Stephen M. Watt
Mathematical Document Classification via Symbol Frequency Analysis

In: Petr Sojka (ed.): Towards Digital Mathematics Library. Birmingham, United Kingdom, July 27th, 2008. Masaryk University, Brno, 2008. pp. 29--40.

Persistent URL: http://dml.cz/dmlcz/702543

Mathematical Document Classification via Symbol Frequency Analysis

Stephen M. Watt

Ontario Research Centre for Computer Algebra
University of Western Ontario
London Ontario, Canada N6A 5B7
E-mail: Stephen.Watt@uwo.ca

Abstract. Earlier work has examined the frequency of symbol and expression use in mathematical documents for various purposes including mathematical handwriting recognition and forming the most natural output from computer algebra systems. This work has found, unsurprisingly, that the particulars of symbol and expression vary from area to area and, in particular, between different top-level subjects of the 2000 Mathematical Subject Classification. If the area of mathematics is known in advance, then an area-specific information can be used for the recognition or output problem. What is more interesting is that although the specifics of which symbols are ranked as most frequent vary from area to area, the shape of the relative frequency curve remains the same. The present work examines the inverse problem: Given the relative frequencies of symbols in a document, is it possible to classify the document and determine the most likely area of mathematics of the work? We examine the symbol frequency "fingerprints" for the different areas of the Mathematical Subject Classification.

Key words: mathematical document classification, mathematical handwriting recognition, statistical methods

1 Introduction

We consider the problem of how to classify mathematical text automatically by its subject area. This problem interests us for a number of reasons.

The first reason is to aid in retro-classification of existing literature: Even if all new documents are written with accurate subject metadata, there is a significant volume of important existing literature that does not have this metadata. Moreover, subject classifications change over time. In the future there will be new subject areas into which current papers will fit, but for which there are not yet classifications.

The second reason is to aid in document understanding. Although we are a long way from machine understanding of general mathematical documents, it is not too ambitious a goal to be able to assign likely interpretations to symbols in a mathematical document. For example, if we know that an article is in the

Petr Sojka (editor): DML 2008, Towards Digital Mathematics Library, pp. 29–40.
 ISBN 978-80-210-4658-0

area of semi-algebraic sets, then $H < G$ would most likely mean that the value of the real-valued variable H is less than the value of the real-valued variable G. If we know that the area is group theory, then the same expression would likely mean that H is a normal subgroup of G.

The third reason relates to pen-based interfaces for mathematical software. If we are able to determine the mathematical subject areas in use, then this can be used to weight handwriting recognition results. For example, the exact same ink trace might "obviously" be interpreted as one symbol in one area of mathematics (say as an "$\imath$" under a summation sign) or as another in a different area (say as "$\dot{z}$" in a differential equation).

Finally, if a human reader can accurately judge the area of a mathematical document with a five second flip through several hundred pages, then that shows there are some macroscopic document properties that we ought to be able to recognize by machine.

We propose that symbol frequency information provides useful information for the classification of mathematical text. In this paper we show that mathematical articles from the `arXiv` [1] preprint service give well separated symbol frequency measures when categorized according to the their top-level subjects in the 2000 Mathematical Subject Classification, MSC 2000. (The Mathematical Subject Classification is used to categorize articles reviewed by Mathematical Reviews and Zentralblatt MATH. See, *e.g.*, [2] for a description.)

In our work we have found it useful to group symbols into two categories: "identifiers", which are letters (typically Latin or Greek) standing for variables, parameters, constants, functions, *etc.*, and "operators", which are other mathematical symbols, or certain letters used in special contexts, such as "$\sum$".

We use the term "document" to describe a piece of mathematical text to be considered. For example, it may be an article, a book, a book chapter, handwritten input, or equations in a computer algebra system worksheet. Each application will have its own particular properties, but we would expect the ideas we discuss here to be useful across such a range of areas.

The outline of the paper is as follows: In Section 2 we describe the bodies of mathematical text we have analyzed and how we computed symbol and n-gram frequencies in the mathematical expressions. In Section 3 we show how the frequency information varies by mathematical area. In Section 4 we explore the idea of distinguishing mathematical area by examining symbol frequency in particular documents. Finally, in Section 5 we conclude the article.

