Evaluation of Similarity-Measure Factors for Formulae based on the NTCIR-11 Math Task

Moritz Schubotz
Database Systems and Information Management Grp.
Technische Universität Berlin
Berlin, Germany
schubotz@tu-berlin.de

Abdou Youssef
Department of Computer Science
The George Washington University
Washington, DC 20052
ayoussef@gwu.edu

Volker Markl
Database Systems and Information Management Grp.
Technische Universität Berlin
Berlin, Germany
volker.markl@tu-berlin.de

Howard S. Cohl
Applied and Computational Mathematics Division
National Institute of Standards and Technology
Gaithersburg, Maryland
20899-8910
howard.cohl@nist.gov

Jimmy J. Li
Richard Montgomery H.S.
250 Richard Montgomery Dr.
Rockville, Maryland 20852
jimmy.jiajian.li@gmail.com

ABSTRACT

In this paper we evaluate the similarity-measure factors proposed by Zhang and Youssef based on the NTCIR-11 gold standard. In contrast to Zhang and Youssef we evaluate them individually. The evaluation indicates that four of five factors are relevant. The fifth factor alone is of lower relevance than the other four factors. However, we do not prove that the fifth factor is irrelevant.

Team Name

Formula Search Engine (FSE)

Keywords

Math Search, MathML, Apache Flink, Query Language, Math Similarity-Measure

1. INTRODUCTION

In our Formula Search Engine (FSE) team contribution for NTCIR-10, we developed a concept which we claimed would enhance math search research significantly [5]. The idea is to separate questions regarding *big data processing* from conceptual questions regarding Math search. This leads to an accelerated development cycle because the processing in regard to math search is not distracted by data organization efforts. In this paper, we take advantage of this accelerated development cycle to evaluate the similarity-measure factors for formulae recently proposed in Zhang & Youssef (2014) [7] based on the NTCIR-11 data set.

This paper makes two major contributions. First, using the large human-generated ground truth in NTCIR (i.e. 2500 manually evaluated document sections), this paper performs a broader evaluation of Zhang and Youssef's similarity search than the one reported in [7]. That is the larger and standardized NTCIR test collection is being used, rather than a hand crafted subset of the Digital Library of Mathematical Formulae. Second, this paper evaluates the contribution of each individual similarity factor out of the five factors identified by Zhang and Youssef. In [7], the similarity measure combines all five factors into a single metric, which was evaluated collectively. In contrast, this paper evaluates the impact of each factor separately, thus providing a more fundamental insight into the contributions of each factor, and leading the way for a more targeted fine-tuning of similarity-search parameters and thus to better optimization of math similarity search.

2. FORMULA SIMILARITY SEARCH

Exact formula search queries can be formalized using languages such as XQuery, XPath, or the MathWebSearch α-equivalence language concept [3]. The result set depends on the data only, and if implemented correctly, is independent of the realization in the language used. This method works well for queries that lead to few results (on the order of 10).

For larger result sets, usual query refinement techniques (such as [2, 4]) known from traditional, general-purpose databases, can be applied in order to reduce the result set size. The authors are not aware of any math query refinement techniques specific to exact search.

In contrast, the concept of *similarity based* search (hereafter similarity search) is that a score is assigned to the Cartesian product of the formulas and search patterns. For each query, a partially score-ordered result set is returned. Since the score calculation might be computationally expensive, approximations to this exact scoring method are usually used. Note that in the worst case, the score for a pattern-formula tuple depends on the full collection of formulae in the data set. It is common that this score will depend on a set of aggregated values derived from the data set. An example of aggregated values are the frequencies of variable occurrence. Regardless of the technical aspects, there is no established way to define similarities between formulae. For example one system which uses similarity search is MIaS [6].

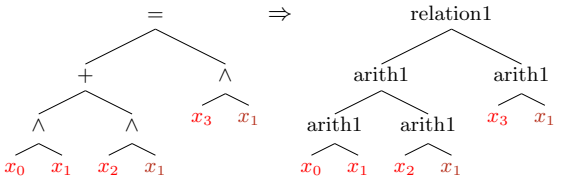

Figure 1: Projection to the Content Dictionary dimension for the search pattern $x_0{}^{x_1} + x_2{}^{x_1} = x_3{}^{x_1}$.

