УДК 004.89 doi: 10.26102/2310-6018/2018.23.4.014

Н.А. Вайнгольц, Г.А. Верещак, Д.М. Коробкин, С.А. Фоменков

АВТОМАТИЗАЦИЯ СРАВНЕНИЯ ХИМИЧЕСКИХ ФОРМУЛ

Волгоградский государственный технический университет, Волгоград, Россия

Эксперту патентного ведомства для установления уникальности патентируемой технологии необходимо провести сравнение патентной заявки с документами патентного массива и удостовериться в отсутствии полных аналогов изобретения. При анализе патентов химических классов требуется сравнивать химические формулы, которые могут быть приведены в различных форматах: MOL, InChi, SMILES, структурная формула, молекулярный отпечаток. В данной работе описывается разработка программного модуля, автоматизирующего процедуры конвертации различных способов формализации химической формулы, сравнения формул химических соединений в патентной заявке и документах патентного массива, выявления патентов-аналогов на основе результатов сравнения химических соединений, содержащихся в патентах. Сравнение химических формул производится на основе вычисления схожести молекулярных отпечатков с использованием коэффициента Танимото. Коэффициент схожести патентов вычисляется на основе максимальных значений коэффициента Танимото для набора сравниваемых химических соединений из патентов. Программный модуль реализован на языке Java с использованием технологии Spring Framework, СУБД H2 и библиотеки Chemistry Development Kit (CDK). Реализованный программный модуль показал высокую эффективность (высокая полнота и точность поиска патентов-аналогов на основе химических формул, низкие значения потери информации и информационный шум) при проверке на тестовом патентном массиве.

Ключевые слова: химическая формула, SMILES, InChi, MDL Molfile, молекулярный отпечаток, анализ патентного массива, коэффициент Танимото.

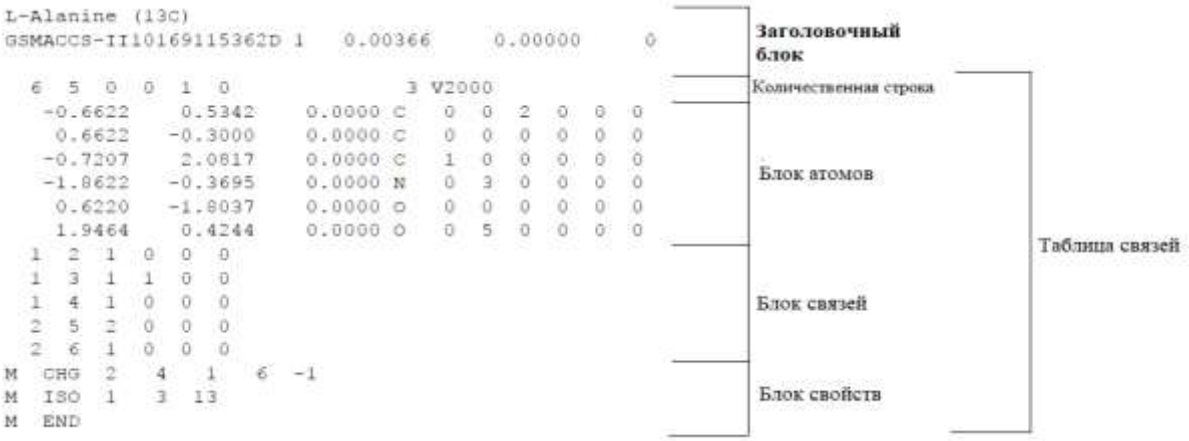
Введение

С увеличением количества патентов (в настоящее время в мире зарегистрировано свыше 20 миллионов патентов) увеличивается время рассмотрения патентной заявки. Эксперту патентного ведомства для установления уникальности патентуемой технологии необходимо провести сравнение со схожими патентами (иногда требуется проанализировать свыше тысячи документов) и удостовериться в отсутствии аналогов изобретения.

Существующие автоматизированные системы помощи эксперту при патентной экспертизе [1,2] не обеспечивают полной автоматизации процесса. Эксперт изучает заявку и с помощью автоматизированной системы производит патентный поиск на основе ключевых слов, выделенных им вручную из патента-заявки.

При анализе патентов химических классов требуется сравнивать химические формулы [3], т.е. формализованные описания химических соединений, содержащие информацию об атомарном составе молекулы отдельно взятого химического соединения.

Химические формулы (Рисунок 1) могут быть приведены в различных форматах: MOL [4], InChi [5], SMILES [6], структурная формула [3], молекулярный отпечаток [7].