2 Computing Symbol and n-gram Frequencies

In earlier work we have examined particular bodies of text to produce empirical measures on the sets of mathematical symbols and n-grams (n symbol sequences) that occur in the mathematical expressions in two sorts of material. Similar studies have been performed by Garain *et al.* [3] and Uchida *et al.* [4].

arXiv ids		arXiv operators		Eng. text ids		Eng. text operators	
Symbol	Freq.	Symbol	Freq.	Symbol	Freq.	Symbol	Freq.
n	48,150	$=$	128,715	x	49,740	$=$	58,988
i	43,280	$-$	116,064	y	29,481	$($	50,843
x	36,240	$,$	112,818	n	21,152	$)$	50,838
k	32,060	$@$	103,090	z	18,859	$-$	38,243
t	25,967	$+$	79,404	t	17,100	$+$	31,297
X	23,369	$\ni$	43,942	f	13,092	$,$	25,350
j	23,038	$*$	29,210	a	12,119	$\square$	17,305
p	22,832	$\rightarrow$	23,818	i	9,179	$.$	16,213
A	22,791	$/$	23,405	u	9,147	$'$	12,401
a	21,435	$\leqslant$	20,088	c	8,985	$\vert$	8,176
d	19,457	$\tilde{}$	16,875	s	8,784	$/$	7,508
m	19,263	$\bigotimes$	14,242	d	8,457	$]$	7,012
f	18,235	$\sum$	13,560	e	8,451	$[$	7,010
M	18,135	$>$	13,528	π	7,664	$\ldots$	4,396
s	17,659	∞	13,138	k	7,194	∂	4,105
r	17,248	$\bar{}$	12,451	m	6,437	$<$	3,922
C	16,915	$<$	12,058	r	5,561	$\leqslant$	3,808
S	16,487	$\ldots$	12,005	b	5,447	$\int$	3,732
G	16,074	∂	11,940	v	4,537	∞	3,490
α	15,943	$\times$	11,294	j	4,491	$\sum$	3,743

Fig. 1. Most frequent identifiers and operators in `arXiv` articles and in 2nd year engineering texts. `arXiv` frequencies for identifiers (operators) are per million identifiers (operators). Engineering text frequencies are per million total symbols (identifiers, operators, digits combined). Parentheses were not counted in the `arXiv` analysis. The operator "@" stands for the MathML invisible "apply function" operator, which is inserted by the conversion process.

The arXiv study. The first study [5,6] examined approximately 20,000 preprints from the mathematical `arXiv` collection from the years 2000 to 2005. This was a near-complete collection for that time period of those articles for which source TEX was available and were classified according to the MSC 2000 subject classification.

TEX documents typically rely heavily on both system- and author-defined macros so it is necessary to expand macros to find the symbols actually appear in a document. To perform this macro expansion we used our TEX to MathML converter [7], and then took the sequence of leaf symbols from the resulting MathML trees.

The leaf symbols were tabulated separately for identifiers and mathematical operators. The frequency with which each subexpression occurred was also analyzed. The most frequently seen identifiers and operators are shown in Figure 1.

#		Subject Classification	#		Subject Classification
19	00	General	34	45	Integral equations
39	01	History and biography	1066	46	Functional analysis
228	03	Math. logic and foundations	543	47	Operator theory
1212	05	Combinatorics	164	49	Calculus of var.; optimization
164	06	Order, lattices, ordered alg. struct.	171	51	Geometry
48	08	General algebraic systems	435	52	Convex and discrete geometry
1383	11	Number theory	1717	53	Differential geometry
108	12	Field theory and polynomials	226	54	General topology
667	13	Commutative rings and algebras	627	55	Algebraic topology
2445	14	Algebraic geometry	1618	57	Manifolds and cell complexes
240	15	Linear and multilin. alg.; matrix thy	920	58	Global analysis, an. on manifolds
861	16	Associative rings and algebras	877	60	Prob. theory and stoch. processes
760	17	Nonassociative rings and algebras	105	62	Statistics
404	18	Category theory; hom. algebra	209	65	Numerical analysis
239	19	K-theory	237	68	Computer science
1169	20	Group theory and generalizations	113	70	Mechanics of particles and systems
472	22	Topological groups, Lie groups	34	74	Mechanics of deformable solids
185	26	Real functions	69	76	Fluid mechanics
123	28	Measure and integration	13	78	Optics, electromagnetic theory
308	30	Functions of a complex variable	6	80	Classical thermodyn., heat xfer
59	31	Potential theory	553	81	Quantum theory
797	32	Several complex var. & anal. spaces	260	82	Stat. mechanics, struct. of matter
312	33	Special functions	48	83	Relativity and gravitational theory
295	34	Ordinary differential equations	6	85	Astronomy and astrophysics
746	35	Partial differential equations	15	86	Geophysics
706	37	Dyn. systems and ergodic theory	96	90	Operations research, math. prog.
52	39	Difference and functional eqns	42	91	Game thy, econ., soc. & behav. sci.
21	40	Sequences, series, summability	35	92	Biology and other natural sciences
88	41	Approximations and expansions	115	93	Systems theory; control
290	42	Fourier analysis	128	94	Info. and comm., circuits
143	43	Abstract harmonic analysis	12	97	Mathematics education
43	44	Integral transforms, op. calculus			

Fig. 2. Count of `arXiv` articles by MR subject classification

The engineering text study. In the second study [8], we examined the symbols and n-grams that occur in the most popular second year university engineering mathematics texts used in North America. The most popular texts (by sales) were by Kreyszig [9,10] (72%), Greenberg [11] (13%) and O'Neil [12] (7%), together making up more than 90% of the second year engineering textbook use.

