIES	1	11	5	4	1	2	1		1			4	1	1			3	1			
WR GRACE	7	1	1	1		3		3	1	4	2	1	4				3	4	4	1	
BOREALIS TECHNOLOGY	4	3	6	2	5	11	2							1	1						

Рис. 7: Тепловая карта распределения патентов

Текущие результаты и планы

План работ на третий семестр

- Создать набор агентов и инструментов для решения задачи извлечения информации на базе одной БЯМ (ChatGPT-3.5Turbo)
- Собрать небольшой датасет (около 50-100 наблюдений) для оценки эффективности извлечения данных
- Оценить эффективность обработки данных с помощью метрики F1-score

Список литературы

1. Ananikov V. Top 20 Influential AI-Based Technologies in Chemistry. Chemistry, 2024.

2. Ramos M.C., Collison C.J., White A.D. A Review of Large Language Models and Autonomous Agents in Chemistry: arXiv:2407.01603. arXiv, 2024.

3. Lála J. et al. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research: arXiv:2312.07559. arXiv, 2023.

4. Zheng Z. et al. ChatGPT Chemistry Assistant for Text Mining and Prediction of MOF Synthesis.