Augmenting Mathematical Formulae for More Effective Querying & Presentation

Moritz Schubotz
Motivation

Proposition 2 (expectation value of waiting time times tunnel rate) For every PDF $f(x)$ in means of definition (0.1) the inequality

$$\langle x \rangle \langle \frac{1}{x} \rangle \geq 1$$

is valid.

I. POSITIVITY OF FANO FACTOR PARAMETERS

For every convex function $f(x)$, we have according to the Jenson inequality

$$f(\langle x \rangle) \leq \langle f(x) \rangle \quad (1)$$

That means that

$$\langle x \rangle^{-k} \leq \langle x^{-k} \rangle. \quad (2)$$

Especially $k = 1$ leads to the fact that $\alpha \geq 0$. 

26th of February 2011

18.03.-26.03.2011

28th of March
Example 1: \[ \frac{1}{\langle x \rangle} \leq \langle \frac{1}{x} \rangle \]

1. Different forms e.g. \( \langle x \rangle \left( \frac{1}{x} \right) \geq 1 \)

2. Different notations e.g.
   \[ \int_x f(x) x \, dx = \langle x \rangle \]

3. Exact match seldom

4. Ambiguity in syntax e.g. \( E \Psi = \hat{H} \Psi \)

5. no TeX-function mean

\[
\frac{1}{\text{mean } ?x} \leq \text{mean } \frac{1}{?x}
\]

\[<\text{apply}>
  \text{leq}/>
  \text{apply}>
  \text{divide}/>
  \text{apply}>
  \text{divide}/>
  \text{apply}>
  \langle qvar \rangle x \langle/ qvar \rangle</apply></apply>

\[<\text{apply}>
  \langle mean / \rangle
  \langle qvar \rangle x \langle/ qvar \rangle</apply></apply>

\[<\text{apply}>
  \langle mean / \rangle
  \langle divide / \rangle
  \langle cn type="integer" \rangle 1 </cn>
  \langle apply \rangle
  \langle mean / \rangle
  \langle divide / \rangle
  \langle cn type="integer" \rangle 1 </cn>
  \langle apply \rangle
  \langle qvar \rangle x \langle/ qvar \rangle</apply></apply>

NTCIR-11
Math-2
WMC-D1
Example 1: $\frac{1}{\langle x \rangle} \leq \langle \frac{1}{x} \rangle$

1. Different forms e.g. $\langle x \rangle \left( \frac{1}{x} \right) \geq 1$

2. Different notations e.g.
   \[ \int_X f(x) x \, dx = \langle x \rangle \]

3. Exact match seldom

4. Ambiguity in syntax e.g. $E \Psi = \hat{H} \Psi$

5. No TeX function mean

\[
\text{DRMF} \quad \text{NTCIR-11} \\
\text{WMC-D1}
\]
Result 1: $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$

\[
\varphi(x) \leq x \leq \frac{1}{\mathbb{E}[X]} \leq \mathbb{E}[\varphi(x)]
\]
Result 1: \( \varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)] \)

\[ \frac{1}{\mathbb{E}[x]} \leq \mathbb{E}[\frac{1}{\varphi(x)}] \]

Not trivial
Solution 1 inexact matches

• Refined query:

\[
\text{\textbackslash superconceptOf}[
  \text{orderby} = \text{editdistance} ]\{ \\
  \frac{1}{\text{mean } \text{x}} \leq \text{mean } \frac{1}{\text{mean } \text{x}} \\
\}
\]

• Computational complexity
• Restriction of the search space
• Check most likely solutions at first
But there are diverse information needs

1. Definition look-up
2. Explanation look-up
3. Proof look-up
4. Application look-up
5. Computation assistance
6. General formula search
, and the data looks like that

\[ \int_{-\infty}^{\infty} f(y(x))p(x)dx \geq f \left( \int_{-\infty}^{\infty} y(x)p(x)dx \right) \]

ルベーグ積分論の観点では、離散の場合も連続の場合も同一に見做せる。

証明は、\( f \)の\( \int_{-\infty}^{\infty} y(x)p(x)dx \)における接線を \( g \)とおいて、常に \( g(x) \) が \( f(x) \) よりも小さいことを使えばよい。

