Applied Mathematical Sciences, Vol. 7, 2013, no. 47, 2347 - 2352
HIKARI Ltd, www.m-hikari.com

The Minimum k-Cover Problem

Anna Gorbenko

Department of Intelligent Systems and Robotics
Ural Federal University
620083 Ekaterinburg, Russia
gorbenko.ann@gmail.com

Vladimir Popov

Department of Intelligent Systems and Robotics
Ural Federal University
620083 Ekaterinburg, Russia
Vladimir.Popov@usu.ru

Abstract

We consider the problem of determining the minimum cardinality collection of substrings, each of given length $k \geq 2$, that "cover" a given string x of length n. We describe an approach to solve this problem. This approach is based on constructing an explicit reduction from the problem to the satisfiability problem.

Keywords: strings, k-covers, satisfiability

Different problems of finding regularities are thoroughly studied in theoretical computer science (see e.g. [1] – [6]). In particular, the minimum k-cover problem was introduced in [7].

Given a nonempty string x of length n, a set $V = \{v_1, v_2, \ldots, v_p\}$ of p substrings of x. We say that V is a cover for x if and only if every position of x lies within an occurrence of some v_i, $1 \leq i \leq p$. In addition, if each string in V has length k, then V is a k-cover of x. If p is the minimum integer for which such a set V exists, then V is said to be a minimum k-cover of x.

THE MINIMUM k-COVER PROBLEM (MCP):
INSTANCE: *An alphabet* Σ*, a string* X *over* Σ*, positive integers* k *and* p.
QUESTION: *Whether there exists a* k*-cover of* X *of cardinality* p*?*

The minimum k-cover problem is **NP**-complete (see [8]). Encoding problems as Boolean satisfiability and solving them with very efficient satisfiability algorithms has recently caused considerable interest (see e.g. [9] – [25]). In this paper, we consider an explicit reduction from MCP to the satisfiability problem. For simplicity, we use $S[i]$ to denote the ith letter in sequence S, and $S[i, j]$ to denote the substring of S consisting of the ith letter through the jth letter. Let $\Sigma = \{a_1, a_2, \ldots, a_{|\Sigma|}\}$. Let

$$\varphi[1, i, j] = \vee_{1 \le l \le |\Sigma|} x[i, j, l],$$

$$\varphi[2, i, j] = \wedge_{1 \le l[1] \le |\Sigma|, 1 \le l[2] \le |\Sigma|, l[1] \ne l[2]} (\neg x[i, j, l[1]] \vee \neg x[i, j, l[2]]),$$

$$\varphi[i, j] = \varphi[1, i, j] \wedge \varphi[2, i, j],$$

$$\varphi = \wedge_{1 \le i \le p, 1 \le j \le |k|} \varphi[i, j],$$

$$\psi[i] = \vee_{1 \le j \le |X| - k + 1} y[i, j],$$

$$\psi = \wedge_{1 \le i \le p} \psi[i],$$

$$\rho[i] = \vee_{1 \le j \le p, h_i \le l \le i, h_i = 1, \text{ if } i \le k, h_i = i - k + 1, \text{ if } i > k} y[j, l],$$

$$\rho = \wedge_{1 \le i \le |X|} \rho[i],$$

$$\tau[1, i] = \wedge_{1 \le j \le |\Sigma|, X[i] = a_l, l \ne j} \neg z[i, j],$$

$$\tau[2] = \wedge_{1 \le i \le |X|, X[i] = a_j} z[i, j],$$

$$\tau = \tau[2] \wedge \wedge_{1 \le i \le |X|} \tau[1, i],$$

$$\eta = \wedge_{1 \le i \le p, 1 \le j \le |X| - k + 1, 0 \le t \le k - 1, 1 \le l \le |\Sigma|} y[i, j] \to z[j + t, l] = x[i, 1 + t, l],$$

$$\xi = \varphi \wedge \psi \wedge \rho \wedge \tau \wedge \eta.$$

Theorem. *Given a fixed alphabet* Σ*, a string* X *over* Σ*, positive integers* k *and* p*. There is a* k*-cover of* X *of cardinality* p *if and only if* ξ *is satisfiable.*