We obtained the TEX sources for the textbooks of Greenberg and O'Neil from the author and publisher respectively. For textbook of Kreyszig, we scanned all the pages of the book, used the Infty [13] document analysis program to generate TEX to the degree that it could, and then hand-corrected the TEX. In each case, the TEX was converted to MathML for analysis as in the `arXiv` study. In principle, MathML could have been generated by Infty from the scans of the Kreyszig pages, but it was easier to correct the generated TEX than MathML. To obtain overall statistics, the symbol and n-gram frequencies from the three texts were combined using the textbook adoption rate as weights. The most frequently seen identifiers and operators from this study are shown in Figure 1 on the previous page. This data has been used to build predictive mathematical character recognizers [14].

Subject	Chapters
Ordinary Differential Equations	Kreyszig 1–6, Greenberg 1–7, O'Neil 1–5 & 10–11
Linear Algebra	Kreyszig 7–8, Greenberg 8–11 & 14, O'Neil 6–9
Vector Calculus	Kreyszig 9–10, Greenberg 16, O'Neil 12–13
Partial Differential Equations	Kreyszig 12, Greenberg 18–20, O'Neil 17–19
Fourier Analysis	Kreyszig 11, Greenberg 17, O'Neil 14–16
Multivariable Calculus	Greenberg 13&15
Complex Analysis	Kreyszig 13–18, Greenberg 12&21–24, O'Neil 20–25
Numerical Analysis	Kreyszig 19–21
Linear Programming	Kreyszig 22
Graph Theory	Kreyszig 23
Probability and Statistics	Kreyszig 24–25, O'Neil 26–27

Fig. 3. Engineering text chapter subject groupings

3 Frequencies by Area

As well as analyzing the most frequently occurring symbols, n-grams and expressions for the entire corpus considered, the earlier studies [5,6,8] also analyzed frequency by mathematical area.

In the `arXiv` study, the symbol frequencies were calculated separately for each top-level subject classification. The number of articles in each top-level subject area for the sample is shown in Figure 2 on the facing page. For the engineering text study, a number of subject areas were identified, as shown in Figure 3, and the relevant chapters were considered together.

In both studies, each subject had its own distinct set of most popular symbols. To illustrate, the most frequent identifiers in three subject classifications in the `arXiv` study are shown in Figure 4 on the following page.

It can be seen that the pattern of relative frequencies for the most popular symbols is similar even though *which* symbols are the most popular is different. In each area the most frequently used identifier occurred about 50,000 times per million and the twentieth most frequently used identifier occurred about 16,000 times per million. These frequencies are graphed in Figure 5 on page 35. Very similar curves are observed in each subject classification, and in all subjects combined.

The same phenomenon is observed in the relative frequency of symbols and n-grams occurring in the engineering mathematics subject areas, as shown in Figure 6 on page 36. In Figure 6(a) the frequency of each symbol from most popular to least popular is shown. Each curve corresponds to one of the subjects listed in Figure 3 and orders the symbols differently (from its own most frequent to its own least frequent) on the X axis. Figure 6(b) shows the same information, but this time with cumulative frequencies, and Figure 6(c) shows the same information as the first but with a logarithmic scale. Figure 6(d) shows the cumulative frequencies of bigrams by author. In this case the horizontal axis enumerates the bigrams, from most frequent to least frequent for each of the curves.

03	
Id	Freq
i	51,565
n	48,239
x	41,042
X	33,862
A	29,845
p	26,292
α	24,604
k	24,374
f	22,671
a	22,030
G	21,983
m	19,893
j	18,062
ω	18,015
M	17,256
S	17,122
C	17,107
F	16,773
y	16,764
t	15,693

11	
Id	Freq
n	58,186
p	40,302
k	38,230
x	35,294
i	35,100
a	25,301
m	23,642
d	22,302
q	21,797
s	21,319
j	21,153
r	19,695
t	19,654
G	19,620
X	19,535
A	19,107
K	18,905
f	18,126
F	16,524
L	15,921

35	
Id	Freq
x	51,773
t	49,859
u	39,841
n	35,705
k	29,924
i	28,941
s	25,234
j	24,968
d	24,095
L	21,094
ϵ	20,740
λ	20,189
p	19,107
C	17,450
α	17,087
r	16,834
ν	16,820
a	15,931
y	15,920
f	15,215

All	
Id	Freq
n	48,150
i	43,280
x	36,240
k	32,060
t	25,967
X	23,369
j	23,038
p	22,832
A	22,791
a	21,435
d	19,457
m	19,263
f	18,235
M	18,135
s	17,659
r	17,248
C	16,915
S	16,487
G	16,074
α	15,943

Fig. 4. The most frequent identifiers (per million) in Logic (03), Number Theory (11) and PDEs (35). The most frequent for all areas combined is shown for comparison.

