Филиал Московского Государственного Университета
имени М.В. Ломоносова в городе Ташкенте
Факультет прикладной математики и информатики
Кафедра прикладной математики и информатики

Бутузов Игорь Владимирович

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

НА ТЕМУ: «Поиск математических утверждений в базе данных» ПО НАПРАВЛЕНИЮ 01.04.02 «Прикладная математика и информатика»

ВКР рассмотрена и рекомендована к защите
зав. кафедрой «ПМиИ» д.ф.-м.н., профессор _____Кудрявцев В.Б.

Научный руководитель
к.ф.-м.н. _____Калачев Г.В.

«____» _____ 2020 год

Ташкент 2020

Аннотация

Данная работа посвящена разработке программы, способной найти само утверждение или похожее на него в базе данных с математическими теоремами. Такая программа сможет по запросам пользователя ранжировать теоремы таким образом, чтобы теоремы, максимально подходящие под запрос, получали более высокий приоритет при выдаче результата.

В первой главе работы дано описание работы, реализованной программы. Рассматриваются этапы её работы, описываются алгоритмы работы основных модулей.

Во второй части работы показаны результаты работы модуля ранжирования при использовании двух алгоритмов. Для этой цели использовался набор из 15 вручную разобранных задач.

Ключевые слова: синтаксический анализ, теорема, функция ранжирования

Abstract

This paper is devoted to the development of a program capable of finding the statement itself or similar to it in a database with mathematical theorems. Such a program will be able to rank the theorems at the user's request so that the theorems that are most suitable for the query receive a higher priority when issuing the result.

The first chapter of the paper describes the work of the implemented program. The stages of its work are considered, the algorithms of the main modules are described.

The second part of the paper shows the results of the ranking module using two algorithms. For this purpose, a set of 15 manually parsed tasks was used.

Keywords: syntax analysis, theorem, ranking function

Содержание

Введение

Математические теоремы обладают уникальной структурой синтаксиса и большим разнообразием семантически эквивалентных конструкций. Одну и ту же теорему можно записать несколькими способами. Например, изменив обозначення в используемых формулах на эквивалентные, так и изменив конструкцию, записанную на естественном языке, на конструкцию с тем же смыслом.

По этой причине поиск по ключевым словам в теоремах может не дать желаемого результата. Если рассматривать только поиск по формулам, входящим в состав математического утверждения, без учета текста, записанного на естественном языке, то такой поиск имеет ряд существенных недостатков: например, формула может встречаться в нескольких теоремах, она может быть записана при помощи других обозначений или другой формулой с тем же семантическим значением.

Формализация структуры математического утверждения способна повысить вероятность найти искомое утверждение в базе данных с такими утверждениями. Для формализации текста теоремы на естественном языке, удобно использовать язык логики предикатов. Перевод расплывчатой словесной формулировки на строгий, не допускающий противоречивых толкований язык логики предикатов способствует четкости и ясности мышления, но в процессе осуществления такого перевода возникает ряд трудностей. Например, такой процесс не может быть полностью автоматизирован, т.к. на результат работы программы может сильно изменится из-за одного неверно истолкованного слова.

Полученное представление текста теоремы на языке логики предикатов является аналогом формулы и оно может быть представлено в виде абстрактного дерева. Объединив такое дерево с формулами из данного утверждения, предварительно представленными в виде деревьев со схожей структуры, можно получить дерево исходного математического утверждения. К такому абстрактному дереву удобно применять алгоритмы работы с деревьями.

Результат перевода на язык логики предикатов будет более качественным, если входной текст для такого перевода будет структурирован и в нем будет содержаться вся необходимая для перевода информация [2],[3]. Такое представление текста могут дать синтаксические анализаторы. Задачей синтаксического анализатора является определение того какие члены предложения являются главными, а какие – подчиненными. Для этого у компьютера на входе есть цепочка символов, которую нужно правильно проинтерпретировать, разбить на слова, связать их между собой и построить синтаксическое дерево.

