ELSEVIER

Theoretical Computer Science 150 (1995) 195–196

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. One potentially valuable tool is indexing. Here is a (subject and author) index for volumes 101–150 of the journal *Theoretical Computer Science*. We, the initiators at Elsevier and the ones who carried it out (Harry Bego, Stijn van Dongen, and the undersigned) by no means claim that this is a perfect index, but we do think that with it we have provided the science community with a valuable tool for accessing the information in these 50 volumes of *TCS*.

The numbers behind the key phrases in the index itself refer to the 806 articles that appeared in these fifty volumes. They are numbered in historical order. This issue also contains the thus numbered list of these articles giving authors(s), title, volume number, and page numbers.

Given the fact that an index can be an invaluable information finding tool, especially for a larger body of text, it is perhaps surprising that so many books appear with indexes that are virtually worthless (and also still a few with no index at all).

First, authors seem to feel, provided their work is important enough and written well enough, that knowledge of it will pass to the right persons in any case in some unspecified way, and that there is no need to slave away at making an index or to provide other information retrieval tools. And to some extent they are no doubt right.

Second, making an index is hard work and quite time consuming (and, once one has reached that stage, things are often well past a deadline). In this respect things are looking up. Most high-end word processors now come with index-making facilities (based on marking suitable phrases in the text) and these are certainly a great help, though it is still a hard job, even with these electronic aids. Very promising indeed, are special indexing software packages which can pick out suitable noun phrases from an ASCII file. It was partly to experiment with such packages that we undertook the present job.

I would like to say a few words on how the present index was put together. We worked on the basis of the abstracts only; key phrases are not used in *TCS*. After suitable preparation of the material, not a trivial task, even though much of the material was available in electronic form, the commercially available program TExtract™, developed by Harry Bego, was used to pick out a suitable collection of noun phrases. At the same time I made a rough and ready index by hand of the first fifty articles. Comparing the

Elsevier Science B.V.

results showed that the hand-made index contains longer phrases as a rule and that most of these were prepositional noun phrases, i.e. noun phrases linked by prepositions. An ad hoc program was written by Harry Bego to admit prepositional noun phrases around the single preposition "of" (no doubt the most frequently occurring one in the kind of key phrases desired for an index). This was a success. After having done all that (several times), about 46 hours of human editing finished the job.

We also asked David Evans from Pittsburgh to do a preliminary run of the material through his system CLARIT. This he obligingly did, and he provided us with a raw list of noun phrases (no cross references to the articles) with weights giving the relative importance (frequency) of the noun phrases listed. After throwing away the obvious garbage in both lists, I compared the results of a not particularly random sample of 378 items in total. Of these 306 appeared in both lists, 51 were in the TExtract™ list but not in the CLARIT list, and 21 were in the CLARIT list but not in the TExtract™ list.

As already mentioned, the dedicated index-making software programs of the current generation tend to produce a fair amount of obvious garbage. Much of this is caused by the fact that many words in English can be both a noun and a verb. For instance the three-word phrase "parameter values covers" could well consist of three nouns. Another major cause of garbage are nouns which are irrelevant for the domain of science at hand, but which still occur frequently enough to be detected, and which could be of great relevance for another branch of science. Apart from these two major groups of garbage, there was a great deal more, and I could now fill several pages with curious examples (to the human mind) that the machine picked out in its inexorable logical way.

After weeding out the obvious garbage there is till quite a bit of irrelevant material left. Let me give a few examples. The nouns "modifications", "optimization", "methodology", and the noun phrase "minimum number" came up 8, 8, 4, 3 times, respectively. All could be part of a good index phrase. For instance "modification" is a technical term in certain parts of mathematics, though perhaps not in computer science. As turned out 2 of the occurrences of "modification" and 5 of "optimization" were of use (usually as part of a longer phrase that had to be found by human inspection) and all others were irrelevant. It is at this stage that good post-editing tools are very important (and good packages like TExtract™ have them).

I think that it is clear from this exercise that there are a good many things that can be done to improve the current generation of indexing tools substantially, and I also think that we have gained some promising ideas from this job. In particular, much should be possible by making more use of natural language linguistic knowledge.

Another thing that has become clear is, that it is a far from trivial task to write a good abstract, and also that author-supplied key-phrases would be a most valuable addition (from the point of view of information retrieval).

MICHIEL HAZEWINKEL
Bussum, 22 August 1995