3. SIMILARITY-MEASURE FACTORS

The following Math similarity-measure factors are listed and explained in [7]:

1. Taxonomic Distance based on Content Dictionaries;
2. Data-Type Hierarchical Level;
3. Match-Depth;
4. Query Coverage; and
5. Formula vs. Expression.

Factor 1 assumes a taxonomy of functions, and assigns more similarity between two functions if they belong to the same taxonomic class (e.g., if both are trigonometric functions), and less similarity if the functions belong to two different classes, within a larger super-class (e.g., one trigonometric function and the logarithm, being in different classes but within the super-class of elementary functions). *Factor 2* assumes a hierarchy of math objects, such as constants (level 0), variables (level 1), functions and operations (level 2), functionals like integration and differentiation (level 3), and so on. The higher the levels of two math objects are, the more weight is assigned to their similarity/dissimilarity. *Factor 3* assigns a larger distance (less similarity) between a query expression/formula Q and a hit expression/formula E when Q is more deeply nested in E. For example, if Q is x^2+y^2, E_1 is x^2+y^2+2xy, and E_2 is $\exp\left(\frac{1}{x^2+y^2+5}\right)$, then Q is assumed to be more similar to E_1 than to E_2. *Factor 4* measures how much of a query Q is present in a potential hit E: the more terms and structure of Q there is in E, the more similarity is assigned between Q and E. Finally, *Factor 5* assigns more weight to hits that are formulas (involving a comparison operator) such as "$\sin^2 x + \cos^2 x = 1$" than to non-formula expressions like "$\sin x + \cos y$".

In our evaluation, we treat each factor as a separate measure and we qualify these factors in the following way. For Factor 1, we perform a reduction of the math objects and search pattern by projecting patterns and formulae to the Content Dictionary dimension. For example, we replace arithmetic operators, such as those in $\{+, -, *, /\}$, by their Content Dictionary `arith1` (cf. Figure 1) to generalize to the taxonomic class.

For Factor 2, we do the same thing but with regard to the data-type dimension. If a reduced pattern matches a reduced formula, we call this a generalized hit. We count the number of generalized hits with regard to the assessor ranking v.

For Factor 3, we calculate the Match-Depth penalty factor (or level-coefficient) [6]. Note that since only 8 of the 55 given search patterns contain exact matches at any depth, the sample size for this evaluation is significantly smaller. Average relevances are calculated for all Match-Depth penalty factors.

For Factor 4, queries and formulas are converted to bags of tokens. We compare each pattern with each formula and count the number of tokens from each pattern which are also tokens of the formulae. We normalize this to the total number of tokens with regard to pattern. In a subsequent step, we group (i.e., quantize) the results into 11 buckets (0-5%, 15-25%, ... 85-95%, 95-100%) and calculate the average relevance ranking for each bucket.

For Factor 5, we apply a method similar to measure factors 1 and 2. For every math object, we determine if it is in the Formula category or in the Expression category. The math object is in the Formula category if it contains a relational operator at the root level, and otherwise is in the Expression category. The following set of relational operators were considered as indicators for the Formula category $\{=, \equiv, \neq, <, >, \leq, \geq\}$.

4. EVALUATION

For the evaluation, we used the document sections originating from the `arXiv`. Those were selected in the pooling process of the NTCIR-11 task. We refer to this data set as the gold standard data set. For each of the 50 topics defined in the NTCIR-11 Math Task, human assessors assigned relevance rankings to 50 document sections. This leads to a collection of 2429 distinct `arXiv` document sections with 5 ranking categories from 0 (not relevant) to 4 (most relevant). For more details concerning the evaluation process, refer to the NTCIR-11 Math overview paper [1]. Our evaluation deals with similarity measures for individual formulae. Most of the 2250 document sections that contain mathematical expressions have more than one math expression. This is unfortunate for the task at hand, since our similarity factors are formula-centric and not document section based. Furthermore, the consideration of keywords that are included in the topics in addition to the formula search patterns listed in table 3 add an additional source of error for our evaluation. However, due to the reasonable size of the data set, those effects might average out. Thus, we are still able to show the correlation between relevance ranking and the presence of similarity factors.

We mapped the relevance ranking for a document section to all mathematical objects contained in that section. For articles with more than one formula, this ensured that the formula that leads to classification of the article as relevant, is also marked as relevant. For example, if a document with two mathematical objects was considered as relevant to a particular topic by the assessors, both formulas are considered as relevant with regard to all math search patterns occurring in this topic. Forty Seven topics include only one search pattern, two topics include two search patterns, and topic 48 contains 4 search patterns (see table 3). The downside of this approach is that formulae that are in the same article and did not influence the ranking of the assessor in a positive way, were marked as relevant, even though they are not relevant.

