PARALLELIZE EDIT DISTANCE ALGORITHM

Artur Niewiarowski, Marek Stanuszek
Cracow University of Technology, Cracow, Poland

About the Levenshtein distance

- The Levenshtein distance between two strings is equal to the minimum number of insertions, deletions and substitutions of chars required to change one string into the second one.
- The algorithm creates a matrix where its last element states as the solution.

Levenshtein distance algorithm is described by the formula:

$$\prod_{i=1}^{N}\prod_{j=1}^{M} d(i,j) = \min(d(i-1,j)+1, d(i,j-1)+1, d(i-1,j-1)+\beta)$$

$$\begin{cases} \beta = 0 : a(i) \equiv b(j) \\ \beta = 1 : a(i) \neq b(j) \\ d(i,0) = i \\ d(0,j) = j \\ d(0,0) = 0 \end{cases}$$

		K	U		K	D	M	'	1	3
	0	1	2	3	4	5	6	7	8	9
K	1	0	1	2	3	4	5	6	7	8
U	2	1	0	1	2	3	4	5	6	7
	3	2	1	0	1	2	3	4	5	6
K	4	3	2	1	0	1	2	3	4	5
D	5	4	3	2	1	0	1	2	3	4
M	6	5	4	3	2	1	0	1	2	3
'	7	6	5	4	3	2	1	0	1	2
1	8	7	6	5	4	3	2	1	0	1
4	9	8	7	6	5	4	3	2	1	1

Fig. 1. Example of Levenshtein matrix

where:
$\prod_{i=1}^{N}$ – symbol for the iteration, for i = (1, ..., N),

d – matrix sizes N+1, M+1, made from two terms,
N, M – length of two terms,
$d(i,j)$ - (i,j) – element of matrix **d**,
min – function returns minimum of two variables,
β – variable that gets values: 0 or 1,
a(i) – i – element in string of term a,
b(j) – j – element in string of term b.

Levenshtein distance K is a minimum number of operations (insertion, deletion, substitution) required to change one term into the other.

$K = d(N,M)$

Details of a problem:

- ❑ Levenshtein distance
- ❑ Levenshtein-Damerau distance
- ❑ very large matrix (e.g. 9,00E+12 elements)

Used technologies:

- ❑ Microsoft .NET (Framework 4.0)
- ❑ Xamarin Mono (for OS Linux)

Examples of the use:

- ❑ texts (documents) analysis
- ❑ analysis of DNA sequences

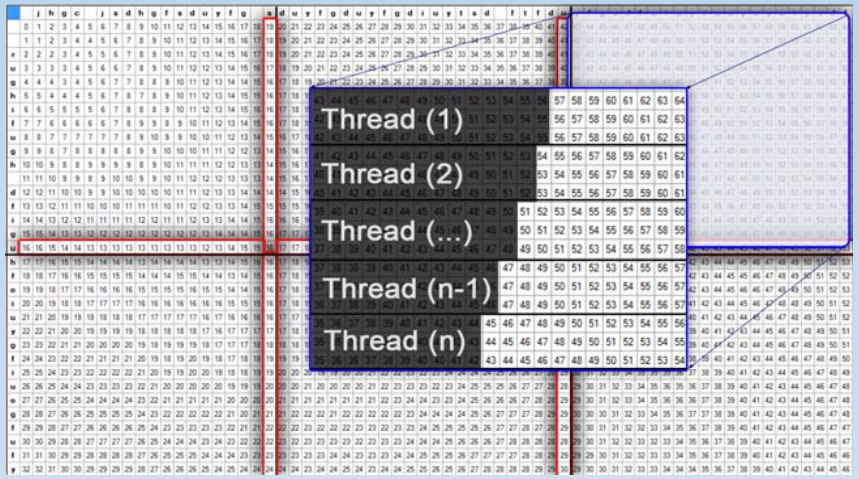

Fig. 1. Parallelized procedure of Levenshtein distance matrix decomposition

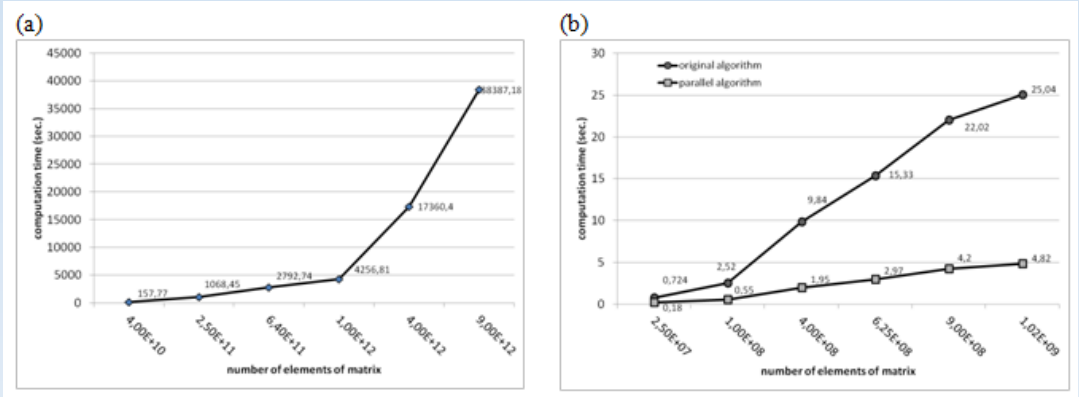

Fig. 2. Computation time of calculations for very long strings. (a) using matrix decomposition and parallel computations, (b) using original algorithm and parallel method for one decomposed matrix