Файлы химических таблиц MDL Molfile

Идентификатор InChi (International Chemical Identifier)

1S/C10H13NO2/c1-10(11,9(12)13-2)8-6-4-3-5-7-8/h3-7H,11H2,1-2H3	1S/C10H13NO2/c1-13-10(12)9(11)7-8-5-3-2-4-6-8/h2-6,9H,7,11H2,1H3

Спецификация SMILES

CC(C1=CC=CC=C1)(C(=O)OC)N	COC(=O)C(CC1=CC=CC=C1)N

Структурные формулы

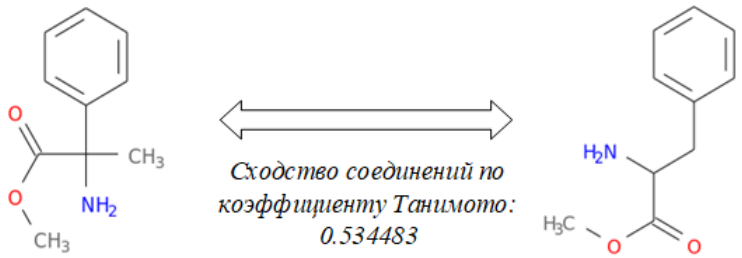


Рисунок 1 – Возможные представления химических соединений (Начало)

Молекулярный отпечаток (первые 1000 символов):

```
1000100101010000001001110010001110000110100001010000110100000010100000000000000000000000000000011010100100101 00
10000100010001010010000000000000000000000001111101000000000000000000000000001111101000000010000000001000000000
00000000000000001100001011000001010000111011001100000000000000000000010111011111110100100101000100010000010101
01000111100011011010111011011101110100110001011010001001010011100111000000111100011111110001010101110011110 1
11100100100001101010101110001101001101010001010010000101000100110101001010011010010000100000101000001110010 0
01110000100111110001010010000000100001110000110010110100010101100001110010101101001010010010011101101100001100
00011001110011101100110100100001100000110100001110000110000000111100000101011110101010010101001010000111000 0
11000001100100110010011001000100100000100001100000110100001100101000001101110011100010000110010110100100100 0
11010001001000010100001100001010000100010011101010100100111000000100100001100000100100001010000111000011000 00
1100011001111001101111011010
```

Рисунок 1 – Возможные представления химических соединений (Продолжение)

Патентное ведомство США (USPTO) [8] предоставляет свободный доступ к файловому хранилищу патентов, которое содержит zip-архивы с файлами в формате XML (текст патента) и связанными документами. Такими документами, в том числе, являются файлы формата MDL Molfile, содержащие список химических соединений для патентов химических классов.

Файлы химических Таблиц MDL Molfile [4] содержат информацию об атомах, связях, валентности и координатах атомов. Файлы MDL состоят из: заголовка; количественной строки; блока атомов; блока связей; блока свойств.

Структурная формула [3] отображает молекулярную структуру химического соединения, показывая атомарные связи и то, каким образом атомы расположены относительно друг друга в пространстве.

Идентификатор InChi (International Chemical Identifier) [5] – международный текстовый идентификатор химических соединений, обеспечивающий способ кодирования информации о составе и структуре молекул. InChi использует множество уровней: формула, связи, изотопы, стереохимия, таутомерия (обратимая изомерия).

Спецификация SMILES (Simplified molecular-input line-entry system) [6] представляет собой систему правил однозначного описания структуры химических соединений строками символов ASCII.

Молекулярный отпечаток [7] – способ кодирования структуры молекулы. Представляет собой битовую строку, где каждый бит соответствует какому-либо химическому свойству, преимущественно, подструктурам. Молекулярные отпечатки обычно используются для поиска схожих химических соединений.

Таким образом, актуальной является задача автоматизации помощи эксперту при рассмотрении патентной заявки химического класса. Поскольку химические формулы в патентной заявке могут быть представлены всеми возможными способами (MOL, InChi, SMILES, структурная формула, молекулярный отпечаток), то необходимо разработать автоматизированные процедуры конвертации различных способов формализации химической формулы и сравнения формул химических соединений в патентной заявке и документах патентного массива.

Материалы и методы

Обзор систем анализа химических формул

Для разрабатываемого программного модуля можно выделить четыре существующих аналога:

A. База данных химических соединений «ChemSpider» [9];
B. База химических соединений «PubChem» [10];
C. База химических соединений «ChemSynthesis» [11];
D. Система распознавания химических соединений «NCI/CADD Chemical Identifier Resolver» [12].

Системы-аналоги оценены по следующим критериям (Таблица 1):

1. Наличие / отсутствие SMILES-представления хим. соединения;
2. Наличие / отсутствие InChi-представления хим. соединения;
3. Наличие / отсутствие MOL-представления хим. соединения;
4. Наличие / отсутствие структурного представления химического соединения;
5. Наличие/отсутствие поиска химического соединения в патентах;
6. Наличие/отсутствие поиска схожих химических соединений.

Таблица 1 – Сводная Таблица оценки систем-аналогов по критериям

Система / Критерий	A	B	C	D
1	1	1	1	1
2	1	1	1	1
3	0	0	0	0
4	1	1	1	1
5	0	1	0	0
6	1	1	0	0

В результате анализа систем-аналогов было принято решение разработать программный модуль, осуществляющий выполнение следующих функций:

- Обработка патентного массива с целью извлечения элементов описания патента и химических формул в форматах MOL, SMILES, InChi, структурное представление, молекулярный отпечаток.
- Конвертация химических формул во все перечисленные представления.
- Сравнение химических соединений на основе молекулярных отпечатков.
- Поиск патентов-аналогов на основе результатов сравнения химических соединений, содержащихся в патентах.

Методика формирования базы данных патентов химических классов

Общий алгоритм формирования базы данных патентов химических классов, содержащей информацию о патентах, представления химических соединений, а также данные о схожих химических соединениях в патентном массиве и выявленных патентах-аналогах, представлен на Рисунке 2.

