# Поиск документов по формулам

**Ю.С. Корухова[1], К.П. Кригер[1], Е.Б. Рязанова[1]**

*[1] Московский государственный университет имени М.В. Ломоносова, факультет Вычислительной математики и кибернетики*

**Аннотация.** В настоящее время большое количество научных статей, исследований и книг преобразованы в цифровую форму и опубликованы на общедоступных ресурсах. Важной составляющей научных работ, особенно в естественных науках и инженерных дисциплинах, являются формулы. Но большинство известных современных поисковых систем не предоставляют возможности поиска по формулам. Математическая поисковая система решает множество проблем, возникающих как при написании собственных научных статей, так и при изучении естественных наук. В данной работе предложен подход к созданию системы математического поиска для коллекций документов, в том числе содержащих отсканированные изображения: подобраны программные инструменты и библиотеки, позволяющие автоматически переводить их в структурированный вид с возможной корректировкой ошибок. В работе представлен метод построения индекса и алгоритм нечёткого поиска. Предложенный подходы реализованы в прототипной программной системе.

**Ключевые слова:** информационный поиск, поиск статей по формулам

# Information Retrieval for Texts with Formulas

**Y.S. Korukhova[1], K.P. Kriger[1], E.B. Riazanova[1]**

*[1] Lomonosov Moscow State University, Computational Mathematics and Cybernetics faculty*

**Abstract.** Nowadays a large number of scientific articles, research texts and books have been converted into digital form and published on open-access web sites. Formulas are an important component of research works, especially in mathematics, in natural sciences and in engineering disciplines. But most of the well-known modern search engines do not implement mathematical retrieval of texts from formulas. A mathematical search engine could solve many problems for authors doing some research and also for educational purposes also. In this paper, an approach to creating a mathematical search system for collections of documents, including those containing scanned images, is proposed: software tools and libraries are selected that allow them to be automatically translated into

a structured form with possible error corrections. The paper presents a method for constructing an index and a fuzzy search algorithm. The proposed approaches are implemented in a prototype software system.

**Keywords:** information retrieval, mathematic information retrieval

## Введение

В настоящее время большое количество научных статей, исследований и книг были преобразованы в цифровую форму и опубликованы на общедоступных ресурсах. Важной составляющей научных публикаций, особенно в математике, естественных науках и инженерных дисциплинах, являются формулы. Однако большинство известных поисковых систем не предоставляют возможности поиска текстов по заданной формуле.

Реализация поиска по формулам является логичным дополнением к возможности поиска статей по автору, ключевым словам и другим данным. Кроме того, по формулам можно найти работы, написанные по решению одной и той же задачи на разных языках, так как язык статьи практически не влияет на вид формулы. Для проведения доказательств и написания научных работ также представляется полезной такая возможность поиска: могут быть обнаружены публикации, где ранее уже была выведена искомая формула или похожая на неё, или доказана некоторая её оценка.

Реализация системы поиска текстов по формулам (назовем ее далее математической поисковой системой) связана с рядом проблем. Во-первых, формулы и статьи представлены в различных форматах: частично - в структурированном виде, но значительная часть ресурсов хранится в виде отсканированных изображений (в формате pdf или jpg), что добавляет к задаче поиска проблему распознавания текста и формул на изображениях. Во-вторых, актуальной представляется реализация именно нечеткого поиска: кроме текстов, содержащих точное совпадение с формулой-запросом, в списке найденных документов ожидаются и те, что содержат прохожие формулы, либо же эквивалентные, но записанные в другом виде. Для реализации поисковой системы необходимо также предложить способ индексирования статей в различных форматах. Наконец, особенностью данной системы является интерфейс, позволяющий задавать формулы для поиска в удобном пользователю виде: как в известных редакторах формул, так и в виде изображения, которое может быть введено с помощью мыши или написано от руки и сфотографировано камерой мобильного устройства. Все из перечисленных вариантов ввода могут привести к появлению опечаток или ошибок распознавания, которые необходимо предлагать скорректировать автоматически.

В работе представлен подход к созданию математической поисковой системы. Сначала рассматриваются известные методы и системы поиска по формулам. Затем описан предлагаемый метод построения индекса формул

на основе синтаксического дерева и алгоритм нечеткого поиска. Далее представлена информация о практической реализации предложенных методов, представлен интерфейс поисковой системы, включающий различные возможности ввода формул, и продемонстрирована работа системы на тестовой коллекции документов. В заключении приведены основные результаты.