For our implementation, we used Apache Flink with the Java API. We published our source code on github.com under the code name `mathosphere2`. Since all the algorithms described here are embarrassingly parallel, the required runtime for a fixed number of formulae scales almost linearly with the available computational resources. This demon-

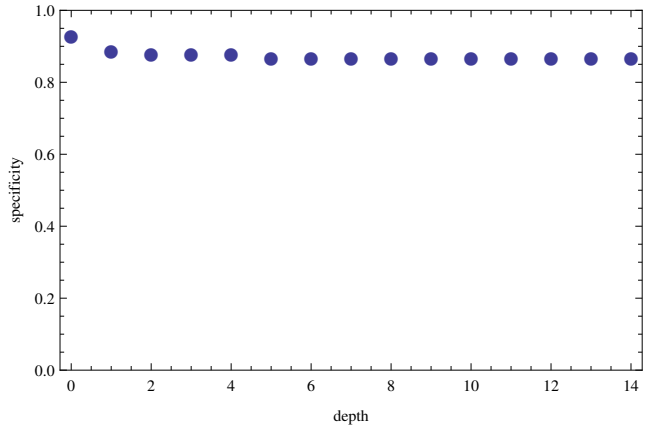

Figure 2: Specificity versus match depth.

strates that the factors evaluated can be used in an interactive application. The remainder of this section presents the evaluation results for each individual factor.

4.1 Taxonomic Distance and Data-Type

The 443 (449) matches for Content Dictionary (data-type) abstraction have the following respective recall, precision and specificity metrics for both factors

$$r = 0.27, \quad p = 0.74, \quad s = 0.91. \tag{1}$$

The high specificity shows that both are relevant similarity factors.

4.2 Match Depth

Only 10 of 55 math patterns (namely 1f1.0, 1f1.1, 12f1.0, 13f1.0, 15f1.0, 18f1.0, 20f1.0, 38f1.0, 45f1.0, 47f1.0 in table 3) contain exact matches. For the 9 underlying topics 479 pages were evaluated by the assessors. We considered documents that contain exact matches to the formula pattern as retrieved documents and calculated the match depths for them.

We measured the recall, precision and specificity for all 10 search patterns and for match depth from 0 to 14. The results are averaged over all the 10 search patterns for each depth, and are presented in Table 1. As evident in Table 1 and Figure 2, the specificity is very high, and it is higher for smaller depths. This indicates that the match depth factor is a relevant similarity factor, and confirms that the smaller the depth, i.e., the less deeply nested the match, the higher the relevance.

4.3 Coverage

To calculate the coverage factor, we took the maximum value of the coverages over all (search-pattern, formula) pairs within a document section. For each coverage level (from 0 to 10), we compute average recall, precision and specificity over all 55 search patterns. The results are presented in Table 2.

As Table 2 and Figure 3 show, the specificity is higher for higher levels of coverage, thus showing that coverage is a relevant similarity measure. Notice that the specificity of the coverage factor is not as high as the specificity of the earlier factors. This could be attributed to the coverage being insensitive to the mathematical structure of expressions.

Table 1: This table lists the match depth d, average recall r, average precision p, and average specificity s, all averaged over 10 search patterns.

d	r	p	s
0	0.15	0.72	0.93
1	0.24	0.67	0.89
2	0.32	0.78	0.88
3	0.38	0.82	0.88
4	0.40	0.82	0.88
5	0.40	0.74	0.87
6	0.40	0.74	0.87
10	0.40	0.74	0.87
13	0.40	0.74	0.87
14	0.40	0.74	0.87

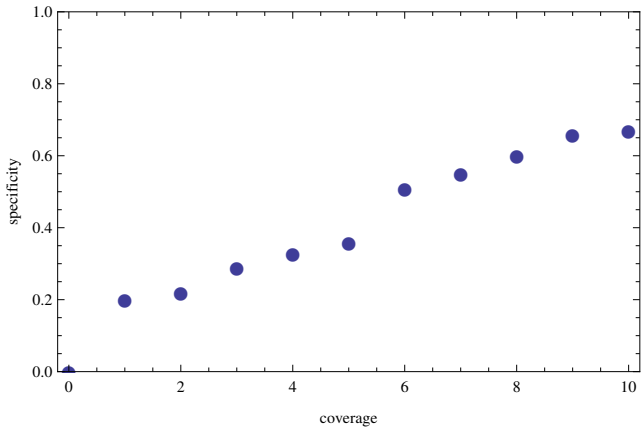

Figure 3: Specificity versus query coverage.