В данной работе используется синтаксический анализатор Stamford NLP [5], для работы которого используется модель СинТагРус [7], [8]. Для преобразования математических формул в форму дерева, используется РБНФ-грамматика (расширенная форма Бэкуса — Наура [4]) и библиотека для Python 3 TatSu [6], которая позволяет по файлу, содержащему правила построения грамматики, построить дерево формулы.

Постановка задачи

Целью данной работы является создание программы, способной найти само утверждение или похожее на него в базе данных с математическими теоремами.

Данная система сможет по запросам пользователя ранжировать теоремы таким образом, чтобы теоремы, максимально подходящие под запрос, получали более высокий приоритет при выдаче результата.

Актуальность данной темы заключается в том, что ежегодно доказывается несколько сотен тысяч теорем и других математических утверждений. Поиск нужного утверждения в столь огромном массиве данных представляется нетривиальной задачей. В отличие от поиска по ключевым словам реализуемая система будет иметь более высокий результат, так как будет опираться не на вхождение определенных слов в запрос, а на смысловое содержание введенной теоремы. Поэтому такая система математического поиска сможет сопоставлять теорему, которую ввел пользователь и теорему из базы данных, даже если они записаны при помощи разных формулировок.

Предполагается, что как заданная теорема, так и теоремы из базы данных записаны на естественном языке (возможно, с использованием математических формул, представленных в формате системы TeX). При этом в программе должно учитываться, что формулировки заданной теоремы и той же теоремы в базе данных могут несколько отличаться.

В процессе работы было решено несколько промежуточных задач:

1. Задача перевода текста, написанного на естественном языке, на язык формул логики предикатов (с помощью синтаксического и семантического анализа), и последующего представления этих формул в виде деревьев;

2. Задача представления в том же виде формул, записанных в формате системы tex;

3. Задача ранжирования формул базы данных, записанных в виде деревьев, по степени похожести на формулу теоремы-запроса.

Кроме того, для проверки работоспособности программы необходима база теорем, а также некоторое множество теорем, выступающих в качестве запросов к программе.

В качестве входных данных для данной программы используется синтаксический анализ Stamford. Программа реализована на языке Python 3.

Результатом работы программы является такое представление базы математических утверждений , в которых формулировка искомой теоремы или похожей на нее должна появиться среди лучших кандидатов, найденных программой.

Сформулируем основные понятия, используемые в данной работе:

Токен — слово, устойчивое выражение или знак препинания. Токен имеет словоформу, лемму и набор морфологических характеристик.

Синтаксическое дерево зависимостей — способ представления синтаксической структуры предложения. В узлах дерева расположены токены данного предложения [?]. Под словами «родительский» и «дочерний» по отношению к токенам в данной работе следует понимать отношения между соответствующими узлами дерева. Каждому токену (исключение составляет корень синтаксического дерева) поставлено в соответствие положительное число – номер другого токена, который является родителем для данного. Для корня синтаксического дерева это число равно нулю. Между «родительским» и «дочерним» токенами устанавливается синтаксическое отношение, его мы будем называть типом связи. У корня синтаксического дерева родителя нет.

Словоформа — форма слова, полученная из основной части слова с помощью склонения и спряжения.

Лемма — словарная форма слова.

1 Описание системы математического поиска

1.1 Описание базы данных с математическими утверждениями

Для тестирования работы реализованной программы были выбраны 300 теорем на русском языке из Википедии и различных учебных пособий. Теоремы были взяты из различных разделов математики.

Предварительно было произведено разделение текстовой части утверждения теорем и содержащихся в них математических формул на языке TeX. Формулы добавлялись в отдельный текстовый файл, а в самих теоремах заменялись на специальные обозначения вида formula_№, где № – порядковый номер формулы в данном файле.