## 1. Системы поиска текстов по формулам

Проблема поиска научных работ по формулам стала актуальна с появлением множества научных работ в открытом доступе в сети Интернет. На данный момент существуют как официальные цифровые библиотеки с множеством статей и учебников, так и отдельные веб-сайты, на которых есть возможность загрузить свою работу в цифровом виде. Так как электронные коллекции научных статей появились достаточно давно, уже существуют поисковые системы с возможностью нечёткого поиска текстов по запросу-формуле. Рассмотрим эти системы.

Один из подходов к решению задачи математического поиска представлен в системе **MathWebSearch** [1]. Система состоит из трёх компонентов: набора веб- сканеров, поискового сервера и веб-сервера. Веб сканеры обходят интернет в поисках подходящего контента, идентифицируют и загружают его в индекс, поисковый сервер осуществляет поиск по индексу, а веб-сервер передаёт результаты пользователю. MathWebSearch работает с документами в формате XML и формулами в формате MathML [2].

Для хранения заранее составленного индекса формул используется дерево подстановок, а дополнительная информация о формуле хранится в базе данных. Все формулы заранее проиндексированы и организованы в виде дерева подстановок. В статье [1] оценено, что время работы поисковой системы остаётся приемлемым даже для достаточно большой коллекции документов, так как многие подформулы разных формул совпадают и записываются в дерево только один раз. Поиск по коллекции порядка 3400 статей выполняется за миллисекунды. Таким образом, дерево подстановок представляется интересным инструментом для индексирования. Однако эксперименты с системой показали, что хотя и возможна выдача в результатах поиска похожих на запрос формул с точностью до переменных, подформул и надформул, но к релевантными результатам для формулы $s^2(x)$ не относится, например, $s(x) * s(x)$. Что демонстрирует необходимость ввода в систему ряда тождественных преобразований формул. Подход, использующий тождественные преобразования на основе волновых правил, был исследован в работе [3], однако не доведен до окончательной реализации.

Другой подход предложен при реализации математической поисковой системы на основе механизма EgoMath [4], написанному в дополнение к текстовому поисковому механизму Egothor и, поскольку он разработан как расширение, его может использовать любая полнотекстовая поисковая система. Поддерживаются формулы в формате MathML. Формулы в индексе хранятся в виде текста. Из сложных структур данных они переводятся в линейные с заранее определенными символами, при этом все подформулы представлены как отдельные слова. Данный подход предоставляет возможности использовать преимущества текстового поиска.

Поиск по формулам, представленным в виде структурированных данных реализован в системах MIaS [5] (формат MathML) и в поиске для Цифровой математической библиотеки Лобачевского [6] (формат LaTeX). В системе LibMeta [7] реализована возможность поиска по формулам, сами же формулы выделяются из текстовых фрагментов, и с ними связываются объекты и понятия тезауруса, что позволяет в дальнейшем искать формулы по ключевым словам. Однако задача поиска по коллекции неструктурированных документов в формате pdf по-прежнему остается актуальной [8].

## 2. Индексирование неструктурированных документов, содержащих формулы

В данной работе основное внимание уделено поиску по коллекции отсканированных статей с формулами, хранящимся в формате pdf. Особенность построения индекса заключается в дополнительном этапе обработки таких документов – распознавании формул в них с возможной корректировкой ошибок. Рассматриваемый документ в формате pdf разделяется на страницы, каждая из которых переводится в графический формат (с помощью библиотеки pdf2image[1]).

Далее используем программный инструмент для распознавания текста Mathpix[2] Так как программа распознавания символов Mathpix OCR может распознавать не только печатный текст, но и текст с изображений, то становится возможной и обработка отсканированных документов, и формул, записанных вручную.

В работе предлагается следующий алгоритм построения индекса.

1. Документу присваивается уникальный идентификатор (id) и этот номер вместе со ссылкой на документ записывается в базу данных
2. С каждой страницы считывается текст с помощью программного инструмента Mathpix.

[1] https://pdf2image.readthedocs.io
[2] https://mathpix.com

3. Из текста выделяются все формулы: делается проход по тексту и ищутся подстроки, совпадающие с выражением формулы в структурированном виде (в формате Mathpix Markdown[3]).

4. Выполняется преобразование формул к стандартному виду: убирается лишняя для поиска информация (например, о том, что формула была набрана жирным шрифтом), корректируются некоторые ошибки (такие как, например, пропущенная операция умножения или неверное имя стандартной функции).