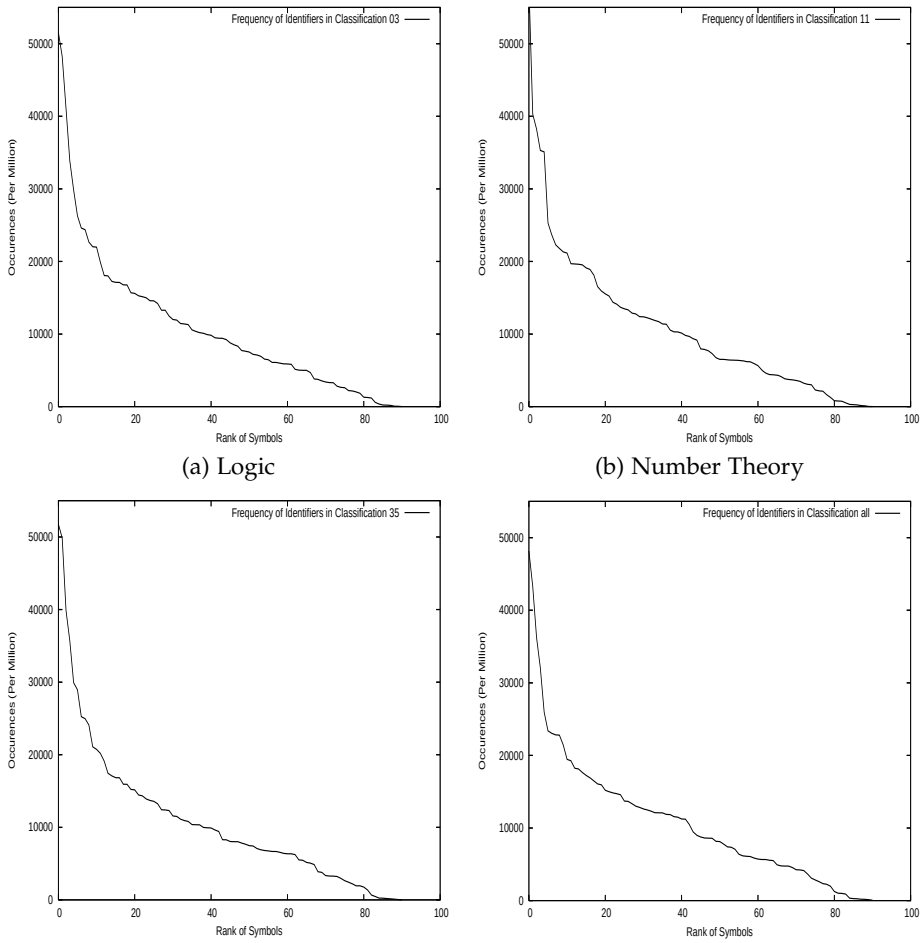
We have seen in each case, for the `arXiv` study and the engineering text study, for subjects separately or combined, for mathematical symbols or n-grams, for identifiers or for operators, that the frequencies roughly follow a Zipf distribution [15].

4 Areas by Frequency

As our empirical analyses have shown that the symbols occurring in mathematical expressions roughly obey Zipf exponential distributions, we expect the frequencies of symbols typically to be well separated. There are certain symbols that usually occur in pairs, such as parentheses, and in this case only one of the symbols need be recorded, keeping the symbol frequencies separated.

With well separated frequencies, and area-dependent rankings of the most common symbols, we expect that frequency ranking of the symbols that occur in a document will give a useful subject characterization.

Using the data collected in earlier work [5,6], we have been able to compare the frequency ranking of symbols according to subject area. Figures 7 on page 37