К каждому преобразованному таким образом математическому утверждению применялся синтаксический анализатор Stamford NLP, который создавал таблицу для каждого слова со следующими полями:

№ Слово Лемма Номер родителя Тип синтаксической связи Часть речи

В полученную таблицу вносился ряд правок. Например, добавление недостающих обозначений для математических объектов, замена местоимений вида «она»,«он» на слова, которые они заменяли. Также слова из словаря заменялись на свои латинские эквиваленты. Несколько подряд идущих токенов, которые образовывали конструкцию («имеет место»,«интегрируема по Риману»), объединялись в одну конструкции.

На следующем этапе данные из таблицы преобразовывались в абстрактное синтаксическое дерево. Абстрактным синтаксическим деревом в данной программе является список вложенных списков, где заголовком каждого списка является родительский токен.

На следующем шаге по полученному абстрактному дереву строилось множество всех возможных путей из корневой вершины. Из данного множества строятся выражения, которые затем и составят формулу логики предикатов.

Для преобразования математических формул в абстрактное синтаксическое дерево, используется РБНФ-грамматика (расширенная форма Бэкуса — Наура) и библиотека для Python 3 TatSu, которая позволяет по файлу с правилами грамматики построить дерево.

На следующем этапе работы специальные обозначения, введенные ранее в формулах логики предикатов, заменялись на деревья соответствующих формул. Полученная формула логики предикатов виде списка вложенных списков вносилась в текстовый файл.

1.2 Описание модуля перевода текста, написанного на естественном языке, на язык формул логики предикатов

Входными данными для данного модуля являются предложения, записанные на естественном языке. Математические формулы в данном предложении заменены на обозначения типа «formula_№», где № – номер соответствующего обозначения или формулы в файле со списком обозначений и формул. Выходными данными для данного модуля является исходное предложение, записанное при помощи логики предикатов.

Входные данные подаются синтаксическому анализатору Stamford NLP. Результат работы анализатора представляет собой таблицу (массив), содержащий подробную морфологическую и синтаксическую информацию, соответствующую обозначениям Национального Корпуса Русского Языка (СинТагРус) [8]. Каждый токен входного предложения представляет собой отдельный массив в данной таблице. В данном массиве содержится порядковый номер токена, его словоформа, лемма, номер родителя, тип связи между родителем и данным токеном, часть речи данного токена.

Для дальнейшей работы в полученную таблицу необходимо внести корректировки.

На первом этапе несколько отдельных токенов объединяются в один и заменяются либо на соответствующее название на английском языке, либо на стандартное математическое обозначение. Несколько подряд идущих токенов в таблице токенов. Например, токены, образующие конструкцию вида «интегрируема по Риману», будут заменены одним токеном R. Конструкция «равномерно непрерывна» будет заменена на конструкцию «uniform_contin», а конструкция «тогда и только тогда, когда» будет заменена на токен «<=>». При каждой такой замене происходит пересчет номеров родителей для дочерних токенов.

Для математических терминов написан словарь, где указаны соответствующая лемма термина и эквивалентное ему слово, на которое его можно заменить. Если произошло совпадение токена и значения в словаре, то будет произведена замена. Также на этом этапе к некоторым токенам при необходимости добавляются соответствующие обозначения. Стандартное обозначение будет добавлено для объектов «точка», «функция», «множество», «последовательность» при отсутствии у них уже существующего обозначения.

Если добавляемое стандартное обозначение уже встретилось в программе и используется для другого объекта, то будет к стандартному обозначению добавляется номер.

Также на предварительном этапе слова вида «его», «эта» заменяются на обозначение термина, который подразумевается под данным словом. Для данного типа токенов смотрится следующий за ним элемент, его значение запоминается и затем осуществляется поиск другого токена с такой же словоформой. Если таких токенов несколько, то будет отдан приоритет тому токену, который имеет такую же синтаксическую связь как у токена, который был запомнен. Если поиск удачен, то среди дочерних элементов найденного токена ищется его обозначение. Данное обозначение подставляется вместо «его», «эта».