5. Формулы вместе с их идентификационными номерами записываются в дерево подстановок, оно хранится в файле в формате json.

6. Формулы и идентификационные номера записываются в таблицу базы данных, связываются с информацией о документах, в которых они найдены

Преобразование формулы из строки в дерево выполняется методом рекурсивного спуска по грамматике, описывающей выражения.

Следующий шаг работы алгоритма - добавление формулы в дерево подстановок. Сначала для формулы строится собственное дерево подстановок, затем оно встраивается в общее для всех формул дерево. Так, например, для формулы c = a + b дерево будет выглядеть как представлено на рис. 1. Дерево подстановок получено обходом дерева формулы в ширину справа налево (см. рис. 2).

Для построения дерева подстановок для всех формул, все деревья объединяются. Результат хранится в словаре, где ключ - узел дерева, значение - пара из списка детей узла и списка идентификаторов формул, которые соответствуют данному пути дерева. Например, для формул d = f, a+b = c и b+a = c будет общее дерево подстановок, представленное на рис. 3, в листьях которого содержатся идентификаторы формул для поиска соответствующих формуле документов в базе данных.

[3] https://mathpix.com/docs/mathpix-markdown/overview

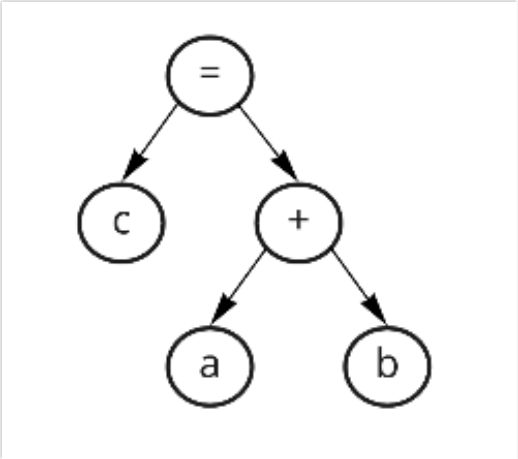


Рис. 1. Представление формулы c = a + b в виде дерева

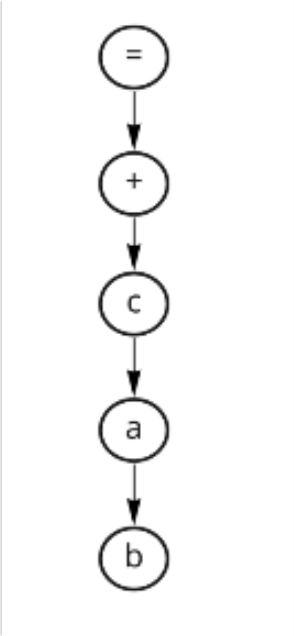


Рис. 2. Дерево подстановок для формулы c = a + b

Подстановки в формуле будут проходить в ширину, тогда вверху дерева в узлах будут операторы, а внизу - переменные. Таким образом расстояние между формулами, отличающимися только переменными, будет меньше.

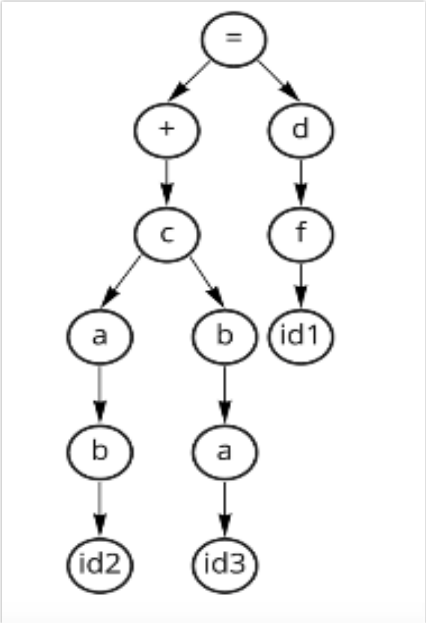


Рис. 3. Общее дерево подстановок для формул d = f, a + b = c и b + a = c

### 3. Алгоритм нечёткого поиска и его реализация

Исходными данными для поиска является формула в виде строки. Последовательно выполняются следующие шаги

1. Строится идентификатор (хэш) для формулы-запроса (с помощью функций библиотеки hashlib [4]) и делается проверка, что нет формул с совпадающим идентификатором в базе данных

2. Для формулы-запроса строится дерево подстановок

3. Осуществляется проход по общему дереву подстановок и поиск наиболее похожих формул.

