Разработка и реализация алгоритмов распознавания и поиска математических формул

Матлин Даниил Сергеевич

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. В. А. Серебряков

17 июня 2020 г.

Постановка задачи

Разработать алгоритм поиска по математическим формулам с использованием векторной модели.

Цели

Исследование существующих стандартов отображения формул в цифровом виде и применимости к ним векторной модели формулы. Реализация алгоритма и создание прототипа поисковой системы.

Области применения

- Поиск по учебным изданиям и публикациям для использования студентами, преподавателями и научными сотрудниками
- Проверка на плагиат содержания научных публикаций, использующих математические выражения, в случае, когда публикации сделаны на разных языках
- Обнаружение семантически схожих формул, записанных разными способами

Существующие решения

- SOJKA, Petr and Martin LÍŠKA. *The Art of Mathematics Retrieval. In Matthew R. B. Hardy, Frank Wm. Tompa. Proceedings of the 2011 ACM Symposium on Document Engineering.*, 2011

- Kenny Davila, Richard Zanibbi, Andrew Kane and Frank Wm. Tompa. *Tangent-3 at the NTCIR-12 MathIR Task.*, 2016.

Способы представления формул в публикациях

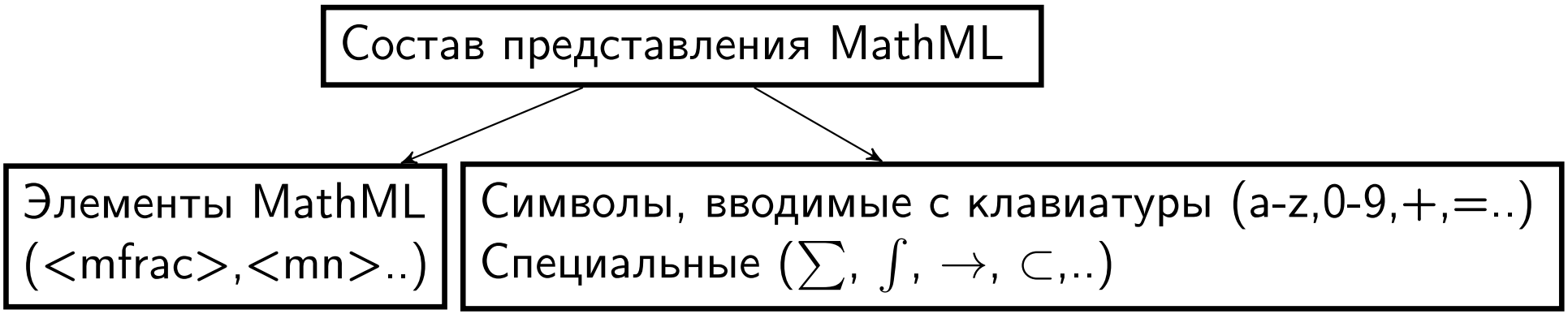
- Графическое изображение (картинка)
- Microsoft Word
- LaTeX
- MathML
  - Content MathML
  - Presentation MathML

Пример представления Presentation MathML:

Пример: $x = y + z$

```
<math>
  <mi>x</mi>
  <mo>=</mo>
  <mrow>
    <mi>y</mi>
    <mo>+</mo>
    <mi>z</mi>
  </mrow>
</math>
```

