NTCIR-12 MathIR Wikipedia Formula Browsing Subtask Queries (NTCIR12-MathWikiFormula)

Richard Zanibbi (rlaz@cs.rit.edu) **(Jan. 2015)**

How users perceive formula similarity is an important open question in Mathematical Information Retrieval (MIR). In the NTCIR-12 MathIR Wikipedia Formula Browsing Subtask (NTCIR12-MathWikiFormula), we imagine an undergraduate or graduate university student is browsing through formulae in Wikipedia, using a search interface that displays a ranked list of formulae in the results. A student might use a formula query to re-locate formulae or articles seen previously (i.e. a navigational query), find articles providing definitions for a given formula, find articles discussing concepts pertinent to the query, or find formulae related to the query formula without a specific goal in mind (i.e. 'browsing' in its truest sense).

In this sub-task we consider a browsing scenario, where a student wishes to find expressions similar to a query formula based on appearance (i.e. symbols and structure), mathematical content (i.e. operators and operands), or a combination of these factors. **There are 40 queries:** 20 formulae are sampled directly from the NTCIR-12 Wiki corpus, and each formula then has one or more subexpressions removed or replaced by wildcard symbols (wildcards are described below).

For the concrete formula queries, we can imagine our student copying-and-pasting a formula from a Wikipedia article into a search box. For wildcard queries, the student might create a formula using placeholders for subexpressions (e.g. when the student forgets specific symbols, or subexpressions), or modifies a formula seen in an article before initiating search, inserting wildcards in the formula's associated LaTeX string, and/or using an equation editor provided by the search interface.

Indexing and Output. Participants should index **formulae contained within `<math>` tags from the NTCIR-12 Wikipedia corpus**. In search results, systems should return individual formulae at specific article locations, using formula identifiers (these are given in the 'id' attribute of `<math>` tags).

Query Language. Each query in this task is an isolated formula. The query language is identical to that for the main NTCIR-12 MathIR Wikipedia task.[1] Each query formula may contain one or more wildcard symbols that represent a subexpression (e.g. $*1*$), and repeated wildcards should be bound to the same subexpression. For example, given query '$*1* + *1*$', the formulae '$\frac{1}{2} + \frac{1}{2}$' and '$a + a$' are valid wildcard matches, but '$a + b$' is not. If no wildcard symbols are repeated, then there are no constraints on wildcard substitutions (e.g. '$*1* + *2*$' may match: '$a + b$', '$\frac{1}{2} + \frac{1}{2}$' or '$\frac{3}{4} + \lambda$').

Relevance Assessment. Search hits will be graded by human evaluators. Evaluators will use a three-point scale to rate the similarity of each returned formula to the query formula. Please note that relevant hits for a formula can change dramatically after removing or replacing subexpressions by wildcards (e.g. '$f(x, y, z) = z\frac{x}{y}$' vs. '$f(*1*) = *2*$' or 'f(*1*)'). **We expect at least the top-5, but at most the top-10 formulae returned by each system to be evaluated.**

Each search hit (formula) that an evaluator assesses will be given one of the following scores:

2. **(Relevant - 'Highly Similar')** Largely or completely matches the appearance, mathematical content, or both for the query formula
1. **(Partially relevant - 'Somewhat Similar')** Some non-trivial parts of the appearance and/or mathematical content are similar to the query formula
0. **(Not relevant - 'Unrelated')** Not similar to the query formula

There will be two evaluators; each search hit's score will be the sum of the evaluator scores. To avoid bias in the design of participant algorithms, queries but *not* the source of the query formulae are visible to participants. Query formula sources will be available sometime after the NTCIR12-MathWiki task has finished.

1. Please consult the document *NTCIR-12 MathIR Wikipedia Task Queries (NTCIR12-MathWiki) for details*

——— NTCIR12-MathWikiFormula Subtask Queries ———

——— Concrete Formulae (20 queries) ———

(NTCIR12-MathWikiFormula–1)

formula $-0.026838601\ldots$

(NTCIR12-MathWikiFormula–2)

formula $\mathfrak{P}$

(NTCIR12-MathWikiFormula–3)

formula $N = \left\lfloor 0.5 - \log_2 \left(\frac{\text{Frequency of this item}}{\text{Frequency of most common item}} \right) \right\rfloor$

(NTCIR12-MathWikiFormula–4)

formula $\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \underbrace{\mu_0 \epsilon_0 \frac{\partial}{\partial t} \mathbf{E}}_{\text{Maxwell's term}}$

(NTCIR12-MathWikiFormula–5)

formula $1 + \cfrac{1}{2 + \cfrac{1}{5 + \cfrac{1}{5 + \cfrac{1}{4 + \ddots}}}}$

(NTCIR12-MathWikiFormula–6)

formula ${}^{238}_{92}\mathrm{U} + {}^{64}_{28}\mathrm{Ni} \rightarrow {}^{302}_{120}\mathrm{Ubn}^{*} \rightarrow$ *fission only*

(NTCIR12-MathWikiFormula–7)

formula $0 \rightarrow G^{\wedge} \xrightarrow{\pi^{\wedge}} X^{\wedge} \xrightarrow{\imath^{\wedge}} H^{\wedge} \rightarrow 0$

(NTCIR12-MathWikiFormula–8)

formula $w = \begin{cases} w^* & \text{if } w^* > \frac{1}{2}, \\ \frac{1}{2} & \text{if } w^* \leq \frac{1}{2}. \end{cases}$

(NTCIR12-MathWikiFormula–9)

formula $\begin{bmatrix} V_1 \\ I_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} I_1 \\ V_2 \end{bmatrix}$

(NTCIR12-MathWikiFormula–10)

formula $L(\lambda, \alpha, s) = \sum_{n=0}^{\infty} \frac{\exp(2\pi i \lambda n)}{(n+\alpha)^s}.$