На начальном этапе осуществляется процесс извлечения из патентного массива информации о патентах химических классов [13,14] и используемых в них химических формулах.

Происходит обработка патентного массива с парсингом патентов, содержащих информацию о химических соединениях, и сохранением определенных полей патентов в БД. Для таких патентов осуществляется извлечение содержащихся в них химических соединений и сохранением их представлений в БД.

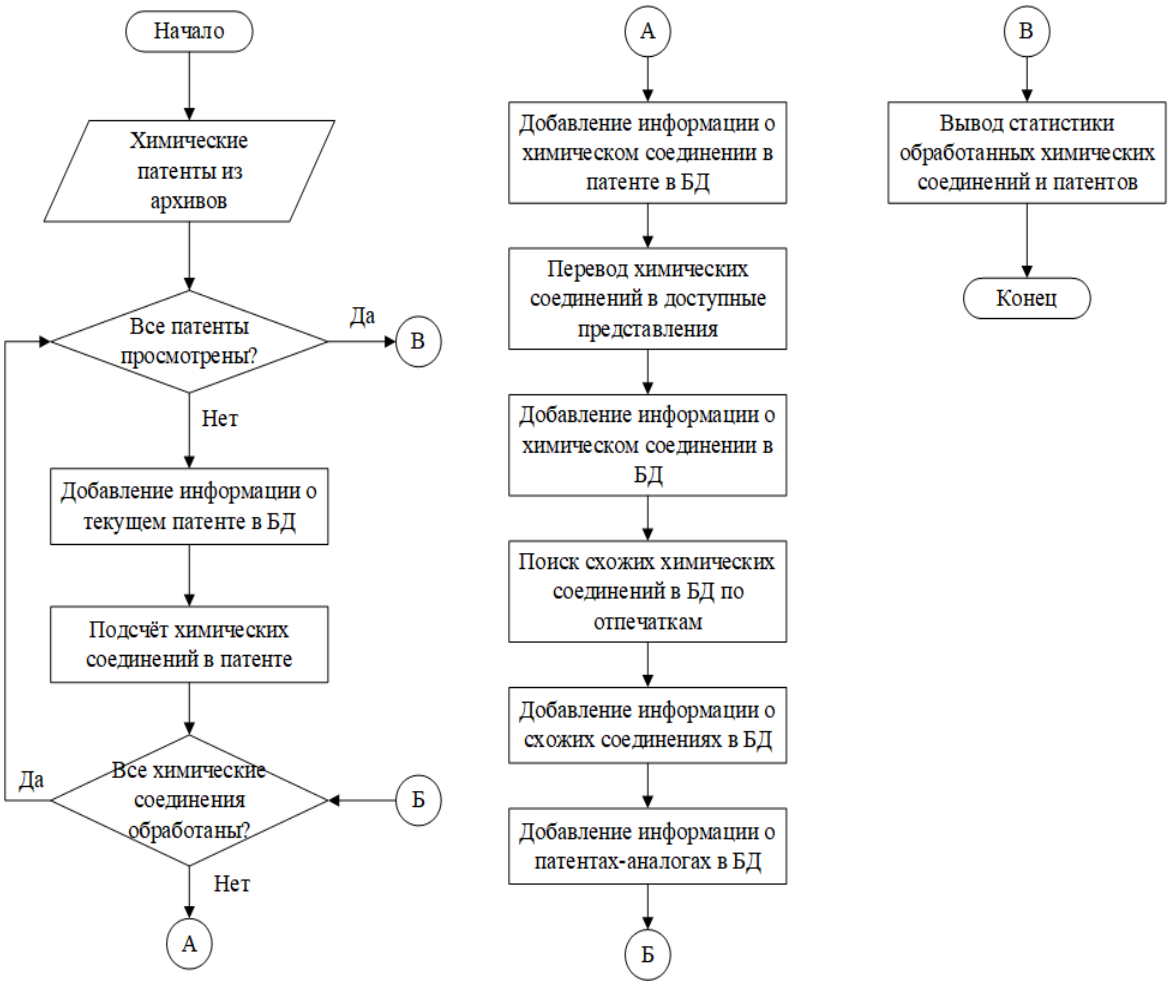


Рисунок 2 - Алгоритм формирования базы данных информации о патентах

Перед сравнением химических формул в различных представлениях осуществляется конвертация химических формул для патентов USPTO из файлов формата MDL Molfile в представления SMILES, InChi, структурные изображения, молекулярные отпечатки. Алгоритм конвертации химических формул в различные представления изображён на Рисунке 3.