This insensitivity makes coverage a less important factor.

4.4 Formula vs. Expression

In order to test the hypothesis that mathematical objects classified as formulae are more relevant compared to non-formula expressions, the search returns all and only articles containing at least one formula.

For this factor we found that the average recall, precision and specificity over all 55 search patterns are:

$$r = 0.28, \quad p = 0.49, \quad s = 0.26. \tag{2}$$

The low specificity shows that many sections that contain expressions but not formulae have been considered as relevant by the assessors. This seems to indicate that this factor is of lower relevance to search ranking than the other 4 factors considered.

5. CONCLUSION AND OUTLOOK

The NTCIR-11 data set provides a good basis for our evaluation. We have found good evidence that four out of five factors are relevant. For the nominal factors 3 (match depth) and 4 (coverage), we demonstrated (anti)-correlation

Table 2: This table lists precision p, recall r, and specificity s, depending on the coverage category c.

c	r	p	s
0	1.00	0.48	0.00
1	0.79	0.48	0.20
2	0.77	0.48	0.22
3	0.69	0.47	0.29
4	0.65	0.47	0.33
5	0.62	0.47	0.36
6	0.50	0.48	0.51
7	0.47	0.49	0.55
8	0.37	0.46	0.60
9	0.34	0.48	0.66
10	0.31	0.46	0.67

to the specificity. This indicates that these factors can be used for result ranking in Math Information Retrial systems. It is not very surprising that the measured categorical values for content and data-type abstraction are almost identical. While these abstractions differ in their conceptual background, their actual implementations are similar. This justifies the approach of Zhang and Youssef to combine both factors and use the taxonomic distance to compare nodes of data type function only. However, with special regard to query refinement and content summarization (which are not part of the task), further research in this direction is needed. At this point, we also note that the aforementioned factors heavily rely on high quality content MathML. Even though the content MathML automatically generated by LaTeXML is a decent starting pointing, we still see improvement potential here.

We have observed two points which will help improve the analysis of our similarity factors and formula-centric math information retrieval systems which incorporate a combination of these factors. The retrieval units in the Math Task are document sections, not individual equations. In our evaluation, we used the best matching result if there was more than one formula. This assumption could easily be dropped if the retrieval unit was more fine-grained. In the NTCIR-11 Math Task, the influence of keywords was considered. Since the analysis presented in this paper does not take keywords into account, the influence of the keywords to the relevance ranking adds random noise from the similarity factor viewpoint. In NTCIR-11, there was a new experimental Wikipedia Subtask. Both of these weak points are not available in the Wikipedia Subtask and therefore we are looking forward to future NTCIR conferences which might incorporate the Wikipedia task as a second main task, and provide human evaluation for the Wikipedia query results.

Acknowledgments. Thanks to Akiko Aizawa, Michael Kohlhase and Bruce Miller for fruitful discussions. Thanks to Holmer Hemsen for proofreading and valuable feedback.

6. REFERENCES

[1] Akiko Aizawa, Michael Kohlhase, Iadh Ounis, and Moritz Schubotz. NTCIR-11 Math-2 Task Overview. In *NTCIR Workshop 11 Meeting*, Tokyo, Japan, 2014.

[2] Surajit Chaudhuri, Gautam Das, Vagelis Hristidis, and Gerhard Weikum. Probabilistic information retrieval approach for ranking of database query results. *ACM Transactions on Database Systems*, 31(3):1134–1168, September 2006.

[3] Michael Kohlhase and Ioan Sucan. A Search Engine for Mathematical Formulae. In Tetsuo Ida, Jacques Calmet, and Dongming Wang, editors, *Proceedings of Artificial Intelligence and Symbolic Computation, AISC'2006*, number 4120 in LNAI, pages 241–253. Springer Verlag, 2006.

[4] Chaitanya Mishra and Nick Koudas. Interactive query refinement. In *Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology - EDBT '09*, page 862, New York, New York, USA, 2009. ACM Press.