4. Из дерева получаются идентификаторы (id) наиболее похожих формул и в базе данных выбираются документы, содержащие эти формулы.

5. Документы ранжируются в зависимости от расстояния между найденными формулами

Остановимся подробнее на пункте 3. Проход по дереву подстановок осуществляется не только по совпадающим с запросом веткам, но по всем веткам дерева. При переходе из одного узла на следующий каждому из узлов-потомков присваивается вес перехода. Вес перехода зависит от совпадения формулы со значением следующего узла дерева. Если значение узла не совпадает со значением узла формулы, то можно либо перейти на следующий узел дерева и на следующий узел формулы, либо перейти на следующий узел дерева и остаться на текущем узле формулы, либо остаться на текущем узле дерева и перейти на следующий узел формулы. Каждому из таких действий соответствует разный вес ошибки и при дальнейшем проходе по дереву эти веса складываются. Если во время прохода на каком-

[4] https://docs.python.org/3/library/hashlib.html

то пути вес становится слишком большим, то эта ветвь дерева убирается из рассмотрения. В конце для каждой формулы из дерева получается финальный вес - расстояние от формулы из запроса до формулы в дереве. Если формула из запроса точно совпадает с формулой в дереве, то расстояние между ними будет равно 0, так как вес каждого перехода будет равен 0. Таким образом получается найти точные совпадения, наиболее близкие формулы, подформулы и надформулы.

Ранжирование результатов, упомянутое в пункте 5 алгоритма, выполняется согласно следующим приоритетам:

1. Вес ошибки, полученный в результате работы алгоритма (от меньшего к большему).
2. Доля документов, в которых встречается формула, по отношению ко всем документам (от большего к меньшему).
3. Количество упоминаний формулы в документе, относительно упоминаний всех формул в документе (от большего к меньшему).

Результатом работы алгоритма является список формул, похожих на входную формулу и список документов, содержащих найденные формулы.

## 4. Реализация системы математического поиска и ее интерфейс

Предложенные подходы были реализованы в виде программной системы на языке Питон, имеющей веб-интерфейс, возможности которого представлены на рис. 4.