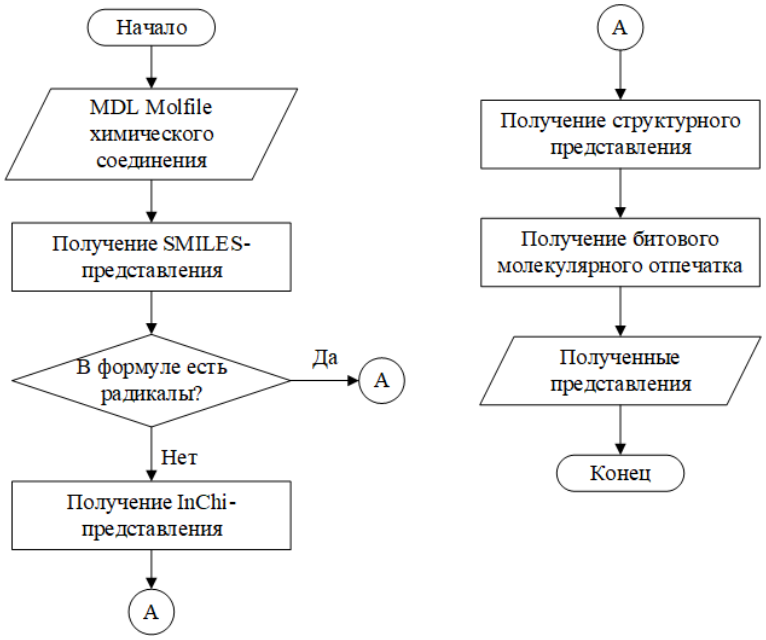


Рисунок 3 - Алгоритм конвертации представлений химических формул

Алгоритм сравнения химических соединений

Сравнение химических формул производится на основе молекулярных отпечатков. Для вычисления меры их схожести используется коэффициент Танимото [15].

Коэффициент Танимото вычисляется по следующей формуле:

$$T_S = \frac{\sum_i (X_i \cap Y_i)}{\sum_i (X_i \cup Y_i)}, \quad (1)$$

где X_i – i-ый бит отпечатка первой молекулы, Y_i – i-ый бит отпечатка второй молекулы, $\cap$ - побитовое И, $\cup$ - побитовое ИЛИ, T_S – коэффициент Танимото для отпечатков X и Y.

Алгоритм сравнения химических формул по молекулярным отпечаткам представлен на Рисунке 4. Рассмотрим работу алгоритма по шагам.

На вход алгоритма подаются два битовых молекулярных отпечатка. Химическая формула 1 имеет следующее SMILES-представление: C1=CC(OC1=O)=O. Химическая формула 2 имеет следующее SMILES-представление: C1=CC(=C(C(=C1*)*)*)O.

Молекулярный отпечаток химической формулы 1 в бинарном виде состоит из 48576 знаков (0 или 1), из которых установленных (равных 1) бит - 15010. Молекулярный отпечаток химической формулы 2 в бинарном виде состоит из 57960 знаков, из которых установленных бит - 20440. Отпечатки выравниваются: к меньшему отпечатку добавляются нули, пока оба отпечатка не сравняются по длине.

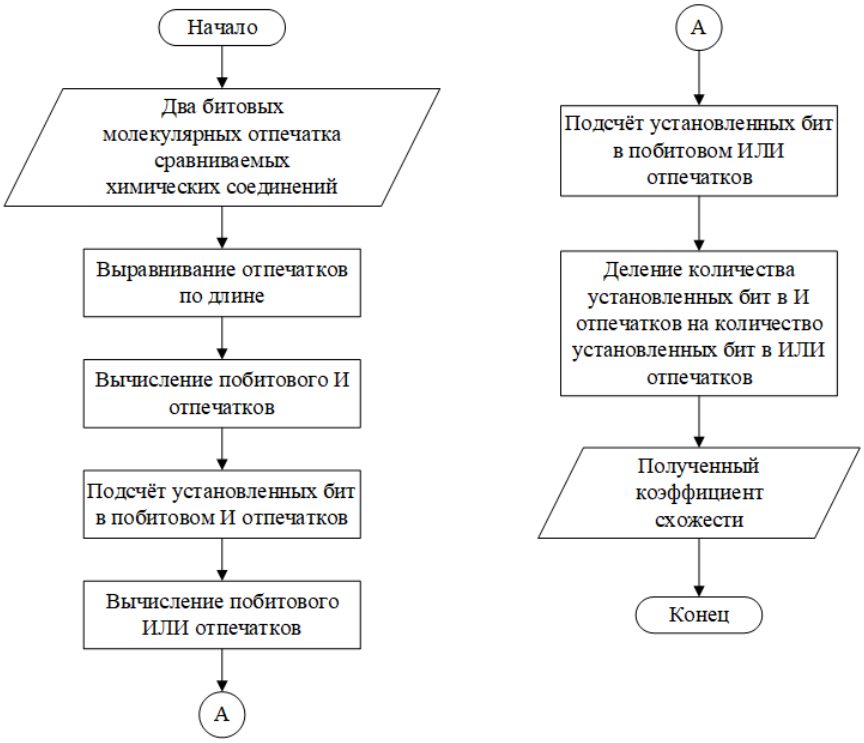


Рисунок 4 - Алгоритм сравнения химических соединений

Затем к двум отпечаткам применяется побитовое «И» и «ИЛИ» по отдельности; рассчитывается количество установленных бит для «И» и для «ИЛИ». Если молекулярные отпечатки идентичны, то количество бит после применения побитового «И» и побитового «ИЛИ» будет одинаковым. Для рассматриваемого примера количество установленных бит после применения операции «И» равно 3800, количество установленных бит после применения операции «ИЛИ» равно 5985. Их отношение дает коэффициент схожести, равный 0.634921. В зависимости от установленного порогового значения на основе коэффициента схожести определяется, являются ли рассматриваемые химические формулы структурно подобными.