統計学において、式の下限を評価するさいに、一定の役割を担っている。例えば、カルバックライフライブラーダイバージェンスが常に 0 より大きいことを証明するときに用いられる。

\( p(x) \) が確率密度関数の場合を考えると、イェンゼンの不等式は次のように書ける。

\[ E[f(y)] \geq f(E[y]) \]

なお、イェンゼンの不等式から、相加相乗平均の不等式などを導くこともできる。
Levels of Abstraction

Semantic

\[ \text{Hamiltonian} \quad \text{Wave-function} \quad \text{Energy-Eigenvalue} \quad \text{Wave-function} \]

Content

\[ \hat{H} \quad \Psi \quad E \quad \Psi \]

Presentation

\[ \hat{H} \Psi = E \Psi \]
Overview

Integrated Queries

Math

Presentation

Content

Semantic

Keywords

Relations

Metadata
# Completed Research

## Querying
- Making Math Searchable in Wikipedia (CICM 2012)
- Evaluation of Similarity-Measure Factors for Formulae (NTCIR 2015)
- Wikipedia Subtask at NTCIR 11 (SIGIR 2015)
- Exploring the single-brain barrier (NTCIR 2016)

## Processing
- Mathematical Language Processing (CICM 2014)
- Digital Repository of Mathematical Formulae (CICM 2014 coauthor)
- Growing the DRMF with generic LaTeX sources
- Mathoid: Accessible Math Rendering for Wikipedia (CICM 2014)
- Semantification of Identifiers in Mathematics for Better Math Information Retrieval (SIGIR 2016)

## Scalability
- Applying Stratosphere for Big Data Analytics (coauthor, BTW 2013)
- Querying large Collections of Mathematical Publications (NTCIR 2013 with Marcus Leich)

## Integrated Queries
- Math
- Text
- Meta-data
- Presentation
- Content
- Semantic
- Keywords
- Relations

## Diagram
- Image
- Image processing
- Presentation
- Structure detection
- Content
- Entity Linkage
- Semantic
Let be a probability space, $X$ an integrable real-valued random variable and $\varphi$ a convex function. Then:

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

- convex function (Q319913, [NDL ID 00573442](#))
- subclass of function
- ja: [凸関数](#)
Exploring the single-brain barrier

• “one-brain barrier” [1]
  – Metaphor: relevant knowledge to conduct math research needs to be co-located in one brain

• Goals of our contribution to NTCIR12:
  – Create a point of reference w.r.t. to this barrier for a trained mathematician
  – Compare the performance of a human to MIR systems and analyse characteristic strengths and weaknesses
  – Derive insights to improve MIR systems
  – Combine the relevant results of the human and the MIR systems to create a gold standard
Exploring the single-brain barrier
Exploring the single-brain barrier

![Graph showing relevance score vs topic with average score (with SD) and our score]
Exploring the single-brain barrier

- **Strengths of MIR systems:**
  - Definition lookup queries
  - Application lookup

- **Weaknesses of MIR systems**
  - Low precision
  - No unified query language to specify query type

- **Gold standard dataset can help to develop a math-aware search engine for Wikipedia**
Semantification of Identifiers in Mathematics for Better Math Information Retrieval

• First step to enable computer to understand mathematicians notations
• Focus on identifiers
• Extract identifier semantics by combining math and
• Use computers to find relevant mathematics
• Computers must understand semantics in math to provide needed information
Math Augmentation Approach

1. Detect formulae
   In physics, mass-energy equivalence is a concept formulated by Albert Einstein that explains the relationship between mass and energy. It states every mass has an energy equivalent and vice versa—expressed using the formula
   \[ E = mc^2 \]
   where \( E \) is the energy of a physical system, \( m \) is the mass of the system, and \( c \) is the speed of light in a vacuum (about \( 3 \times 10^8 \) m/s). In words, energy equals mass multiplied by the speed of light squared. Because the speed of light is a very large number