Формирование словаря



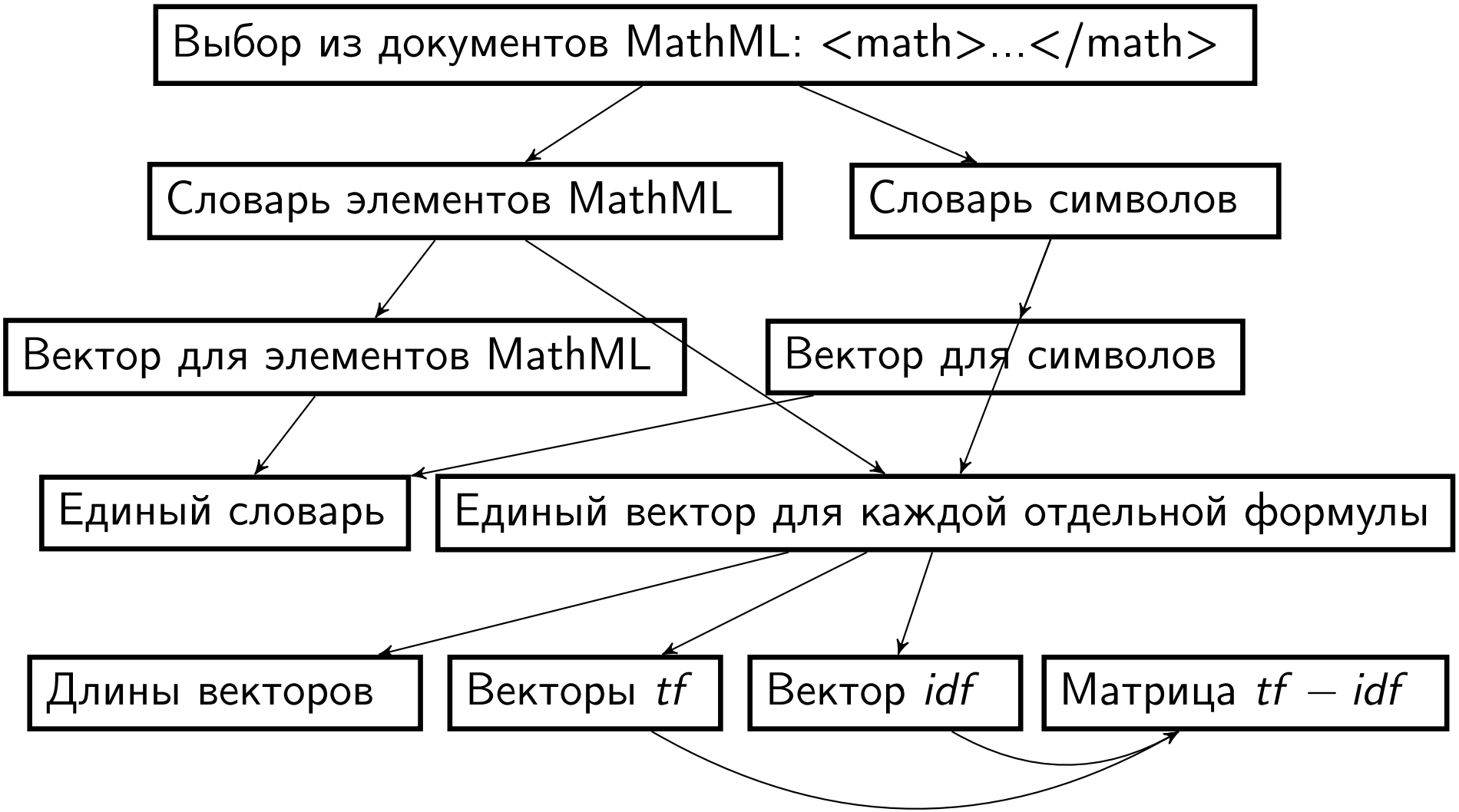
Подготовка словаря:

- Не рассматриваются стилистические элементы MathML, не несущие смысловой нагрузки
- Устранение неоднозначностей, порождаемых элементами из разных редакций MathML и ошибок ввода
- Символы в <mo> и <mi> - индивидуальные элементы, переменные унифицированы по регистру
- Унификация чисел по элементу <mn>