[5] Moritz Schubotz, Marcus Leich, and V Markl. Querying large Collections of Mathematical Publications-NTCIR10 Math Task. *ntcir-math.nii.ac.jp*, pages 667–674, 2013.

[6] Petr Sojka and Martin Líška. Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In James H. Davenport, William M. Farmer, Josef Urban, and Florian Rabe, editors, *Intelligent Computer Mathematics Lecture Notes in Computer Science*, volume 6824 of *Lecture Notes in Computer Science*, pages 228–243, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[7] Qun Zhang and Abdou Youssef. An Approach to Math-Similarity Search. *Intelligent Computer Mathematics, Lecture Notes in Artificial Intelligence 8543*, pages 404–418, 2014.

Table 3: Query data. This table first lists query IDs followed by the queries, where the `qvar` elements (universal variables see `https://trac.mathweb.org/MWS/wiki/MwsQuery`) are listed in red. The columns $v = 0 - 4$ represent the relevance ranking (from the non relevant to the most relevant). Columns F_1 through F_5 correspond to similarity-measure factors $1-5$. The number of Content Dictionary matches is F_1, the number of data type matches is F_2, the number of exact matches at any depth is F_3, the average coverage is F_4, and the number of formulae (as opposed to expressions) is F_5.

ID	query	$v=0$	$v=1$	$v=2$	$v=3$	$v=4$	F_1	F_2	F_3	F_4	F_5
1f1.0	$square(phi) = id$	447	207	784	209	11	12	13	1	0.19	441
1f1.1	$phi \neq id$	447	207	784	209	11	44	8	8	0.19	441
2f1.0	$ImP^{+}_{\gamma} = C^{+}_{\mu}(\gamma)$	399	219	19	0	14	0	0	0	0.07	106
3f1.0	$L\prime_{d-k} = L_k$	764	80	99	10	0	0	0	0	0.12	239
4f1.0	$B\sigma_3 B = \sigma_3$	1307	19	72	11	7	5	5	0	0.11	383
5f1.0	$S_{EH} = \frac{1}{G_3} od^3x\sqrt{-g^{(3)}}$	75	187	313	48	73	0	0	0	0.14	198
6f1.0	$S = -T_p \int d^{p+1}x\sqrt{g}$	936	249	118	90	205	0	0	0	0.36	445
7f1.0	$x\frac{y}{z} - u\frac{v}{w}$	847	232	12	42	0	0	0	0	1.00	246
8f1.0	$x \leq \frac{6}{2^n} + 12\epsilon$	864	119	14	24	0	0	0	0	0.15	225
9f1.0	$x^i_n = (1-\epsilon)f + \frac{\epsilon}{2}[g+h]$	1726	79	224	0	0	0	0	0	0.18	527
10f1.0	$f(x) = \frac{1}{\sigma\sqrt{2\pi}}z$	458	160	68	8	60	0	0	0	0.17	231
11f1.0	$p^2 + x^2(ix)^{\epsilon}$	649	466	9	25	84	11	11	0	0.22	369
12f1.0	L_∞	515	135	66	47	36	196	205	122	0.30	117
13f1.0	(D)	645	7	17	60	6	723	727	281	0.38	103
14f1.0	$-tr(xlnx)$	427	147	42	7	17	0	0	0	0.23	164
15f1.0	$\frac{1}{n^s}$	666	10	115	0	43	87	78	52	0.32	157
16f1.0	$f(x) = x$	909	66	4	72	9	22	76	0	1.00	322
17f1.0	$f(z) = z^d + c$	579	50	14	83	70	8	12	0	0.35	211
18f1.0	$\frac{az+b}{cz+d}$	531	30	0	44	290	34	34	33	0.27	285
19f1.0	$H_{n-k}(X)$	245	60	10	213	431	13	23	0	0.25	175
20f1.0	$x^2 - x - 1 = 0$	1562	11	47	44	123	6	6	5	0.29	502
21f1.0	$f(ax+by) < af(x) + bf(y)$	418	265	59	35	24	0	0	0	0.32	195
22f1.0	$\int_M fdS$	183	109	97	52	328	29	31	0	0.23	182
23f1.0	$\langle\cdot,\cdot\rangle$	114	108	174	239	112	141	141	0	0.04	92
24f1.0	$\widehat{\mathrm{CH}}^p(A) \cong Y$	324	48	11	80	236	1	1	0	0.16	72
25f1.0	$\widehat{\deg}(x_1^{k_1}x_2^{k_2}\cdots x_n^{k_n})$	1016	15	14	10	6	0	0	0	0.30	256
26f1.0	$det(a_1b_2 - a_2b_1 + c)$	738	193	12	20	0	0	0	0	0.28	249
27f1.0	$E(\lambda) = -m^2_{\mathrm{dyn}}(\lambda)$	484	496	19	12	99	3	3	0	0.10	323
28f1.0	Φ^4	426	21	8	181	130	207	347	0	0.09	146
29f1.0	$\sum_{n=0}^{\infty} t^m a_k(x)$	373	665	13	222	61	0	0	0	0.28	349
30f1.0	$\mathbb{C}P^n$	240	15	0	16	319	123	128	0	0.13	80