Алгоритм сравнения патентов

Выявление патентов-аналогов производится на основании данных проведённой процедуры сравнения молекулярных отпечатков химических формул для сравниваемых патентов.

Рассмотрим пример работы алгоритма, на вход которого подаются патенты, информационные карты которых представлены в Таблицах 2 и 3.

Таблица 2 – Выходная карта патента 1

Поле	Значение
Название	Moldable compositions containing carbinol functional silicone resins or anhydride functional silicone resins
Номер публикации	US07807012
Дата публикации	2010-10-05
Заявитель	Dow Corning Corporation

Таблица 3 – Выходная карта патента 2

Поле	Значение
Название	Isolated aqueous enzymatic preparation and the use thereof for the functionalization of the surface of paper or cellulosic substrates
Номер публикации	US09702087
Дата публикации	2017-07-11
Заявитель	UNIVERSITAT POLITECNICA DE CATALUNYA; RatnerPrestia; Universitat Politecnica de Catalunya; Nopco Paper Technology Holding AS

Список химических соединений для сравниваемых патентов USPTO содержится в виде файлов формата MDL Molfile.

В патенте 1 содержатся следующие химические формулы:

1. C1(=CC(C2C(C1)C(OC2=O)=O)[Si](C)(C)OC)C;
2. C1(=CC(C2C(C1)C(OC2=O)=O)C)C;
3. C1=CC(OC1=O)=O.

В патенте 2 содержатся следующие химические формулы:

1. C1=CC(=C(C(=C1*)*)*)O;
2. C1=C(C(=C(C=C1C(=O)O*)O*)O)*;
3. C1(=C(C(=C2C(=C1*)CCC(O2)(CCC=C(C)CCC=C(C)CCC=C(C)C)C)*)*)O;
4. C1=C(C(=C(C=C1C(=O)O*)O*)O)*;
5. C1(=C(C(=C2C(=C1*)CCC(O2)(CCCC(C)CCCC(C)CCCC(C)C)C)*)*O;
6. C1=C(C(=C(C=C1C(=O)O*)O*)O)*;
7. C1=C(C=CC=C1)C(C=2C(=C(C=C(C2)C(C=3C=CC=CC3)C)C(C=4C=CC=CC4)C)O[H])C;
8. C1(=CC=C(C=C1)OC=2C=CC(=CC2)O)C(F)(F)F.

Для каждого файла с химическим соединением, относящегося к патенту, рассчитывается его молекулярный отпечаток. Затем происходит попарное сравнение молекулярных отпечатков соединений из патентов 1 и 2 между собой (Таблица 4).

Таблица 4 – Сравнение химических соединений из патентов

Химическая формула из патента 1	Химическая формула из патента 2	Коэффициент Танимото
C1(=CC(C2C(C1)C(OC2=O)=O)[Si](C)(C)OC)C	C1=CC(=C(C(=C1*)*)*)O	0.564756
C1(=CC(C2C(C1)C(OC2=O)=	C1=C(C(=C(C=C1C(=O)O*)O*)O)*	0.487217

Химическая формула из патента 1	Химическая формула из патента 2	Коэффициент Танимото
O)[Si](C)(C)OC)C		
C1(=CC(C2C(C1)C(OC2=O)=O)[Si](C)(C)OC)C	C1(=C(C(=C2C(=C1*)CCC(O2)(CCC=C(C)CCC=C(C)CCC=C(C)C)C)*)*)O	0.455891
C1(=CC(C2C(C1)C(OC2=O)=O)[Si](C)(C)OC)C	C1=C(C(=C(C=C1C(=O)O*)O*)O)*	0.458193
C1(=CC(C2C(C1)C(OC2=O)=O)[Si](C)(C)OC)C	C1(=C(C(=C2C(=C1*)CCC(O2)(CCCC(C)CCCC(C)CCCC(C)C)C)*)*)O	0.634921
C1(=CC(C2C(C1)C(OC2=O)=O)[Si](C)(C)OC)C	C1=C(C(=C(C=C1C(=O)O*)O*)O)*	0.612731
C1(=CC(C2C(C1)C(OC2=O)=O)[Si](C)(C)OC)C	C1=C(C=CC=C1)C(C=2C(=C(C=C(C2)C(C=3C=CC=CC3)C)C(C=4C=CC=CC4)C)O[H])C	0.598742
C1(=CC(C2C(C1)C(OC2=O)=O)[Si](C)(C)OC)C	C1(=CC=C(C=C1)OC=2C=CC(=CC2)O)C(F)(F)F	0.473124

Для первой химической формулы из патента 1 наиболее схожей является пятая химическая формула из патента 2 (коэффициент Танимото равен 0.634921).