2. Extract identifiers

3. Find identifiers

4. Find definients candidates
   \[ E = mc^2 \]
   where \( E \) is the energy of a physical system, \( m \) is the mass of the system, and \( c \) is the speed of light in a vacuum (about \( 3 \times 10^8 \) m/s). In words, energy equals mass multiplied by the speed of light squared. Because the speed of light is a very large number

5. Score all identifier-definients pairs

6. Generate feature vectors

7. Cluster feature vectors

8. Map clusters to subject hierarchy
In physics, mass-energy equivalence is a concept formulated by Albert Einstein that explains the relationship between mass and energy. It states every mass has an energy equivalent and vice versa—expressed using the formula

\[ E = mc^2 \]

where \( E \) is the energy of a physical system, \( m \) is the mass of the system, and \( c \) is the speed of light in a vacuum (about \( 3 \times 10^8 \) m/s). In words, energy equals mass multiplied by the.
(2) Extract identifiers

\[ E = mc^2 \]
(3) Find identifiers

\[ E = mc^2 \]

where \( E \) is the energy of a physical system, \( m \) is the mass of the system, and \( c \) is the speed of light in a vacuum (about \( 3 \times 10^8 \) m/s). In words, energy equals mass multiplied by the square of the speed of light.
(4) Find definiens candidates

\[ E = mc^2 \]

where \( E \) is the energy of a physical system, \( m \) is the mass of the system, and \( c \) is the speed of light in a vacuum (about \( 3 \times 10^8 \) m/s). In words, energy equals mass multiplied by the speed of light squared. Because the speed of light is a very large number...
(5) Score all identifier-definiens pairs

<table>
<thead>
<tr>
<th>Identifier</th>
<th>$E$</th>
<th>$m$</th>
<th>$c$</th>
<th>$m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>energy</td>
<td>2</td>
<td>4</td>
<td>13</td>
<td>22</td>
</tr>
<tr>
<td>physical system</td>
<td>5</td>
<td>1</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>mass</td>
<td>10</td>
<td>2</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>speed of light</td>
<td>19</td>
<td>11</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>vacuum</td>
<td>21</td>
<td>14</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>
(6) Generate feature vectors

\[
\begin{pmatrix}
E_{\text{energy}} \\
E_{\text{expected\_value}} \\
m_{\text{mass}} \\
m_{\text{natural\_number}} \\
\vdots
\end{pmatrix}
\begin{pmatrix}
d_1 \\
\vdots \\
\end{pmatrix}
\begin{pmatrix}
d_2 \\
\vdots \\
\end{pmatrix}
\begin{pmatrix}
2 \\
d \_ \_ \\
2 \\
\vdots \\
\end{pmatrix}
\begin{pmatrix}
2 \\
d \_ \_ \\
- \\
\vdots \\
\end{pmatrix}
\begin{pmatrix}
\_ \\
\vdots \\
\end{pmatrix}
\begin{pmatrix}
\_ \\
\vdots \\
\end{pmatrix}
\begin{pmatrix}
- \\
\vdots \\
2 \\
\vdots \\
\end{pmatrix}
\begin{pmatrix}
- \\
\vdots \\
0 \\
\vdots \\
\end{pmatrix}
\begin{pmatrix}
\_ \\
\vdots \\
\end{pmatrix}
\begin{pmatrix}
\_ \\
\vdots \\
\end{pmatrix}
\]

(7) Cluster feature vectors
(8) Map clusters to subject hierarchy
Wikipedia Subtask at NTCIR 11

- NTCIR 11 Wikipedia dataset*
- 30k Wikipedia Articles
- 280k Formulae
- 100 queries

Wikipedia Subtask at NTCIR 11

- CICM 2012 (2 Participants)
- NTCIR 2013 (pilot) (6 Participants)
- NTCIR 2014 arXiv (8 Participants)  Wikipedia (7 Participants)
Wikipedia Task results

Part III / IV Completed Research
(1) Van der Waerden’s theorem: $W(2,k) > 2^k / k^\varepsilon$

$W$  Van der Waerden number

$k$  integer : number that can be written without a fractional or decimal component

$\varepsilon$  positive number (real number . . . )