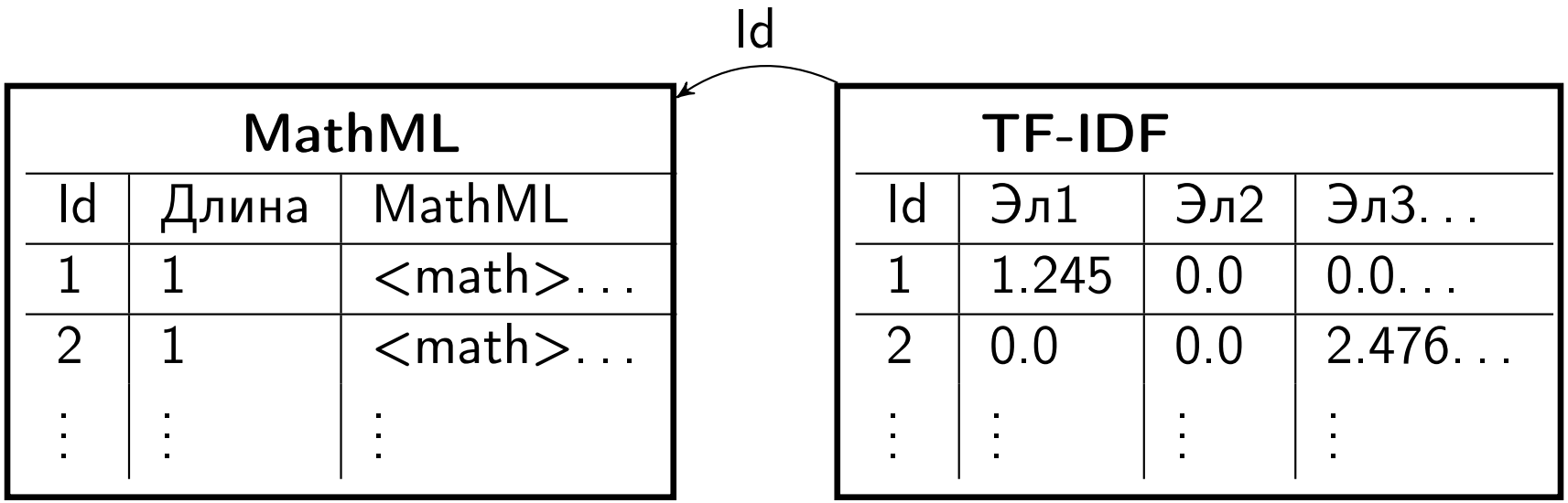
Описание представления

- Введём нормированное пространство размерности n, где n - количество элементов словаря,
- Метрика в пространстве определяется как $d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$,
- Для элементов матрицы рассчитываются веса $tf - idf$, где $tf(t, f) = \frac{n_t}{\sum_k n_k}$, $idf(t, F) = ln\frac{|F|}{|\{f_i \in F | t \in f_i\}|}$.