За результирующее значение коэффициента схожести патентов K_T принимается отношение суммы максимальных значений коэффициента Танимото для сравниваемых химических соединений из меньшего по размеру набора с химическими соединения из другого патента к размерности минимального набора.

$$K_T = \frac{\sum_i \max_j T_S}{\min(|C_1|,|C_2|)}, \quad (2)$$

где C_1, C_2 - множества химических соединений для патента 1 и 2 соответственно;

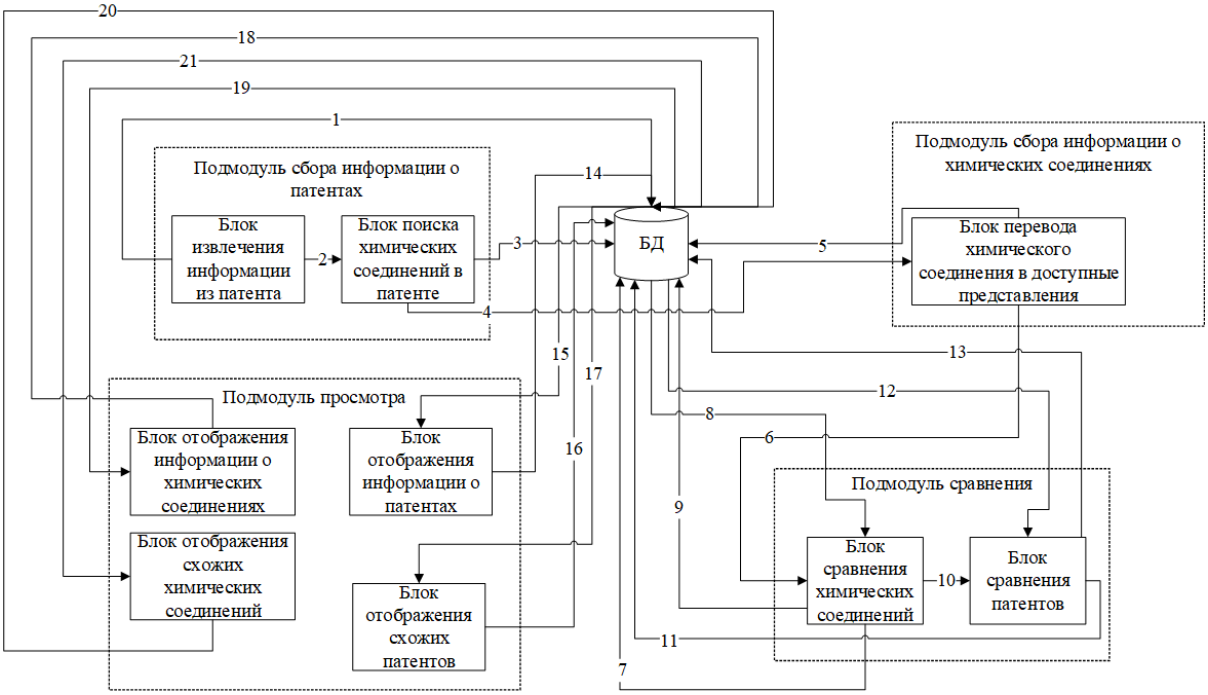
$\min(|C_1|,|C_2|)$ –минимальный размер множеств химических соединений для двух патентов;

i – индекс перебора по минимальному по размеру множеству химических соединений;

$\max_j T_S$ - максимальное значение коэффициента Танимото для i – ого химического соединения.

Результаты

Программный модуль (архитектура представлена на Рисунке 5) реализован на языке Java с использованием технологии Spring Framework 5.0.1 [16]. В качестве СУБД используется H2 [17] вследствие качественной интеграции со Spring Framework. Библиотека Chemistry Development Kit (CDK) [18] выбрана в качестве коллекции средств для обработки химической информации.



1 – сохранение в БД записи с информацией о патенте;
2 – вызов блока поиска химических соединений в патенте;
3 – сохранение в БД записи с информацией о том, что текущие химические соединения содержатся в патенте;
4 – вызов блока перевода химического соединения в доступные представления;
5 – сохранение в БД записи химического соединения;
6 – вызов блока сравнения химических соединений;
7 – запрос к БД на извлечение записей химических соединений;
8 – получение из БД записей химических соединений;
9 – сохранение результатов сравнения химических соединений в БД;
10 – вызов блока сравнения патентов;
11 – запрос к БД на извлечение записей патентов;
12 – получение из БД записей патентов;
13 – сохранение результатов сравнения патентов в БД;
14 – запрос к БД на извлечение записей патентов;
15 – получение из БД записей патентов;
16 – запрос к БД на извлечение записей схожих патентов;
17 – получение из БД записей схожих патентов;
18 – запрос к БД на извлечение записей химических соединений;
19 – получение из БД записей химических соединений;
20 – запрос к БД на извлечение записей схожих химических соединений;
21 – получение из БД записей схожих химических соединений.

Рисунок 5 - Архитектура модуля

Разработанный модуль отображает:

- информацию о патентах и химических соединениях, содержащихся в патентах (Рисунок 6);

- характеристики химического соединения и список патентов, в которых содержится данное химическое соединение;
- список схожих химических соединений;
- список патентов, которые сходны с выбранным патентом на основании содержащихся в нём химических соединений.

Рисунок 6 - Web-страница с информацией о выбранном химическом соединении

Для проверки эффективности разработанного модуля была составлена выборка с 53 патентами (187 химических соединений) из списка цитирования отдельно взятых 15 патентов. Найдена 41 пара схожих патентов, 120 пар схожих химических соединений.