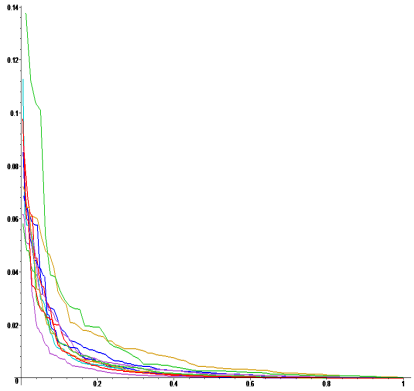
(a) Logic (b) Number Theory

(c) Partial Differential Equations (d) All areas combined

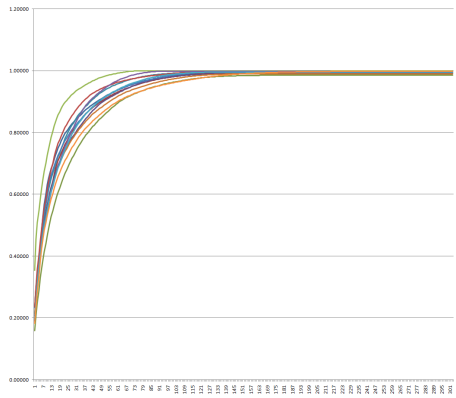
Fig. 5. `arXiv` frequencies The most frequent identifiers in representative areas. The horizontal axis gives the symbol, from most frequent to least frequent, and the vertical axis gives the number of occurrences. The symbol order is different in each case.

and 8 on page 38 show the ranking of the most frequently used identifiers and operators, respectively, for each top-level MSC 2000 subject classification.

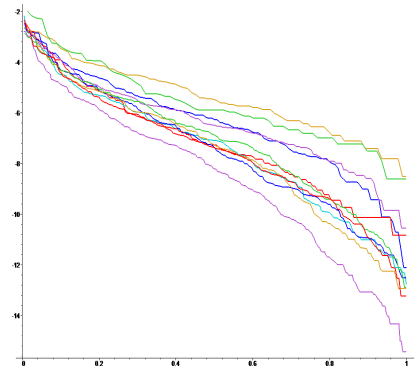
For each top-level classification, the twenty most frequently occurring identifiers and operators are listed, ordered by frequency. For example, in subject 35 (PDEs) x is the most frequently occurring identifier, t the next most, and so on. In each row of the tables a "¶" marker separates the part of the symbol ranking that is common to more than one classification and the part that is unique to the classification. For example, having x most frequent and t



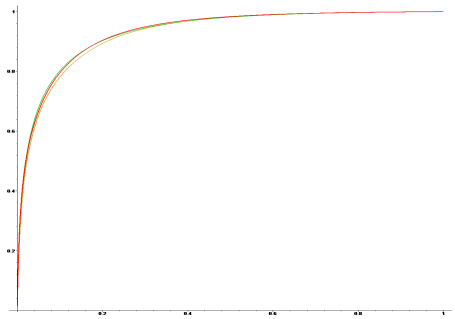
(a) Symbol frequency by area



(b) Cumulative symbol freq. by area



(c) Log symbol frequency by area



(d) Cumulative bigram freq. by author

Fig. 6. Engineering text frequencies For (a), (b), (c), each curve is for a subject area. For (d), each curve is for an author. The horizontal axis gives the symbols/bigrams from most frequent to least frequent, independently for each curve.

second most frequent is common to classifications 34, 35 and 49, but no others. Each of these three subject areas is seen to have a unique third most common identifier: for ODEs it is n; for PDEs it is u; for the calculus of variations it is i.

Examining the rankings for identifiers and for operators we see that the frequencies of the most common identifiers tend to be more area-specific than the operator frequencies. We therefore concentrate on the identifier frequencies. In some cases, however, the frequent use of a particular operator will indicate a certain subject, *e.g.* ∂ or ∇.

Comparing all subject areas we see that identifying the top six most frequently used identifiers will uniquely identify a top-level subject classification. In many cases identifying the top two or three most frequently used identifiers will correctly give the subject area.