31f1.0	$k + 1/(3k + c)$	877	44	96	38	0	0	0	0	0.43	397
32f1.0	$\|u \cdot v\| \leq \|\|u\|\|\|\|v\|\|$	729	74	27	14	0	2	2	0	0.00	227
33f1.0	$\|\|fg\|\|_1 \leq \|\|f\|\|_p \|\|g\|\|_q$	878	34	25	11	42	2	4	0	0.22	259
34f1.0	$\lim_{n\to\infty} \int_X f_n du = \int_X \lim_{n\to\infty} f_n du$	833	57	69	51	0	0	0	0	0.24	242
35f1.0	$\|\|x - a\|\| \leq \frac{1}{\|\|a^{-1}\|\|}$	1089	82	27	9	6	0	0	0	0.38	328
36f1.0	$\rho(A) = \lim_{n\to\infty} \|\|A^n\|\|^{1/n}$	384	210	22	38	90	10	10	0	0.22	217
37f1.0	$A = USV^T$	521	249	85	60	151	10	10	0	0.30	293
38f1.0	$\|\|x + y\|\|_p \leq \|\|x\|\|_p + \|\|y\|\|_p$	260	337	70	169	33	7	6	1	1.00	264
39f1.0	$\mathbb{P}[X \geq t] \leq \frac{\mathbf{E}[X]}{t}$	196	303	52	14	37	1	1	0	0.05	132
40f1.0	$\lim_{n\to\infty} \mathbb{P}[\|A_n - \mathbf{E}[X]\| > e] = 0$	443	381	39	6	3	0	0	0	0.12	194
41f1.0	$\mathbb{P}[\lim_{n\to\infty} A_n = \mathbf{E}[X]] = 1$	103	255	158	3	17	0	0	0	0.19	123
42f1.0	$E = \bigoplus_{i=0}^{\infty} E_i$	197	152	340	42	22	14	14	0	0.22	156
43f1.0	$\oint_C \mathbf{B} \cdot \mathrm{d}\ell = \mu_0 I$	1000	1251	78	16	31	0	0	0	0.03	606
44f1.0	$x^n + y^n = z^n$	1250	16	3	58	24	5	7	0	1.00	292
44f1.1	$x, y, z, n \in \mathbb{N}$	1250	16	3	58	24	0	0	0	0.00	292
45f1.0	$\frac{1+\sqrt{5}}{2}^n$	565	103	40	5	77	9	10	9	0.30	198
46f1.0[1]	$1024k^10 - 2560k^9 + 3840k^8$ $-4480k^7 + 4096k^6 - 2944k^5$ $+1696k^4 - 760k^3 + 236k^2 - 40k$	890	19	23	19	2	0	0	0	0.16	220
47f1.0	$P_n = 2P_{n-1} + P_{n-2}$	948	142	10	18	82	7	7	3	0.39	383
48f1.0	$\dot{x}(t) = Ax(t) + Bu(t)$	78	759	45	0	25	3	3	0	0.22	162
48f1.1	$t \in \mathbb{R}$	78	759	45	0	25	63	63	0	0.15	162
48f1.2	$x(t) \in \mathbb{R}^n$	78	759	45	0	25	4	4	0	0.18	162
48f1.3	$u(t) \in \mathbb{R}^m$	78	759	45	0	25	4	4	0	0.16	162
49f1.0	$\sum_{n=1}^{2*k-1} (-1)^n * cos(1/4 * \pi) * n^{2/k} = R$	1074	155	9	0	0	0	0	0	0.25	348
50f1.0	$\chi'_a(G) \leq \Delta(G) + 6$	462	7	121	84	228	4	3	0	0.02	223

[1]We added line breaks to the query 46f1.0 to improve readability.