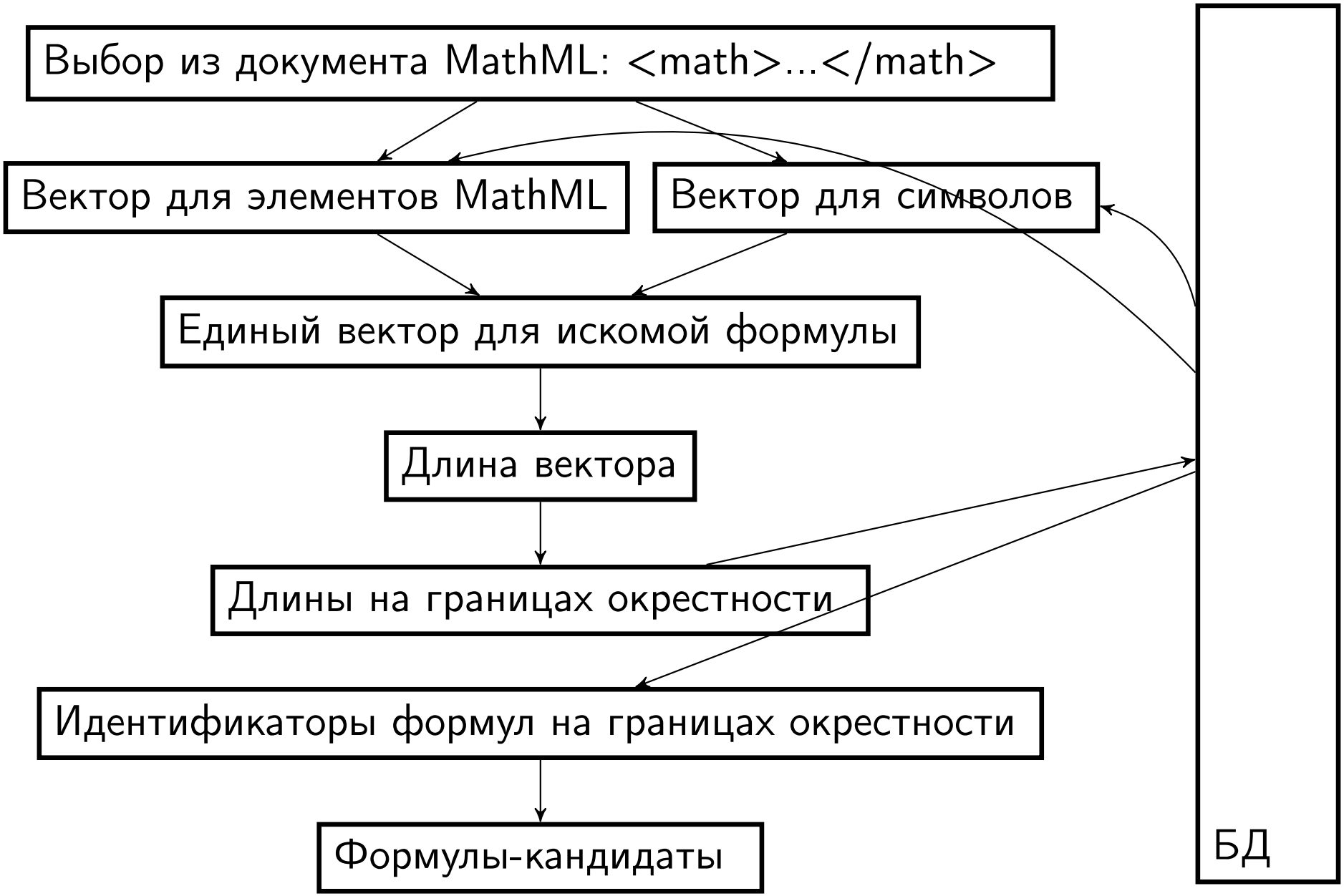
Описание алгоритма: подготовка данных



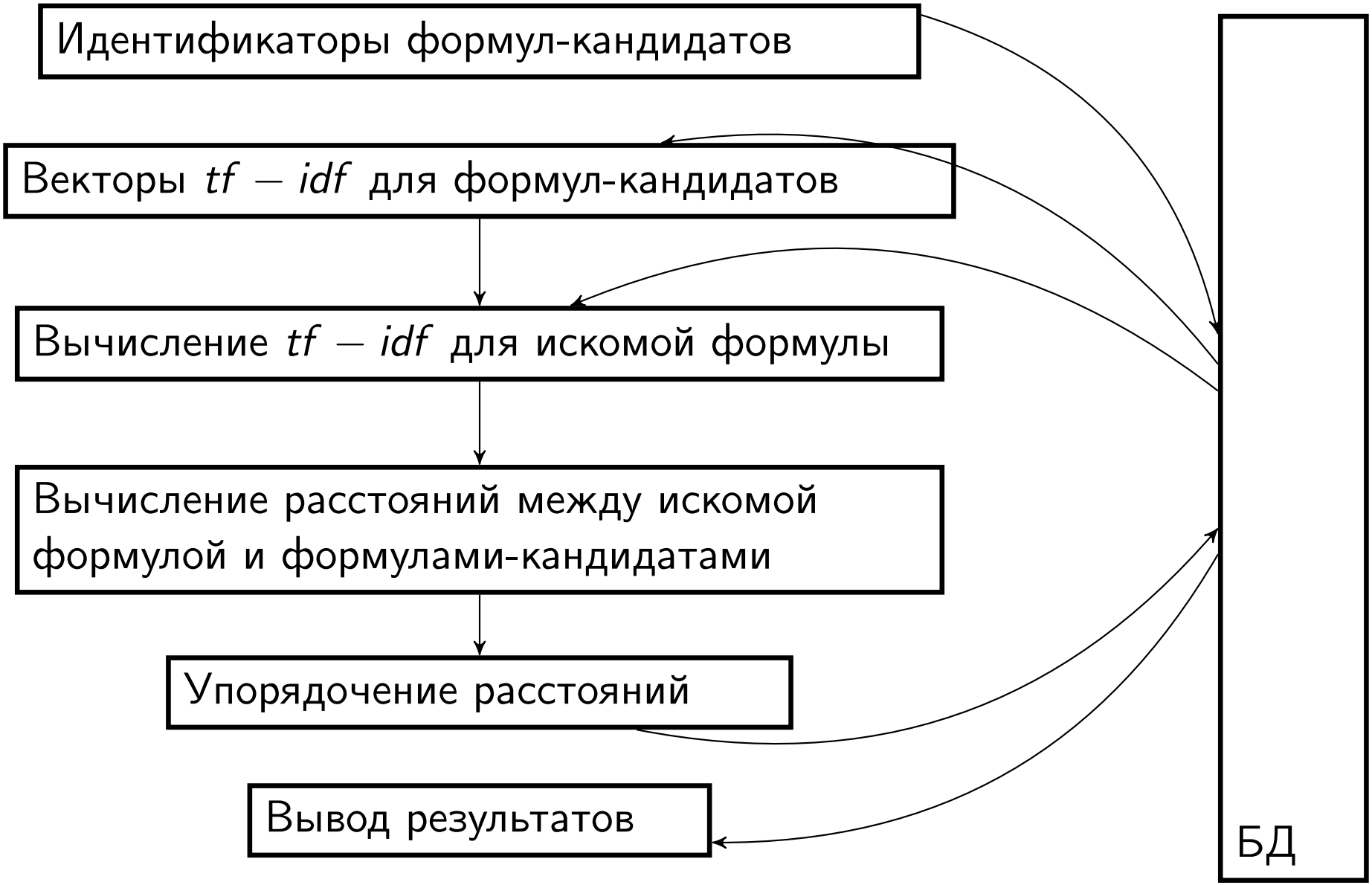
Описание алгоритма: описание хранилища



Описание алгоритма: процедура поиска



Описание алгоритма: процедура поиска



Искомая формула:

$\sum_{i=1}^{k_n} a_{ni} X_{ni} \to 0$	$\{X_n, n \geq 1\}$	$\int_0^\infty \alpha(s,x) ds$

Результаты:

$\sum_{i=1}^{k_n} a_{ni} X_{ni} \to 0$	$\{X_n, n \geq 1\}$	$\int_0^\infty \alpha(s,x) ds$
$\sum_{k=1}^{k_n} a_{nk} Z_{nk} \to 0$	$\{U_n, n \geq 1\}$	$\int_0^t R(t,s) b(s) ds$
$\sum_{i=1}^{k_n} a_{ni} X_{ni}$	$\{Y_n, n \geq 1\}$	$v = \int_0^1 f^2(x) dx$
$S_n = \sum_{i=1}^{k_n} a_{ni} X_{ni}$	$\{a_n, n \geq 1\}$	$y_n = \int_{n-1}^{n} \frac{dt}{t^{1+\alpha}}$
$h = \sum_{i=1}^{n} h_i(x) e_i$	$\{k_n, n \geq 1\}$	$C_b^\infty(M_n, x)$

Искомая формула:

$$\int\limits_0^\infty f(x)\mathrm{dx} \qquad a^2 + b^2 = c^2 \qquad \frac{a+\sqrt{b}}{c}$$

Результаты:

$$\int_0^\infty \alpha(s,x)ds \qquad x^p + y^p = z^p \qquad \frac{r}{\sqrt{1+r^4}}$$

$$\int_0^t R(t,s)b(s)ds \qquad x^{r-1} + y^r = 0 \qquad \frac{1}{\sqrt{2}}T$$