00	n	x	k	¶	m	i	α	d	P	s	e	y	W	B	L	θ	A	p	S	z	w
01	x	n	¶	t	y	a	f	k	λ	p	i	C	b	w	g	m	A	z	c	r	M
03	i	n	x	¶	X	A	p	α	k	f	a	G	m	j	ω	M	S	C	F	y	t
05	n	i	k	¶	x	j	a	m	q	t	r	p	A	d	G	s	b	λ	S	ν	P
06	n	x	i	¶	a	k	y	p	A	S	b	P	j	L	f	ν	u	F	t	c	m
08	n	¶	A	a	i	x	f	X	G	k	V	b	B	W	U	y	M	α	m	g	p
11	n	¶	p	k	x	i	a	m	d	q	s	j	r	t	G	X	A	K	f	F	L
12	n	x	i	¶	K	k	p	a	A	f	G	F	X	r	j	M	t	H	m	q	y
13	i	n	¶	R	x	A	I	k	m	d	M	a	t	X	j	p	f	S	r	C	H
14	i	¶	X	n	k	x	p	d	m	C	A	S	j	r	t	a	H	G	f	E	F
15	n	i	A	¶	k	x	j	a	r	m	B	N	t	p	d	C	S	X	s	P	f
16	i	¶	A	n	x	k	a	M	H	R	C	j	X	p	t	S	m	B	α	f	r
17	i	n	k	a	¶	j	x	q	A	m	α	r	p	λ	g	e	s	V	t	L	b
18	¶	A	i	n	X	C	M	x	f	G	k	F	p	S	a	j	R	B	m	H	d
19	n	i	A	¶	G	X	K	k	p	C	H	M	x	t	a	S	B	E	f	F	R
20	i	n	¶	G	x	k	a	p	A	X	j	t	s	m	S	q	H	r	g	α	ν
22	G	n	¶	i	x	X	g	k	A	a	p	λ	s	H	M	t	α	f	L	d	C
26	x	n	¶	i	f	a	t	k	p	m	j	b	y	d	α	s	z	A	λ	g	X
28	n	x	k	i	¶	t	j	X	A	μ	d	f	G	m	r	a	T	λ	p	B	e
30	n	¶	z	x	k	i	f	t	j	m	a	d	r	g	S	A	p	y	α	P	w
31	x	n	¶	z	d	k	r	u	j	t	i	Ω	D	p	y	h	f	α	A	C	K
32	n	i	k	¶	X	x	z	j	f	p	t	M	d	m	A	C	a	S	r	H	D
33	n	¶	q	k	x	i	a	j	z	m	t	b	s	r	p	λ	d	y	c	α	A
34	x	t	¶	n	k	i	a	y	j	z	d	u	s	λ	m	α	p	f	A	e	q
35	x	t	¶	u	n	k	i	s	j	d	L	e	λ	p	C	α	r	ν	a	y	f
37	n	x	i	k	t	¶	f	d	p	z	j	X	s	T	A	m	u	a	α	M	C
39	n	x	¶	q	z	i	k	t	m	j	a	A	p	f	λ	X	α	d	P	y	μ
40	k	n	¶	z	x	d	a	e	b	t	m	j	i	f	q	r	w	s	u	α	p
41	n	x	¶	a	i	k	z	t	j	f	b	m	s	p	d	A	N	r	c	P	L
42	n	x	k	¶	j	f	i	d	t	p	m	L	z	N	ξ	C	α	T	a	y	A
43	G	n	¶	x	f	i	k	X	L	A	p	d	t	π	z	S	j	a	m	g	α
44	x	n	k	¶	f	d	m	t	α	y	z	i	g	a	j	p	ξ	w	q	b	s
45	t	¶	n	k	x	d	L	p	s	A	e	R	u	C	m	f	α	T	z	N	λ
46	n	i	x	¶	A	k	X	t	p	a	j	f	C	G	M	s	E	B	m	α	d
47	n	x	k	i	¶	A	t	j	f	T	p	a	X	H	z	C	d	α	λ	L	m
49	x	t	¶	i	u	n	k	p	d	y	s	T	ν	Ω	h	j	K	M	f	L	α
51	i	n	x	¶	k	j	p	a	d	z	r	t	S	A	ν	P	X	y	m	C	q
52	n	i	x	k	¶	d	A	j	P	p	m	S	a	ν	t	X	r	F	C	f	M
53	i	n	M	¶	x	k	t	X	g	d	p	j	A	α	S	a	s	T	H	f	r
54	X	¶	x	n	f	i	k	G	A	y	p	C	g	S	U	α	F	j	K	L	a
55	n	¶	X	i	A	k	p	G	S	C	H	M	x	f	B	m	j	a	F	t	P
57	i	n	M	¶	k	S	x	X	p	t	H	A	C	G	K	j	a	T	f	m	L
58	n	x	i	k	t	¶	M	d	X	j	p	A	g	f	s	λ	a	u	C	α	L
60	n	¶	t	x	i	k	d	s	j	X	N	p	y	m	u	T	A	f	α	a	r
62	n	x	t	¶	k	i	j	f	X	m	d	α	h	s	y	θ	N	P	p	u	A
65	n	x	t	¶	i	k	j	u	a	f	s	d	h	m	y	p	z	α	N	r	A
68	n	i	x	k	¶	X	t	p	j	a	α	m	r	d	f	A	P	S	s	z	c
70	i	¶	k	n	x	t	q	j	s	d	A	p	m	L	M	H	X	a	T	r	ν
74	t	¶	u	n	x	Ω	i	k	e	K	s	ν	d	h	a	H	δ	g	L	T	j
76	t	x	¶	u	s	T	n	k	d	L	e	p	i	ω	f	α	y	C	r	τ	w
78	k	n	¶	x	l	t	i	j	d	ξ	N	r	m	σ	u	α	f	p	β	s	π
80	k	¶	d	t	x	q	j	τ	X	p	n	a	α	c	m	ν	M	b	s	u	ν
81	i	n	k	a	¶	x	t	j	A	q	p	m	s	z	d	α	λ	L	H	b	f
82	n	x	i	¶	t	k	N	d	j	p	m	a	s	y	z	b	r	q	c	u	L
83	i	x	¶	M	t	k	j	a	n	g	r	d	α	p	X	ν	u	L	e	E	s
85	t	¶	N	i	x	s	M	p	r	d	f	k	y	ν	W	n	e	S	R	Y	B
86	t	¶	e	x	T	y	s	u	θ	ω	d	k	L	D	R	τ	S	H	μ	h	δ
90	i	x	¶	n	k	j	t	m	A	f	d	P	c	ν	p	r	y	s	u	z	a
91	t	x	¶	n	i	T	k	d	s	q	X	j	p	a	f	y	u	G	R	h	θ
92	k	n	¶	i	t	r	j	x	P	a	N	s	A	S	X	I	θ	T	α	d	y
93	t	x	¶	i	n	s	j	k	T	α	d	a	m	p	A	X	z	u	f	y	V
94	n	x	i	k	¶	f	p	m	t	d	j	a	z	A	X	y	N	C	e	P	g
97	x	n	¶	θ	m	A	d	B	y	N	L	p	S	k	C	e	π	a	j	f	t