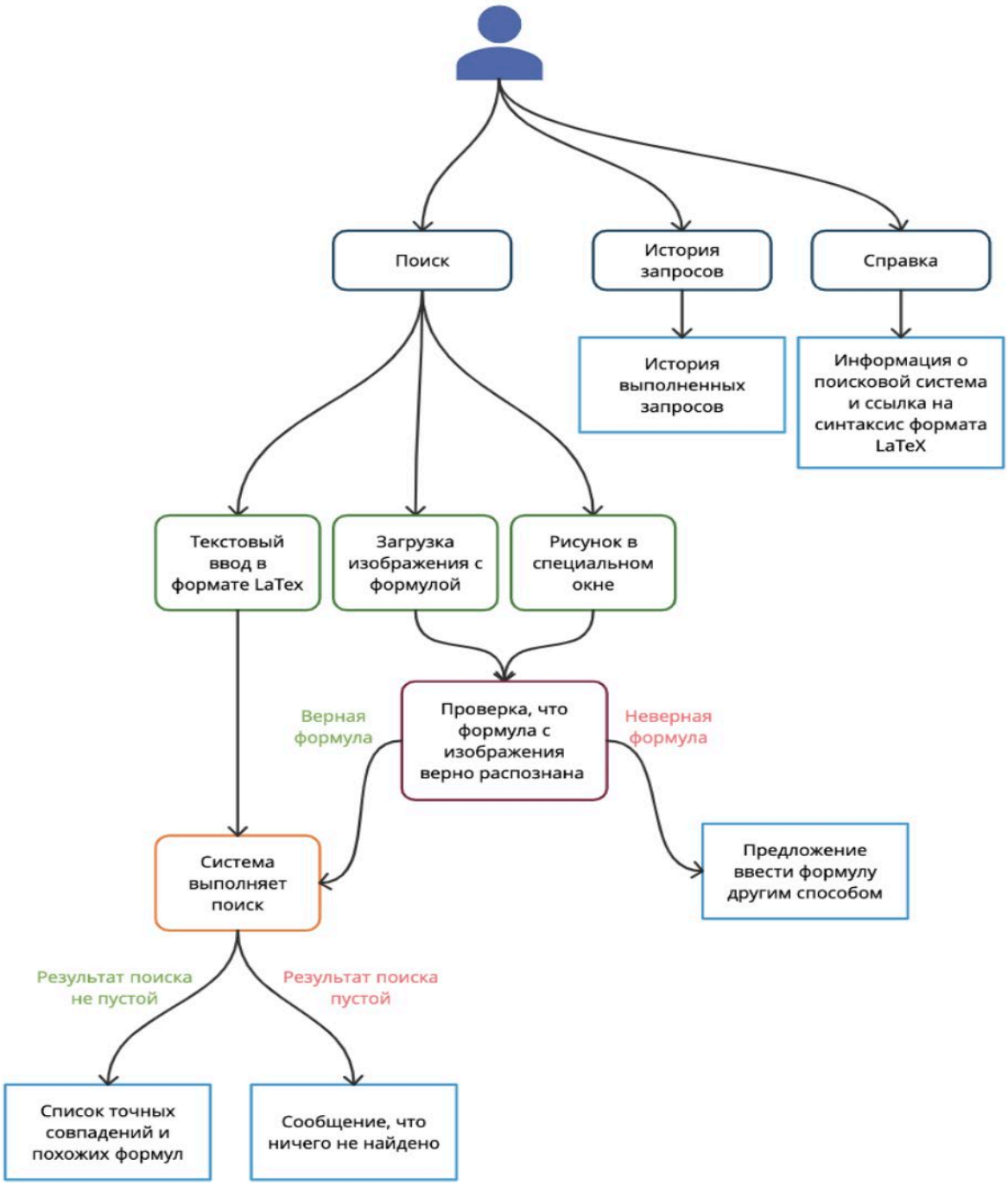


Рис. 4. Сценарии взаимодействия пользователя с системой математического поиска через веб-интерфейс

Также в процессе реализации находится подсистема ввода и преобразования формул, работающая на мобильных устройствах (ОС Android), позволяющая считывать формулу – запрос с помощью камеры мобильного устройства и выполнять ряд эквивалентных преобразований для сопоставления запроса с формулами индекса и корректировки возможных ошибок распознавания и записи.

Для исследования работы метода была проиндексирована тестовая библиотека документов, содержащих формулы. В нее включено 100 работ на русском языке из электронной библиотеки препринтов ИПМ им. Келдыша [5] и 200 работ на английском языке из электронного архива научных статей arxiv.org. В обоих источниках есть возможность скачивания файла в формате pdf, а также разделение статей по тематике. Результатом работы модуля индексирования стало дерево, содержащее около 32 000 узлов, размер файла 1,6 Мб.

Для работы с формулами и статьями создана база данных, схема которой представлена на рис. 5.

