Outline current state of formula recognition
(informal preliminary report)

by

Michiel Hazewinkel
CWI
POBox 94079
1090GB Amsterdam
The Netherlands

Abstract.

MSCS:

Key words and key phrases:

1. Description of the current state of affairs.

Mathematical formulas can carry quite a good deal of information. Indeed, it has been said that a professional mathematician, unacquainted with Japanese, can tell what a mathematical monograph in Japanese is about just by looking at the formulas. Also whether it is a book worthy to be translated or not.

Thus there is substantial interest in being able to recognize mathematical (and physical and chemical, ...) formulas automatically.

There would appear to be four stages involved as follows:

(i) Separate formulas, including in-line formulas, from the other parts of a text. This used to be a relatively well known problem in (optical) document analysis and has been researched adequately in that context, see e.g. [13].

Nowadays a substantial and increasing portion of science material on the web is in the form of .ps (Postscript) and .pdf (Portable Data Format) files. Separating out the forrmula parts from the text parts in these types of documents is an entirely different matter, especially in the .pdf (and .html or .xml) case. A start towards solving this problem (and there is no doubt that it can be solved efficiently) is in [45].

(ii) Once the formula parts have been recognised, it is needed to do some form of OCR (Optical Character Recognition) to see what glyphs are present; also their relative position, coded in some way, because mathematical formulas are two dimensional animals. In the case of .ps or .pdf, or .html files this is simple. In the case of scanned images this is a far from trivial matter and it has been addressed in numerous publications. Even to the point of recognising handwritten formulas.

(iii) Syntactical analysis. Once the 'glyph information' alluded to in (ii) above is available, plus relative position, the resulting two-dimensional image needs to be parsed. That is a representation must be found in terms of some mathematical formula display language like TeX of Math ML (or any other two-dimensional formula grammar).

This problem has also been well researched and there are a number of ways of dealing with it. For an up t odate survey see [6]. Sometimes step (ii) and (iii) are combined in various implemented systems.

(iv) Semantic analysis. The final step is to recognise the meaning of the various two dimensional collections of formula glyphs have. Presentation languages like TeX do not tell one what a formula is supposed to represent. The representation is ambiguous (both ways, because there are usually several ways to encode the pattern and they may very well not be faithful to the underlying mathematics. For instance, depending on the TeX encoding, the meaning of a formula can be effectively obscured. For instance, for display purposes one can elect to code presups to a symbol or treat them as postsubs to a previous symbol. Such things are perfectly permissible in display languages and in postscript for instance, there can be no hint as to what has been done.)

Thus it remains to attach some semantics to whatever has been recognised. The problem arises even at the level of a single glyph. Suppose π (Greek letter) has been recognised. To a group theorist this would probably mean a permutation, to a topologist it would probably mean a projection and to a geometer or real analist it would probably mean the number 3.14... . The surrounding context could probably settle it.

Thus the final step is to use some context analysis to determine the semantics of a formula.

The idea is to use a concept called "identification clouds" which has come up before in the context of automatic classification and automatic key phrase assignment, [20, 21]. This concept is described briefly in the next section.

2. Re identification clouds.

The rough idea of an identification cloud (for a key pharse, or concept or formula) is the following.

It may well happen that a very good key phrase for a given article is simply not present or so lingusitically mangled that even the best NLP techniques (Natural Language Processing) are not good enough to recognize it. Yet professionals in the field in question will have no difficulty in attributing that key phrase to that article. They do that by context.

The basic idea is that a good key phrase can be recognised by (part of) the collection of characteristic words (and short phrases) that one would normally expect to find in its neighbourhood. Thus the idenfification cloud of a (standardized) key phrase is a collection of words and short key phrases that belong to it and can be expected to show up in the immediate surrounding text. The same idea applies to formulas. If, for instance, the surrounding text is all about projective systems and projections and/or inverse ssytems then a 'π' in that neeighbourhood is likely to be a projection type mapping; if it is all about groups and representation and compositions the same 'π' is likely to be a permutation (or possibly a composition).

A well known quote is

> "Tell me what company thou keepest, and I will tell thee what thou art"
> Miguel de Cervantes (1547-1616),
> Don Quixote, Pt II, Ch. 23

The same applies to formulas and (missing) key phrases.