Fig. 7. Most frequent identifiers in descending order, by subject.
The ranking preceding the ¶ in each row is shared with at least one other row. The frequency rankings following the ¶ are unique to their rows.

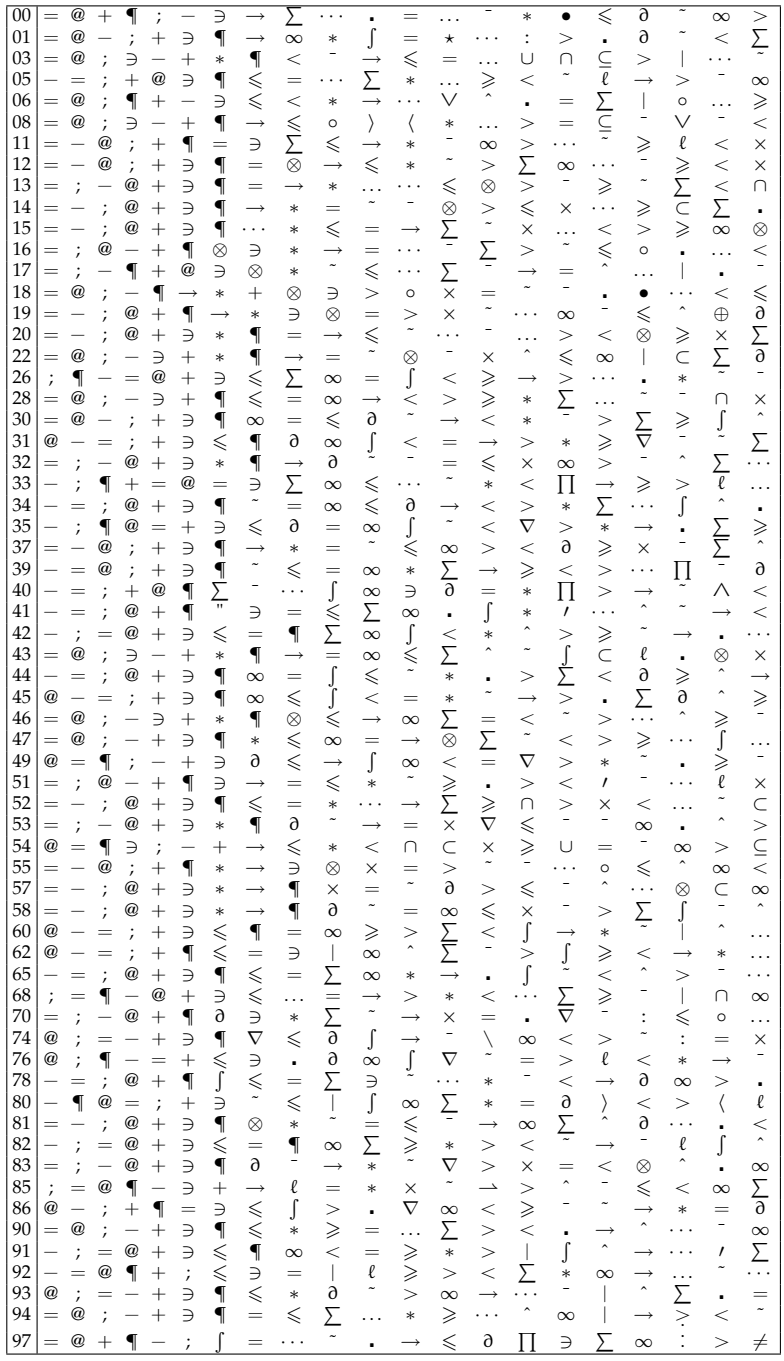


Fig. 8. Most frequent operators in descending order, by subject.
The ranking preceding the ¶ in each row is shared with at least one other row. The frequency rankings following the ¶ are uniqe to their rows. The symbol "@" stands for the invisible MathML "ApplyFunction" operator.