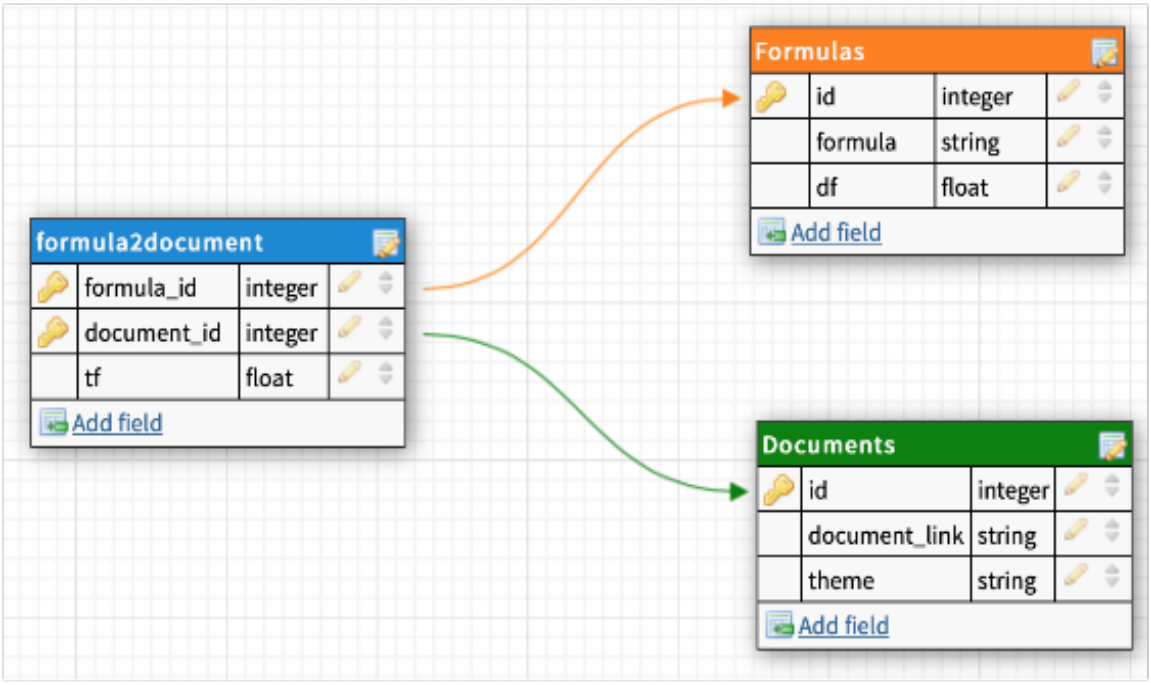


Рис. 5. Схема базы данных математической поисковой системы

В базе данных представлена информация

— о формулах (id - уникальный идентификатор, formula - строковое представление формулы, df - доля документов, в которых встретилась формула по отношению ко всем документам);

— о документах (id - уникальный идентификатор, document - ссылка на статью, theme - тематика статьи), в дальнейшей работе таблица может быть расширена, когда при поиске будет задействоваться дополнительная информация, в частности предполагается в дальнейшем учитывать предметную область, к которой работа относится;

— о связи формул и документов (formula_id - идентификатор формулы, document_id - идентификатор документа, в котором содержится

[5] https://library.keldysh.ru/

формула, tf - доля данной формулы по отношению ко всем формулам в данном документе).

Пример работы поисковой системы представлен на рис. 6. По заданной в текстовом виде формуле в коллекции было найдено два документа (препринта), ссылки на которые выданы в качестве результатов поиска.