Слова не имеющие смысловой важности, такие как «неравенство», «условие» могут быть удалены из таблицы токенов при условии, что следующий за ними токен является обозначением и это обозначение позволяет однозначно идентифицировать слово, которое предполагается удалить. Если удаление допустимо, предварительно все дочерние связи данных слов переносятся на обозначение данного слова. Например, для конструкции «выполнено неравенство formula_4», где родителем токена «неравенство» является токен «выполнено», необходимо проверить, что formula_4 однозначно идентифицировать неравенство. После корректировки родителем токена formula_4 станет токен «выполнено», formula_4 получит тот же тип связи, который связывал «неравенство» и «выполнить». После этого токен «неравенство» будет удален.

№	Название	Эквивалент в программе	Добавляемое обозначение
1	Функция	function	f
2	Множество	set	a
3	Отрезок	closed_interval	-
4	Непрерывный	C (continuity)	-
5	Монотонный	M (monotonic)	-

Таблица 1: Примеры заменяемых слов и добавляемых обозначений

Если в таблице с синтаксическим разбором есть «,» и после нее следуют специальные слова «a»,«то»,«тогда», то данная таблица будет разделена по этой запятой на две части с соответствующим пересчетом синтаксических связей. Следующий этап для каждой из частей будет проходить отдельно, а затем добавлены во временный массив. На следующем этапе данные из таблицы преобразовываются в абстрактное синтаксическое дерево (список вложенных списков). Используя номера токенов и их родителей рекурсивно строится такое представление дерева, в котором номер родителя является заголовком списка. Затем в полученном дереве номера токенов заменяется на конструкцию вида «лемма : тип синтаксического отношения».

Затем из полученных данных строится множество всевозможных путей из корневой вершины. Будем называть такое множество множеством путей.

Опишем некоторые преобразования из данного модуля:

1. Из множества путей удаляются элементы, которые полностью содержатся в других.

2. Если элемент множества путей без специального слова («if»,«then») полностью совпадает с другим элементом этого множества, то оставляется только специальное слово. Иначе, происходит разделение элемента множества на два: специальное слово и остальную часть.

3. Если 2 элемента множества путей начинаются с одинакового слова и не входят в список объектов («функция»,«точка»), то эти два элемента объединяется.

4. Сочетания объект и соответствующее определение, объект и обозначение добавляются в множество путей.

5. Для каждого из специальных слов (кванторов) «if»,«then»,«forall»,«exists» создается выражение для данного слова. Чтобы элемент множества путей вошел в данное выражение он должен либо содержать переменную (обозначение) на которую распространяется

действие квантора (это обозначение следует за кванторов), либо не содержать никак специальных слов, тогда оно будет отнесено в область действий предыдущего квантора. Если количество элементов в полученном выражении больше одного, то в заголовок данного выражения будет добавлена «&».

1.3 Описание модуля преобразования математических формул

На вход данному модулю поступает математическая формула, записанная в формате системы TeX. На предварительном этапе работы модуля при помощи регулярных выражений из нее удаляются слова, которые не несут никакой смысловой нагрузки: displaystyle, textstyle, right, left, limits, nolimits, quad. Некоторые обозначения заменяются на эквиваленты.

№	Обозначение	Эквивалент в программе	Значение
1	{ .. }	set	множество
2	[..]	closed_interval	отрезок
3	|| .. ||	norm	норма
4	| .. |	abs	модуль

Таблица 2: Примеры заменяемых обозначений

Для построения абстрактного синтаксического дерева формулы в данном модуле используется библиотека TatSu . Это библиотека генерирует код на языке Python из грамматик в вариации РБНФ (расширенная форма Бэкуса — Наура).

Описание грамматики в РБНФ представляет собой набор правил, определяющих отношения между терминальными символами (терминалами) и нетерминальными символами (нетерминалами).