(69) Engine efficiency: $\eta = \frac{\text{work done}}{\text{heat absorbed}} = \frac{Q_1 - Q_2}{Q_1}$

$\eta$  energy efficiency

$Q_1$  heat (energy)

$Q_2$  heat (energy)
Gold standard details

310 Identifiers

- Wikidata item: 174
- Two Wikidata items: 8
- Wikidata item + NP: 97
- Individual NP: 27
- Multiple NP: 4

31/07/2016
Results: Identifiers

- 294/310 correctly extracted (94.8%)
- 57 false positive (fp)
- Problems
  - Incorrect markup (8fn, 33fp) \( \eta = \frac{Q_1 - Q_2}{Q_1} \)
  - Symbols (9fp) \( \frac{d}{dx} \)
  - Sub-super script (3fp, 2fn) \( \sigma^2_y \)
  - Special notation (10fp, 2fn) \( \mathbf{u} \times \mathbf{v} = \epsilon^i_{\ jk} u^j v^k e_i \)
$E = mc^2$

where $E$ is the energy of a physical system, $m$ is the mass of the system, and $c$ is the speed of light in a vacuum (about $3 \times 10^8$ m/s). In words, energy equals mass multiplied by the speed of light squared. Because the speed of light is a very large number...
Distribution of identifier counts
Results: Definitions with Namespace support

Definitions

- Exact match: 147
- Partial match: 103
- not found: 60

Legend:
- Green: Exact match
- Yellow: Partial match
- Orange: not found
Discovered namespaces

- 250 clusters -> 167 mapped to classification schemata
- 5618 definitions with $s > 1$ (2124 Wikidata concepts)
- Evaluate 6 randomly sampled namespaces
Impression from namespace samples

Definitions in Namespaces

- Correct: 144
- Wrong: 8
- Unspecific: 7
- Cannot say: 129

31/07/2016 www.formulasearchengine.com 40
Classical mechanics of discrete systems 45.00 (PACS)
Categories: Physics, Mechanics, Classical mechanics
Purity: 61%, matching score: 31%,
identifiers 103, semantic concepts 50, Ø 58, ✓ 4, ? 42, X 1
Identifier-definitions:
m mass (quantitative measure of a physical object’s
resistance to acceleration by a force . . . ) [s≈29] Ø
F force (influence that causes an object to change) [s≈25] Ø
v velocity (rate of change of the position of an object
. . . and the direction of that change) [s≈24] Ø
t time (dimension in which events can be ordered the past
through the present into the future) [s≈19] Ø

Stochastic analysis 60Hxx (MSC)
Categories: Stochastic processes, Probability theory
Purity: 92%, matching score: 62%, identifiers 54, semantic
concepts 32, Ø 18, ✓ 0, ? 30, X 0
Identifier-definitions:
a stochastic process ( . . . random variables) [s≈12] Ø
X stochastic process ( . . . random variables) [s≈10] Ø
. . .
E expected value [s≈2] Ø
. . .
E expected value s < 1
v function s < 1

• Physics
• Identifiers are significant for formulae

• Mathematics
• Identifiers might be less significant for formulae
Conclusions

• For 10% of the identifiers used in Wikipedia (en) we could assign the associated Wikidata item

• → 90% ahead
  – More specific Wikidata items needed
  – Combine data from different language versions
  – Improve recognition rate within a document

• Namespaces for mathematical identifiers could be identified
Next steps

• The identifier information is available from the Wikipedia API today [http://en.wikipedia.org/api](http://en.wikipedia.org/api)

• Develop tools to augment the user experience for math in Wikipedia and beyond
  – Tooltips (ongoing)
  – Physical Dimensions (ongoing)
  – Translation to Computer Algebra Systems (started)
  – Math Question and Answering (ongoing)
  – Author assistance (ongoing with WMF)
  – Related formulae search (started)

• →Justification for semantification effort
Contact

Moritz Schubotz (now at Universität Konstanz)

moritz@schubotz.de

+49 7531 88 4438

Mobile: +49 1578 047 1397

www.isg.uni-konstanz.de

www.formulasearchengine.com