На основе полученных данных были вычислены показатели эффективности поиска [19]:

$$R = \frac{a}{a+c} \times 100\%, \quad (3) \qquad P = \frac{a}{a+b} \times 100\%, \quad (4)$$

$$L = \frac{c}{a+c} \times 100\%, \quad (5) \qquad N = \frac{b}{a+b} \times 100\%, \quad (6)$$

где R - полнота выдачи патентов-аналогов; P - точность выдачи;
L - потери информации;
N - информационный шум;
a - количество релевантных и выданных модулем патентов;
b - количество нерелевантных, но выданных модулем патентов;
c- количество релевантных, но не выданных модулем документов;
a = 41, b = 0, c = 12.

Полнота поиска патентов-аналогов = 77,3%; точность = 100%; потери информации = 22,6%; информационный шум = 0%.

Обсуждение

В результате разработки модуля автоматизации сравнения химических формул были достигнуты следующие цели:

- разработана методика формирования базы данных патентов химических классов;
- разработаны алгоритмы конвертации химических соединений в различные представления, сравнения химических соединений, поиска патентов-аналогов на основе результатов сравнения химических соединений;
- спроектирован модуль анализа химических формул, произведены его реализация и тестирование.

Перспективы развития программного модуля:

- разработка поиска патентов и химических соединений по заданному ключу (представлению);
- разработка алгоритма поиска патентов-аналогов с учётом списка цитирования;
- расширение списка доступных представлений химических соединений, выделение наименований IUPAC из текста;
- кластеризация патентов химических классов по различным признакам и выделение патентных трендов.

Заключение

Работа выполнена при поддержке РФФИ (проекты №18-07-01086 А, № 16-07-00534 А).

ЛИТЕРАТУРА

1. Д.М. Коробкин, Н.А. Гордеев, С.А. Фоменков, М.А. Дыков. Метод выявления патентных трендов на основе описаний технических функций. Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. - Волгоград, 2018. - № 5 (215). - С. 56-60.
2. Д.М. Коробкин, С.А. Фоменков, И.А. Кобликов, Г.А. Карачунова. Методика семантического патентного поиска. Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. - Волгоград, 2017. - № 1 (196). - С. 65-73.

3. Chemical formula | Britannica.com [Электронный ресурс]. – Режим доступа : https://www.britannica.com/science/chemical-formula (дата обращ. 15.05.2018).
4. MDL Information Systems, Inc. CTFile Formats / MDL Information Systems, Inc. – San Leandro : MDL Information Systems, 2003. – 106 с.
5. Heller, R. The IUPAC International Chemical Identifier (InChI) / R. Heller, Alan D. McNaught // CHEMISTRY International. – 2009. – № 1. – С. 7-9.
6. Daylight Theory: SMILES [Электронный ресурс]. – Режим доступа : http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (дата обращ. 15.05.2018).
7. Dalke, A. Molecular fingerprints, background [Электронный ресурс] // A. Dalke. – Режим доступа : http://www.dalkescientific.com/writings/diary/archive/2008/06/26/fingerprint_background.html (дата обращ. 15.05.2018).
8. Bulk Data Storage System [Электронный ресурс]. – 2018. – Режим доступа : https://bulkdata.uspto.gov/ (дата обращ. 25.10.2018).
9. ChemSpider reaches 50 million compounds [Электронный ресурс]. – Режим доступа: http://www.rsc.org/journals-books-databases/librarians-information/librarians-notes/all-articles/2016/jun/chemspider-reaches-50-million-compounds/ (дата обращ. 18.05.2018).
10. PubChem Docs – About [Электронный ресурс]. – Режим доступа: https://pubchemdocs.ncbi.nlm.nih.gov/about (дата обращ. 18.05.2018).
11. ChemSynthesis – Chemical Database [Электронный ресурс]. – Режим доступа: http://www.chemsynthesis.com/ (дата обращ. 18.05.2018).
12. NCI/CADD Chemical Resolver – Chemical Identifier Resolver documentation [Электронный ресурс]. – Режим доступа: https://cactus.nci.nih.gov/chemical/structure_documentation (дата обращ. 18.05.2018).
13. Д.М. Коробкин, Е.А. Тюлькина, С.А. Фоменков, С.Г. Колесников. Система извлечения технических функций из патентного массива. ИТНОУ: Информационные технологии в науке, образовании и управлении. - 2017. - № 2 (2). - С. 24-30.
14. И.А. Кобликов, Д.М. Коробкин, С.А. Фоменков, В.А. Яровенко. Методика извлечения описаний реализуемых в патенте технических функций. Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. - Волгоград, 2017. - № 8 (203). - С. 55-59.
15. Tanimoto (cdk 2.1-SNAPSHOT API) [Электронный ресурс]. – Режим доступа: http://cdk.github.io/cdk/2.1/docs/api/org/openscience/cdk/similarity/Tanimoto.html (дата обращ. 21.05.2018).

16. Spring Framework Overview [Электронный ресурс]. – Режим доступа: https://www.tutorialspoint.com/spring/spring_overview.htm (дата обращ. 18.05.2018).
17. Spring Boot and H2 in memory database – Why, What and How? – Spring Boot Tutorial [Электронный ресурс]. – Режим доступа: http://www.springboottutorial.com/spring-boot-and-h2-in-memory-database (дата обращ. 19.05.2018).
18. Chemistry Development Kit [Электронный ресурс]. – Режим доступа: https://cdk.github.io/ (дата обращ. 19.05.2018).
19. Гопта Е.А., Фоменков С.А., Карачунова Г.А. Автоматизация процесса линейного синтеза физического принципа действия. Известия Волгоградского государственного технического университета. 2010. № 11 (71). С. 129-133.