3. Background.

In the following list there is a collection of references that gives in my opnion an adequate idea of the current state of the art as regards formula recognition. As stated, stages 1,2,3 of the proces seem to be adequately researched; stage 4 still requires work, and, ideed, seems to have been completey neglected so far.

[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45]1999 #36; Kosmala, 2000 #13; Kosmala, 2000 #17; Kosmala, 2000 #34; Krenn, 1997 #3; Lavirotte, 1997 #15; Lavirotte, 1997 #37; Lavirotte, 1998 #32; Lawrence, 1998 #26; Lawrence, 1999 #27; Lee, 1997 #8; Lenart, 1992

#24; Pottier, 1994 #25; Saarland University, 1995 #18; Silagadze, 1997 #6; Smeaton, 1992 #28; Suzuki, 2000 #12; Yang, 2003 #38]

References

1. Steven Abney, *Statistical methods and linguistics* . In: Judith Klavans and Philip Resnik (ed.), The balancing act, MIT Press, 1996,

2. R H Anderson, *Syntax-directed recognition of hand-printed two-dimensional mathematics* , Dept. Engineering and Applied Physics. 1968.

3. R H Anderson, *Syntax-directed recognition of hand-printed two-dimensional mathematics* . In: Interactive systems for experimental aplied mathematics, Academic Press, 1968,

4. A Belaid, J P Haton, *A syntactic approach for handwritten mathematical formula recognition*, IEEE Trans. Pattern Analysis and Machine Intelligence **6**(1984), 105-111.

5. Kam-Fail Chan, Dit-Yan Yeung, *An efficient syntactic aproach to structural analysis of on-line handwritten mathematical expressions* , Pattern Recognition **33**:3(2000), 375-384.

6. Kam-Fail Chan, Dit-Yan Yeung, *Mathematical expression recognition: a survey* , International Journal on Document Analysis and Recognition **3**:1(2000), 3-15.

7. Kam-Fail Chan, Dit-Yan Yeung, *Error detection, error correction and performance evaluation in on-line mathematical expression recognition* , Pattern Recognition **34**:8(2001), 1671-1684.

8. S-K Chang, *A method for the structural analysis of two-diemnsional mathematical expressions* , Information sciences **2**(1970), 253-272.

9. E Cherniak, *Statistical techniques for natural language parsing* , Preprint, Brown university, 1997.

10. Ian Crowlesmith, *Creating a treasure trove of words* , Elsevier Science World. 14-15, 1993.

11. Ian Crowlesmith, *The development of a biomedical thesaurus* , NBBI Thesaurus Seminar. 1993.

12. D A Evans, *Preface to G Grefenstette, Explorations in automatic thesaurus discovery, KAP, 1994*. In: 1994,

13. Richard J Fateman, *How to find mathematics on a scanned page* . In: Daniel P Lopresti and Jiangying Zhou (ed.), Proc. SPIE Vol. 3967: Document recognition and retrieval VII, SPIE, 1999, 89-109.

14. Richard J Fateman, Taku Tokuyasu, *Progress in recognizing typeset mathematics* . In: Proc. SPIE Vol. 2660, Document Recognition III, San Jose, 1996, 1996, 37-50.

15. R J Fateman, T Tokuyasu, B P Berman, N Mitchell, *Optical character recognition and parsing of typeset mathematics* , Journal of Visual Communication and Image Representation **7**:1(1996), 2-15.

16. J T Favata, *General word recognition using aproximate segment-string matching* . In: International conference on Document Analysis and Recognition, Ulm, 1997, IEEE, 1997, 92-96.

17. U Garain, B B Chaudhuri, *A syntactic approach for processing mathematical expressions in printed documents* . In: Proc. International Conf. on Pattern Recognition 2000, IEEE, 2000, 4051-4651.

18. Gregory Grefenstette, *Explorations in automatic thesaurus discovery* , KAP, 1994.

19. J Hartmann, G Hotz, R Loos, R Marzinkewitsch, J Quapp, F Weigel, A Weber, *The "optical formula recognition" system for handprinted input* , J Symbolic Computation **11**(1995), 1-8.

20. Michiel Hazewinkel, *Key words and key phrases in scientific databases. Aspects of guaranteeing output quality for databases of information* . In: Proceedings of the ISI conference on Statistical Publishing, Warsaw, Uugust 1999, ISI, 1999, 44-48.