$$v = \int_0^1 f^2(x)dx \qquad e^2, e^3, e^4 \qquad \frac{r^*}{2\sqrt{2}} + i\frac{r^*}{2\sqrt{2}}$$

$$Z = \int_0^{+\infty} e^{\lambda t}\varphi_{t*}Ydt \qquad R = R^0 + \mathcal{J}R^0\mathcal{J} \qquad \frac{r^*}{\sqrt{2}} + i\frac{r^*}{\sqrt{2}}$$

$$\int\limits_\Omega M(2z(x))dx < +\infty \qquad y^3 + py + q = 0 \qquad i\frac{r^*}{\sqrt{2}}$$

Искомая формула: $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Результаты:

1. $g = \begin{pmatrix} 0 & x & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ 2. $B = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$ 3. $\left\{ \begin{pmatrix} 0 & x & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right\}$

4. $\left\{ \begin{pmatrix} 0 & x & 0 \\ 0 & 0 & x \\ 0 & 0 & 0 \end{pmatrix} \right\}$ 5. $\left\{ \begin{pmatrix} x & 0 & 0 \\ 0 & 0 & x \\ 0 & 0 & 0 \end{pmatrix} \right\}$

Выносится на защиту

1. Предложен способ формирования словаря для представления Presentation MathML и реализован алгоритм поиска по математическим формулам с использованием векторной модели, обеспечивающие высокую точность поиска.