N.A. Vayngolts, G.A. Vereshchak, D.M. Korobkin, S.A. Fomenkov

THE AUTOMATIZATION OF CHEMICAL FORMULAS COMPARISON

Volgograd State Technical University
Volgograd, Russia

An expert of the patent office to establish the uniqueness of the patented technology, it is necessary to compare the patent application with the patents and make sure that there are no analogues of the invention. When analyzing patents of chemical classes, it is required to compare chemical formulas that can be given in different formats: MOL, InChi, SMILES, structural formula, molecular fingerprint. This paper describes the development of a software that automates the procedures: conversion of various formalization of the chemical formula, comparison of chemical formulas from the patent application and patents, identification of patents-analogues based on the results of comparison of chemical formulas. Comparison of chemical formulas is based on the calculation of the similarity of molecular fingerprints using the Tanimoto coefficient. The coefficient of similarity of patents is calculated based on the maximum values of the Tanimoto coefficient for a set of compared chemical compounds from patents. The software is developed on Java using the Spring Framework technology, the H2, and the Chemistry Development Kit (CDK). The software showed a high performance (high recall and precision of the patent search on the basis of chemical formulas, the lowest values of the information loss and noise).

Keywords: chemical formula, SMILES, InChi, MDL Molfile, molecular fingerprint, patent database analysis, Tanimoto Coefficient.

REFERENCES

1. D.M. Korobkin, N.A. Gordeev, S.A. Fomenkov, M.A. Dykov. Metod vyyavleniya patentnyh trendov na osnove opisanij tekhnicheskih funkcij. Izvestiya VolgGTU. Ser. Aktual'nye problemy upravleniya, vychislitel'noj tekhniki i informatiki v tekhnicheskih sistemah. - Volgograd, 2018. - № 5 (215). - C. 56-60.
2. D.M. Korobkin, S.A. Fomenkov, I.A. Koblikov, G.A. Karachunova. Metodika semanticheskogo patentnogo poiska. Izvestiya VolgGTU. Ser. Aktual'nye problemy upravleniya, vychislitel'noj tekhniki i informatiki v tekhnicheskih sistemah. - Volgograd, 2017. - № 1 (196). - C. 65-73.
3. Chemical formula – https://www.britannica.com/science/chemical-formula.
4. MDL Information Systems, Inc. CTFile Formats / MDL Information Systems, Inc. – San Leandro : MDL Information Systems, 2003. – 106 p.
5. Heller, R. The IUPAC International Chemical Identifier (InChI) / R. Heller, Alan D. McNaught // CHEMISTRY International. – 2009. – № 1. – pp. 7-9.
6. Daylight Theory: SMILES - http://www.daylight.com/dayhtml/doc/ theory/theory.smiles.html.
7. Dalke, A. Molecular fingerprints, background - http://www.dalkescientific.com/writings/diary/archive/2008/06/26/ fingerprint_background.html.
8. Bulk Data Storage System - https://bulkdata.uspto.gov.
9. ChemSpider reaches 50 million compounds - http://www.rsc.org/journals-books-databases/librarians-information/librarians-notes/all-articles/2016/jun/ chemspider- reaches-50-million-compounds.
10. PubChem Docs – About – https://pubchemdocs.ncbi.nlm.nih.gov/about.
11. ChemSynthesis – Chemical Database [EHlektronnyj resurs]. – Rezhim dostupa: http://www.chemsynthesis.com/ (data obrashch. 18.05.2018).
12. NCI/CADD Chemical Resolver – Chemical Identifier Resolver documentation – https://cactus.nci.nih.gov/chemical/structure_documentation.
13. D.M. Korobkin, E.A. Tyul'kina, S.A. Fomenkov, S.G. Kolesnikov. Sistema izvlecheniya tekhnicheskih funkcij iz patentnogo massiva. ITNOU: Informacionnye tekhnologii v nauke, obrazovanii i upravlenii. - 2017. - № 2 (2). - C. 24-30.
14. I.A. Koblikov, D.M. Korobkin, S.A. Fomenkov, V.A. YArovenko. Metodika izvlecheniya opisanij realizuemyh v patente tekhnicheskih funkcij. Izvestiya VolgGTU. Ser. Aktual'nye problemy upravleniya, vychislitel'noj tekhniki i informatiki v tekhnicheskih sistemah. - Volgograd, 2017. - № 8 (203). - C. 55-59.

15. Tanimoto (cdk 2.1-SNAPSHOT API) – http://cdk.github.io/cdk/2.1/docs/api/org/openscience/cdk/similarity/Tanimoto.html.
16. Spring Framework Overview – https://www.tutorialspoint.com/spring/spring_overview.htm.
17. Spring Boot and H2 in memory database – Why, What and How? – Spring Boot Tutorial –http://www.springboottutorial.com/spring-boot-and-h2-in-memory-database.
18. Chemistry Development Kit –https://cdk.github.io.
19. Gopta E.A., Fomenkov S.A., Karachunova G.A. Avtomatizaciya processa linejnogo sinteza fizicheskogo principa dejstviya. Izvestiya Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta. 2010. № 11 (71). S. 129-133.