ELSEVIER

Theoretical Computer Science 213/214 (1999) 1–3

Theoretical
Computer Science

# Preface

The value of a body of knowledge depends very much on whether it is accessible; more precisely, on how well it is accessible. Very roughly the value of two hundred volumes of the journal Theoretical Computer Science (TCS) is practically zero if one has no way to selectively find information in it. One potentially valuable tool is comprehensive indices. Here is a (subject and author) index for volumes 1–200 of the journal TCS. It incorporates the index of volumes 101–150 which appeared in TCS 150 (2) (1995) (but the opportunity has been taken to add some information to that part, particularly in the form of explanations of acronyms).

The numbers behind the key phrases in the index itself refer to around 3 200 articles that have appeared in these 200 volumes. They are numbered in more or less chronological order. This volume also contains the thus numbered list of these articles giving author(s), title, volume number, and page numbers. There are gaps in the numbering sequence; this is just the result of the procedure that assigns numbers to articles and thus it does not mean that there are any articles missing; the list is complete.

There is no claim that this is a perfect index and that all information one could possibly want can be found using it; I do believe that it will turn out to be quite helpful in many cases. I think it is pretty complete as regards meaningful (short) phrases from theoretical computer science in so far as they have turned up in the abstracts of the articles in these 200 volumes; it is not complete in terms of single words for the field; it is also not particularly complete as regards all possible linguistic variations of a phrase that one could think of.

Future updates are planned and eventually I expect and hope (using also other corpora) to generate an adequate list of phrases for the field "Theoretical Computer Science". Meanwhile this list is available to the community at large (also electronically) to experiment with. All those interested are also most cordially invited to send in additional worthwhile phrases (<mich@cwi.nl>). I am particularly interested in testing whether current authors can find enough phrases in this list to describe their publications in the form of key phrases (apart from new terms that they have just invented).

The question arises how far along this index of some 42 000 phrases (not counting inversions) is on the road to a more or less adequate collection for the field. One can make a rather simple stochastic model for how a field evolves, and a corresponding index, as indexing proceeds (chronologically). At starting time (time zero) there is an (unknown) collection of key phrases that is adequate for the field in question. In addition, there is

an infinite universe of potential terms that can be dreamed up by authors and others of new (important) key phrases. Both the field and the index grow stochastically according to easily imagined recipes. (For a precise description of this model see the write-up of the talk "Topologies and metrics on information spaces" that I gave last year in Osnabrück. It is available through the electronic proceedings of that workshop. [1] ) Doing a rather crude preliminary analysis (that still needs to be justified mathematically) one finds a differential equation that is actually explicitly solvable. On the basis of that differential equation and the dynamic evolution of the numbers of terms found as this index progressed, I would guess that this index is around halfway towards an adequate list for the field. Maybe a little further.

One goal in doing these index exercises (there is also one for the journal Artificial Intelligence, Volumes 1–89 [2] ), is to use them as a basis for various software information finding tools. See again the talk already referred to for more details.

The present version of this index, however, is mainly for human use. Thus, for instance, single words that do carry meaning but have 60 or more references have been left out. For instance, 'languages', 'algorithms', 'automata', 'sets', 'families', 'graphs'. That only means that such words do not occur as single-word-complete-phrases in this index; they do frequently occur as parts of two or more word phrases. In the future list to be used in connection with information retrieval tools such single rather frequent (but meaningful) words will play important roles; but for human use they are useless.

Let me make a few remarks on the particular lexicographic ordering that has been used:

- Greek letters are ordered as if they were written out in full as their Latin names. Thus '$\alpha$' is treated as 'alpha' and '$\lambda\eta$' as 'lambdaeta'.
- Diacriticals are ignored. Thus 'Bröcker' is placed immediately near 'Brocker'; in fact, just after.
- The various mathematical and logical signs are ignored lexicographically and so are '(', ')', '-', '%', '#', '*', '/', and the like. Thus 'semi-automatic' is like 'semiautomatic'; '$2 \times 2$' is like 22; and '$\wedge$-nodes' and '$\exists\forall\exists$ formulas' will be found just after 'nodes' and 'formulas', respectively.
- The field of theoretical computer science seems very fond of acronyms, like ACT, FIFO, cpo, CCS, BCF, etc. So much so that there are quite a few cases where there are several meanings. Frequently some explanation has been added between brackets. For instance, 'BCF (breadth-first context free)', 'NL ((a) complexity class)', 'CSS (calculus of sequential systems)'. As regards the lexicographical ordering, these "explanations" are ignored (together with their containing brackets).

Little has been done in the way of linguistic normalization of phrases. Thus, for instance, singulars and plurals can both occur. For instance, there is 'arrays, encoding $\sim$ into trees' and also 'array pattern matching'; also there is both 'polynomial time...' and 'polynomial-time...'. This partly finds its reason in the future uses to be made of this and similar lists of phrases. There is also the matter of such variations like 'fixpoint', 'fixed point' 'fixed-point', 'fix point', 'least fixed-point', 'least-fixed-point',..., where it is not very

[1] http://www.mathematik.uni-osnabrueck.de/projects/workshop97/proc.html.

[2] Artificial Intelligence 96 (1) (1997) 3–189.

clear whether there is a "preferred term". By and large, the variations as they occur in the original abstracts have been retained.

As already indicated, this index is just a first step. It is an offering to the community to use, to play with, and to criticize and improve.

Michiel Hazewinkel
Bussum, 30 May 1998