21. Michiel Hazewinkel, *Topologies and metrics on information spaces* , CWI Quarterly **12**:2(1999), 93-110. Preliminary version: http://www.mathematik.uni-osnabrueck.de/projects/workshop97/proc.html

22. G Hotz, Joachim Quapp, Frank Marvin Weigel, *Optical formula recognition*, Inst. Informatik, Univ. des Saarlandes, 1995.

23. A Kacem, A Belaid, M Ben Ahmed, *Embedded formulas extraction* . In: International conference on pattern recognition, Barcelona 2000, IEEE, 2000, 676-680.

24. Andreas Kosmala, *HMM-basierte online Handschrifterkennung: ein integrierter Ansatz zur Text- und Formelerkennung* , 2000. Chapter 7

25. Andreas Kosmala, Stephane Lavirotte, Loïc Pottier, Gerhard Rigoll, *On-line handwritten formula recognition using hidden Markov models and context dependent graph grammars* . In: Proc. 5-th international conference on document analysis and recognition, Bangalore 1999, 1999,

26. Andreas Kosmala, Gerhard Rigoll, *On-line handwritten formula recognition* . In: Seong-Whan Lee (ed.), Advances in handwriting recognition, World Scientific, 1999, 539-548.

27. A Kosmala, G Rigoll, *On-line handwritten formula recognition using statistical methods* . In: International conference on pattern recognition, Brisbane 1998, IEEE, 2000, 1306-1308.

28. A Kosmala, G Rigoll, A Brakensiek, *On-line handwritten formula recognition with integrated correction recognition and execution* . In: International conference on pattern recognition, Barcelona 2000, IEEE, 2000, 590-593.

29. B Krenn, C Samuelson, *The linguists guide to statistics* , Brown University, 1997. Also with this: E Cherniak, Statistical techniques for natural language parsing, and S Abney, Statistical methods and linguistics

30. S Lavirotte, L Pottier, *OFR: optical formula recognition*, 1997.

31. S Lavirotte, L Pottier, *Optical formula recognition.* In: International conference on Document Analysis and Recognition, 1997, 357-361.

32. Stephane Lavirotte, Loic Pottier, *Mathematical formula recognition using graph grammar*. In: Daniel P Lopresti and Jiangying Zhou (ed.), Proc. SPIE Vol. 3305: Document recognition and retrieval V, SPIE, 1998, 44-52.

33. Steve Lawrence, C Lee Giles, *Searching the world wide web* , Science **280**(1998), 98-100.

34. Steve Lawrence, C Lee Giles, *Accesibility of information on the web* , Nature **400**(1999), 107-109.

35. H J Lee, J S Wang, *Design of a mathematical expression understanding system* , Pattern Recognition Letters **18**:3(1997), 289-298.

36. G Lenart, *Fuzzy trainable classifiers* , Preprint, Fac. of Mathematics and Informatics Babes-Bolyai Univ., 1992.

37. Ralph Nörenberg, *Character recognition and mathematical formula recognition* , Preprint, Institut für expirementelle Mathematik, Univ. Essen, 2002.

38. Loïc Pottier, *Planar formulas recognition*, Preprint, INRIA Sophia-Antipolis, 1994.

39. Martin Proulx, *A solution to mathematics parsing* , 1996.

40. Saarland University, *Das Optical-Formula-Rcognition (OFR) System der Universität des Saarlandes. Dokumentation* , Inst. Informatik der Univ. des Saarlandes, 1995.

41. Z K Silagadze, *Citations and the Zipf-Mandelbrot law* , J Complex systems **11**:6(1997), 487-499.

42. Alan F Smeaton, *Progress in the application of natural language processing to information retrieval tasks* , The Computer Journal **35**:3(1992), 268-278.

43. T Suzuki, S Aoshima, K Mori, Y Suenaga, *A new system for the real-time recognition of handwritten mathematical formulas* . In: International conference on pattern recognition, Barcelona 2000, IEEE, 2000, 515-518.

44. Z X Wang, C Faure, *Structural analysis of handwritten mathematical expressions* . In: International conference on pattern recognition, IEEE, 1988, 32-34.

45. Michael Yang, Richard Fateman, *Extracting mathematical expressions from postscript documents*. In: ISSAC, 2003,