Терминальные символы — это минимальные элементы грамматики, не имеющие собственной грамматической структуры. В РБНФ терминальные символы — это либо предопределённые идентификаторы (имена, считающиеся заданными для данного описания грамматики), либо цепочки — последовательности символов в кавычках или апострофах. Нетерминальные символы — это элементы грамматики, имеющие собственные имена и структуру. Каждый нетерминальный символ состоит из одного или более терминальных и/или нетерминальных символов, сочетание которых определяется правилами грамматики. В РБНФ каждый нетерминальный символ имеет имя, которое представляет собой строку символов.

Основным преимуществом парсера TatSu является то, что каждая операция парсинга грамматики при необходимости может быть отредактирована при помощи написания функции на языке Python 3.

1.4 Модуль функции ранжирования

В данной работе сравнивались два алгоритма ранжирования. Под коэффициентом ранжирования будем понимать некое число от 0 до 1. Чем это число ближе к 1, тем выше схожесть между двумя объектами.

В качестве входных данных для данного модуля поступают два дерева (формула), разобранных при помощи модуля перевода текста, написанного на естественном языке, на язык формул логики предикатов и модуля преобразования математических формул A и B. Для каждой из формул рекурсивно построим множество всех путей, начиная от корневой. Назовем эти два множества A' и B'.

Опишем первый алгоритм.

Для множеств A' и B' вычисляется коэффициент Жаккара по следующей формуле:

$$K = \frac{n(A' \cap B')}{n(A' \cup B')}, \text{где } n() \text{ – количество элементов в множестве } C$$

Если обе формулы одинаковы и отличаются лишь обозначениями для переменных, то коэффициент K равен 1. Чем больше различий между формулами, тем значение K ближе к 0.

Теперь опишем второй алгоритм.

Для всех элементов множества A' ищем идентичную конструкцию в B'. Идентичной конструкцией при сравнении двух множеств будем называть такое множество, которое может отличаться от того множества, с которым оно сравнивается только переобозначением переменных. За каждую совпадение увеличиваем значение функции c.

Затем для всех пар элементов A'_i и B'_j вычисляется коэффициент Жаккара $K_{a,b}$ по следующим формулам:

$$K_i = \frac{n(A'_i \cap B'_j)}{n(A'_i \cup B'_j)}$$

$$K_{a,b} = \sum_{i=1}^{n(A)} \frac{c \cdot K_i \cdot i}{n(A)},$$

где i, j – индексы, указывающий насколько далеко ушли от вершины формулы A,B соответственно; $n(A)$ – количество элементов в множестве A .

Повторим процедуру поменяв A'_i и B'_j местами в качестве входных параметров функции. Получим коэффициент Жаккара $K_{b,a}$. Полученные результаты $K_{a,b}$ и $K_{b,a}$ могут не находиться в пределах от 0 до 1. Поэтому проведем их нормировку. Итого, окончательное значение коэффициента ранжирования:

$$\text{Коэффициент ранжирования} = \min\left(\frac{K_{a,b}}{K_{a,a}}, \frac{K_{b,a}}{K_{b,b}}\right)$$

2 Полученные результаты

Заключение

Список литературы

[1] Игошин В.И. Математическая логика и теория алгоритмов. Москва. Издательский центр «Академия» 2008. Страница 196.

[2] Hiroyuki Yamauchi. Processing of syntax and semantics of natural language by Predicate Logic. Institute of Space and Aeronautical Science, University of Tokyo. https://www.aclweb.org/anthology/C80-1059
Страница 390

[3] Julian E. Herwitz. Natural Language to Predicate Logic https://www.cs.rochester.edu/b̃rown/173/exercises/samples/ParseTranslate10_2.pdf. Страница 5.

[4] https://www.cl.cam.ac.uk/ mgk25/iso-14977.pdf

[5] https://stanfordnlp.github.io/stanfordnlp/

[6] https://tatsu.readthedocs.io/en/stable/intro.html

[7] https://tatsu.readthedocs.io/en/stable/intro.html

[8] http://www.ruscorpora.ru/new/instruction-syntax.html