5 Conclusions and Future Work

We have seen that the frequencies of the symbols and of the n-grams occurring in practice in mathematical expressions are close to Zipf distributions.

In the cases we have examined, the frequency distribution holds when subject subsets of a document corpus are analyzed. Although the frequency distribution remains similar, the ranking of the most frequent symbols frequent changes dramatically according to the subject. We have observed this both in arXiv articles classified by MSC 2000 subject area and second year engineering mathematics textbooks with chapters grouped by related topics.

We have seen that in all MSC 2000 subject areas the most frequently occurring identifiers are Latin or Greek letters, and that the different subject areas have quite distinct usage patterns. Indeed, the frequency raking of the most commonly used few (2–6) identifiers appears to be give a different for each subject area. This contrasts with the frequency ranking of operator symbols, which does not vary as much by subject.

We propose using the symbol-frequency ranking for fast automatic pre-classification of mathematical documents. This would allow more specialized methods to then verify or refine the classification. Determining subject area by symbol-frequency ranking can also aid in document recognition, where identifying the subject area can allow area-specific information to be used for disambiguation.

There are a number of interesting questions for future investigation. It would be useful to analyze the typical variance of documents within subject areas and to test the robustness of these symbol frequency measures. It remains an open question as to which classification strategy (Bayesian model, support vector machine, k nearest neighbors, *etc*) works best in this application. In natural language, given a specific set of distributions for word frequencies, it is possible to find an optimal classification scheme. It remains an open question to what degree does this remain practical for symbols in mathematical equations. Finally, the Mathematical Subject Classification and the body of mathematical literature are both moving targets. It would be useful to understand how stable the mathematical symbol frequencies are over decades in the literature and the degree to which they differ in the previous (1991) and subsequent (2010) Mathematical Subject Classifications.

References

1. arXiv e-Print archive, http://arxiv.org.
2. 2000 Mathematics Subject Classification. American Mathematical Society, http://www.ams.org/msc.
3. Garain, U., Chaudhuri, B. B.: A corpus for OCR research on mathematical expressions, International Journal on Document Analysis and Recognition, Vol. 7, Issue 4, pp. 241–259. (September 2005).
4. Uchida, S., Nomura, A., Suzuki, M.: Quantitative analysis of mathematical documents, International Journal on Document Analysis and Recognition, Vol. 7, Issue 4, pp. 211–218. (September 2005).

5. Clare M. So, Watt, S. M.: Determining Empirical Properties of Mathematical Expression Use, Proc. Fourth International Conference on Mathematical Knowledge Management, (MKM 2005), July 15–17, 2005, Bremen Germany, Springer Verlag LNCS 3863, pp. 361–375.
6. Clare M. So: *An Analysis of Mathematical Expressions Used in Practice*, Masters Thesis, University of Western Ontario, 2005.
7. Watt, S. M.: Exploiting Implicit Mathematical Semantics in Conversion between TEX and MathML, Proc. Internet Accessible Mathematical Communication, `http://www.symbolicnet.org/conferences/iamc02`, July 7, 2002, Lille, France.
8. Watt, S. M.: An Empirical Measure on the Set of Symbols Occurring in Engineering Mathematics Texts, Proc. 8th IAPR International Workshop on Document Analysis Systems, (DAS 2008), Sept 17–19, 2008, Nara, Japan, (IEEE, to appear).
9. Kreyszig E.: *Advanced Engineering Mathematics, 8^{th} ed.*, Wiley & Sons 1999.
10. Kreyszig E.: *Advanced Engineering Mathematics, 9^{th} ed.*, Wiley & Sons 2006.
11. Greenberg M.: *Advanced Engineering Mathematics, 2^{nd} ed.*, Prentice Hall 1998.
12. O'Neil P.: *Advanced Engineering Mathematics, 5^{th} ed.*, Thomson-Nelson 2003.
13. Suzuki, M., Tamari, F., Fukuda, R., Uchida, S., Kanahori, T.: Infty—an integrated OCR system for mathematical documents, Proceedings of ACM Symposium on Document Engineering 2003, Grenoble, 2003, pp. 95–104.
14. Smirnova, E., Watt, S. M.: Context-Sensitive Mathematical Character Recognition, Proc. International Conference on Frontiers in Handwriting Recognition, (ICFHR 2008), August 19–21, 2008, Montreal, Canada, (IEEE, to appear).
15. Zipf, G. K.: *Human Behavior and the Principle of Least-Effort*, Addison-Wesley, 1949.