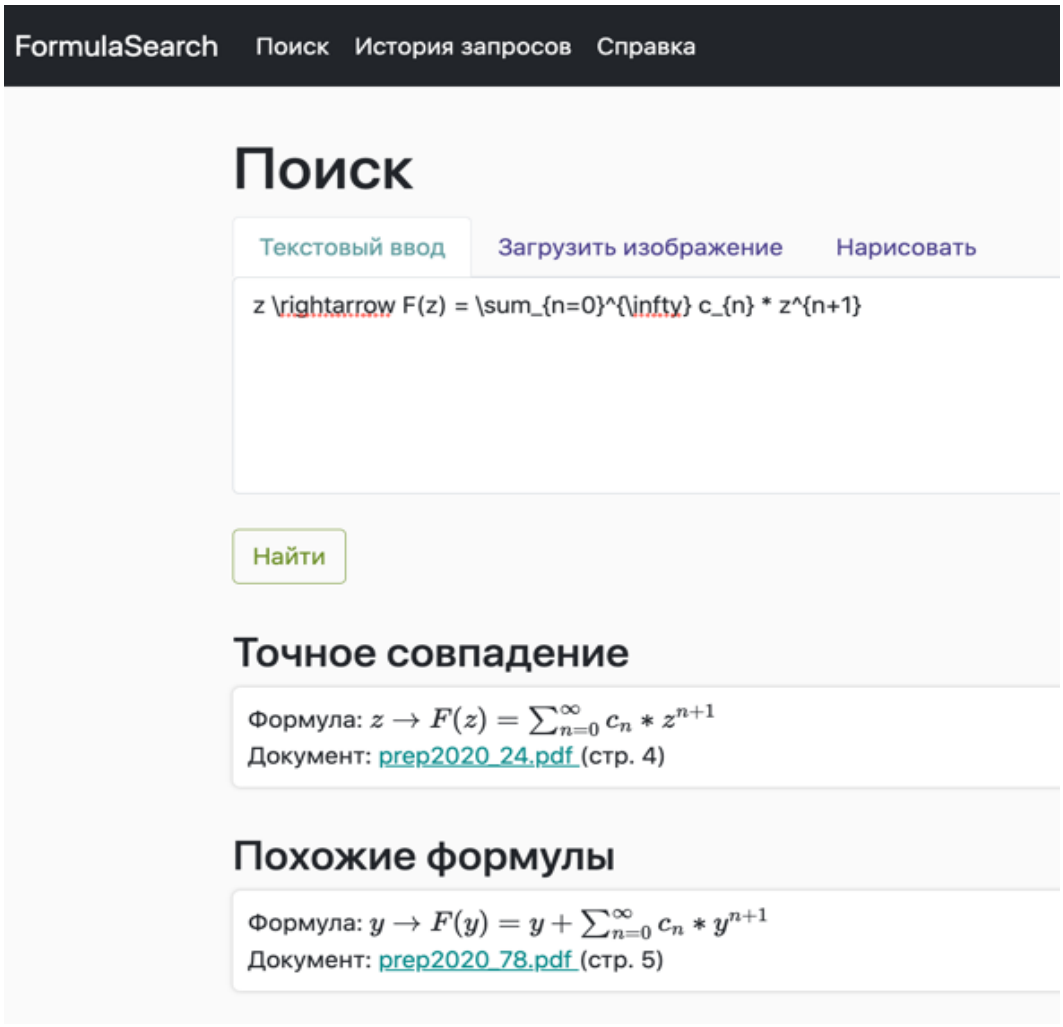


Рис. 6. Пример работы поисковой системы

В процессе тестирования системы было измерено, что время поиска оказывается в пределах 1 секунды, и зависит от длины формулы-запроса.

**Заключение**

В рамках данной работы предложен подход к решению задачи поиска по заданной формуле текстов, в которых содержится либо сама формула, либо похожие на нее. В результате исследования различных методов индексирования, было выбрано представление индекса в виде дерева подстановок формул со ссылками на базу данных. Работа прототипной

реализации системы математического поиска продемонстрирована на тестовой коллекции документов. В дальнейшем предполагается расширение коллекции проиндексированных документов, доработка интерфейса системы на мобильных устройствах и расширения набора преобразований, которые допустимы при сопоставлении формул.

**Литература**


1. Kohlhase M, Sucan I. A Search Engine for Mathematical Formulae. Proceedings of the 8th international conference on Artificial Intelligence and Symbolic Computation. 2006.
2. MathML documentation https://developer.mozilla.org/en-US/docs/Web/MathML
3. Широкий Р.В. Система поиска научных текстов по формулам – Тезисы лучших дипломных работ факультета ВМК МГУ – Издательский отдел факультета ВМК МГУ, 2014
4. Jozef Misutka, Leo Galambos. Extending Full Text Search Engine for Mathematical Content. Towards Digital Mathematics Library. 2008.
5. Petr Sojka, Michal Ruzicka, Vıt Novotny MIaS: Math-Aware Retrieval in Digital Mathematical Libraries. Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018.
6. Цифровая математическая библиотека Лобачевского https://lobachevskii-dml.ru
7. Атаева О.М., Серебряков В.А., Тучкова Н.П. Цифровая библиотека по обыкновенным дифференциальным уравнениям на основе LibMeta // Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (18-23 сентября 2017 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2017. — С. 21-33. — URL: http://keldysh.ru/abrau/2017/65.pdf doi:10.20948/abrau-2017-65
8. Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. Издательство Вильямс. 2011.


**References**


1. Kohlhase M, Sucan I. A Search Engine for Mathematical Formulae. Proceedings of the 8th international conference on Artificial Intelligence and Symbolic Computation. 2006.
2. MathML documentation https://developer.mozilla.org/en-US/docs/Web/MathML
3. Shirokiy R.V. A System for Formulas Retrieval in Scientific Texts – Abstracts of best diploma works of MSU CMC faculty – Editorial board of CMC faculty of MSU, 2014
4. Jozef Misutka, Leo Galambos. Extending Full Text Search Engine for Mathematical Content. Towards Digital Mathematics Library. 2008.

5. Petr Sojka, Michal Ruzicka, Vıt Novotny. MIaS: Math-Aware Retrieval in Digital Mathematical Libraries. Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018.
6. Lobachevskii Digital Mathematics Library —https://lobachevskii-dml.ru
7. Ataeva O.M., Serebryakov V.A., Tuchkova N.P. Digital library of Ordinary Differential equations based on LibMeta // Scientific services in the Internet: proceedings of the XIX Russian scientific conference (18-23 September 2017 г., Novorossiysk). — M.: Keldysh Institute of Applied Mathematics, 2017. — C. 21-33. —
URL: http://keldysh.ru/abrau/2017/65.pdf doi:10.20948/abrau-2017-65
8. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze Introduction to Information Retrieval – Williams, 2011.