(NTCIR12-MathWikiFormula–11)

formula $ax^2 + bx + c = 0$

(NTCIR12-MathWikiFormula–12)

formula $O(mn \log m)$

(NTCIR12-MathWikiFormula–13)

formula $A \oplus B = (A^c \ominus B^s)^c$

(NTCIR12-MathWikiFormula–14)

formula $\cos\alpha = -\cos\beta\cos\gamma + \sin\beta\sin\gamma\cosh\frac{a}{k},$

(NTCIR12-MathWikiFormula–15)

formula $\forall x, y \in A\ [x \neq y \rightarrow \neg\exists z \in X\ [z \leq x \wedge z \leq y]].$

(NTCIR12-MathWikiFormula–16)

formula $\tau_{\mathrm{rms}} = \sqrt{\frac{\int_0^\infty (\tau - \overline{\tau})^2 A_c(\tau) d\tau}{\int_0^\infty A_c(\tau) d\tau}}$

(NTCIR12-MathWikiFormula–17)

formula $x - 1 - \frac{1}{2} - \frac{1}{4} - \frac{1}{5} - \frac{1}{6} - \frac{1}{9} - \cdots = 1$

(NTCIR12-MathWikiFormula–18)

formula $P_i^x = \frac{N!}{n_x!(N-n_x)!} p_x^{n_x} (1-p_x)^{N-n_x}$

(NTCIR12-MathWikiFormula–19)

formula $H_{ij} = \begin{bmatrix} \frac{\partial^2 V_{ij}}{\partial x_i \partial x_j} & \frac{\partial^2 V_{ij}}{\partial x_i \partial y_j} & \frac{\partial^2 V_{ij}}{\partial x_i \partial z_j} \\ \frac{\partial^2 V_{ij}}{\partial y_i \partial x_j} & \frac{\partial^2 V_{ij}}{\partial y_i \partial y_j} & \frac{\partial^2 V_{ij}}{\partial y_i \partial z_j} \\ \frac{\partial^2 V_{ij}}{\partial z_i \partial x_j} & \frac{\partial^2 V_{ij}}{\partial z_i \partial y_j} & \frac{\partial^2 V_{ij}}{\partial z_i \partial z_j} \end{bmatrix}$

(NTCIR12-MathWikiFormula–20)

formula $r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1) s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}},$

———— Queries with Wildcards (20 queries) ————

(NTCIR12-MathWikiFormula–21)

formula $-0.*1*\ldots$

(NTCIR12-MathWikiFormula–22)

formula $\mathfrak{P}^{*1*}$

(NTCIR12-MathWikiFormula–23)

formula $*1* \left(\frac{\text{Frequency } *2*}{\text{Frequency } *3*} \right)$

(NTCIR12-MathWikiFormula–24)

formula $\underbrace{*1*}_{*2*}$

(NTCIR12-MathWikiFormula–25)

formula $1 + \cfrac{1}{*1* + \cfrac{1}{*2*}}$

(NTCIR12-MathWikiFormula–26)

formula ${}^{238}_{92}\mathrm{U} + *1* \rightarrow *2*$

(NTCIR12-MathWikiFormula–27)

formula $*1*^{\wedge} \overset{*2*^{\wedge}}{\rightarrow} *3*^{\wedge}$

(NTCIR12-MathWikiFormula–28)

formula $*1* = \begin{cases} *2* & \text{if } *3* > \frac{1}{2} \\ *4* & \text{if } *3* \leq \frac{1}{2} \end{cases}$

(NTCIR12-MathWikiFormula–29)

formula $*1* = \begin{bmatrix} *2* & *3* \\ *4* & *5* \end{bmatrix} \begin{bmatrix} *6* \\ *7* \end{bmatrix}$

(NTCIR12-MathWikiFormula–30)

formula $L(*1*, *2*, *3*) = \sum_{*4*=0}^{\infty} \frac{\exp(2\pi i *1* *4*)}{(*4* + *2*)^{*3*}}$

(NTCIR12-MathWikiFormula–31)

formula $*1* x^2 + *2* x + *3* = 0$

(NTCIR12-MathWikiFormula–32)

formula $O(*1* \log *2*)$

(NTCIR12-MathWikiFormula–33)

formula $A \ *1* \ B = (*2*)^{*3*}$

(NTCIR12-MathWikiFormula–34)

formula $*1*\alpha = -*1*\beta*1*\gamma + *2*\beta*2*\gamma*3*$

(NTCIR12-MathWikiFormula–35)

formula $\forall x, y \in *1* [x \neq y \rightarrow *2*]$

(NTCIR12-MathWikiFormula–36)

formula $*1* = \sqrt{\frac{\int_0^\infty *2**2**3* \ d*1*}{\int_0^\infty *3* \ d*1*}}$

(NTCIR12-MathWikiFormula–37)

formula $\frac{1}{*1*} - \frac{1}{*2*} - \frac{1}{*3*} - \frac{1}{*4*} - \frac{1}{*5*}$

(NTCIR12-MathWikiFormula–38)

formula $\frac{N!}{*1*!(N-*1*)!} p^{*1*} (1-p)^{N-*1*}$

(NTCIR12-MathWikiFormula–39)

formula $H_{*1*} = \left[\frac{\partial^2 *2*}{*3*} \right]$

(NTCIR12-MathWikiFormula–40)

formula $\frac{\sum_{i=1}^{n} (*1*_i - *\bar{1}*)(*2*_i - *\bar{2}*)}{\sqrt{\sum_{i=1}^{n} (*1*_i - *\bar{1}*)^2 \sum_{i=1}^{n} (*2*_i - *\bar{2}*)^2}}$
