Computer Science > Computation and Language

Good parts first - a new algorithm for approximate search in lexica and string databases

Stefan Gerdjikov, Stoyan Mihov, Petar Mitankin, Klaus U. Schulz

(Submitted on 4 Jan 2013 (v1), last revised 3 Dec 2015 (this version, v2))

We present a new efficient method for approximate search in electronic lexica. Given an input string (the pattern) and a similarity threshold, the algorithm retrieves all entries of the lexicon that are sufficiently similar to the pattern. Search is organized in subsearches that always start with an exact partial match where a substring of the input pattern is aligned with a substring of a lexicon word. Afterwards this partial match is extended stepwise to larger substrings. For aligning further parts of the pattern with corresponding parts of lexicon entries, more errors are tolerated at each subsequent step. For supporting this alignment order, which may start at any part of the pattern, the lexicon is represented as a structure that enables

immediate access to any substring of a lexicon word and permits the extension of such substrings in both directions. Experimental evaluations of the approximate search procedure are given that show significant efficiency improvements compared to existing techniques. Since the technique can be used for large error bounds it offers interesting possibilities for approximate search in special collections of "long" strings, such as phrases, sentences, or book ti

Subjects: **Computation and Language (cs.CL)**; Data Structures and Algorithms (cs.DS)
Cite as: **arXiv:1301.0722 [cs.CL]**
(or **arXiv:1301.0722v2 [cs.CL]** for this version)

Submission history

From: Stefan Gerdjikov [view email]
[v1] Fri, 4 Jan 2013 13:45:35 GMT (154kb)
[v2] Thu, 3 Dec 2015 10:53:17 GMT (154kb)

Good parts first-a new algorithm for approximate search in lexica and string databases, chizelevanie rejects the rhythm.