Proof. Suppose that there is $V = \{v_1, v_2, \ldots, v_p\}$ that is a k-cover of X of cardinality p. Let $x[i, j, l] = 1$ where $1 \le i \le p$, $1 \le j \le k$, $v_i[j] = a_l$; $x[i, j, l] = 0$ where $1 \le i \le p$, $1 \le j \le k$, $v_i[j] \ne a_l$; $y[i, j] = 1$ if and only if $X[j, j + k - 1] = v_i$ where $1 \le i \le p$, $1 \le j \le |X| - k + 1$; $z[i, j] = 1$ where $1 \le i \le |X|$, $1 \le j \le |\Sigma|$, $X[i] = a_j$; $z[i, j] = 0$ where $1 \le i \le |X|$, $1 \le j \le |\Sigma|$, $X[i] \ne a_j$.

Since $V \subseteq \Sigma^k$, for all i and j there is l such that $x[i, j, l] = 1$. Therefore, $\varphi[1, i, j] = 1$. In view of $x[i, j, l] = 0$ where $1 \le i \le p$, $1 \le j \le k$, $v_i[j] \ne a_l$, it is clear that there is no more than one value of l such that $x[i, j, l] = 1$. Hence either $x[i, j, l[1]] = 0$ or $x[i, j, l[2]]$ for all i, j, $l[1] \ne l[2]$. Therefore, $\varphi[2, i, j] = 1$. So, $\varphi = 1$.

Note that V is a set of substrings of X. Since $y[i,j] = 1$ if and only if $X[j, j+k-1] = v_i$, it is easy to see that $\psi[i] = 1$. By definition, $\psi[2,i] = 1$. So, $\psi = 1$.

Since V is a k-cover of X, $X[r, r+k-1] = v_j$ for some r and j such that $1 \leq j \leq p$, $r \leq i \leq r+k-1$. Therefore, $\rho[i] = 1$. So, $\rho = 1$. Since $z[i,j] = 1$ where $1 \leq i \leq |X|$, $1 \leq j \leq |\Sigma|$, $X[i] = a_j$; $z[i,j] = 0$ where $1 \leq i \leq |X|$, $1 \leq j \leq |\Sigma|$, $X[i] \neq a_j$, it is easy to check that $\tau = 1$. Since V is a k-cover of X, it is clear that $\eta = 1$. Therefore, $\xi = 1$.

Suppose now that $\xi = 1$. Hence $\xi = \varphi = \psi = \rho = \tau = \eta = 1$. Since $\varphi = 1$, by definition, $\varphi[1,i,j] = 1$, $\varphi[2,i,j] = 1$. It is easy to check that for all i and j there is only one value of l such that $x[i,j,l] = 1$. Let $v_i[j] = a_l$. Since $\eta = 1$ and $\tau = 1$, it is clear that if $y[i,j] = 1$, then $X[j, j+k-1] = v_i$. In view of $\rho = 1$, we obtain that V is a k-cover of X. □

In view of the theorem, we obtain an explicit reduction from MCP to PSAT.

Note that $\alpha \to \beta \Leftrightarrow \neg\alpha \vee \beta$, $\alpha = \beta \Leftrightarrow (\neg\alpha \vee \beta) \wedge (\alpha \vee \neg\beta)$. Therefore, $\eta = \Leftrightarrow \eta'$ where

$$\eta' = \wedge_{1 \leq i \leq p, 1 \leq j \leq |X|-k+1, 1 \leq t \leq k-1, 1 \leq l \leq |\Sigma|} (\neg y[i,j] \vee \neg z[j+t, l] \vee x[i, 1+t, l]) \wedge$$
$$(\neg y[i,j] \vee z[j+t, l] \vee \neg x[i, 1+t, l]).$$

Let $\xi' = \varphi \wedge \psi \wedge \rho \wedge \tau \wedge \eta'$. It is clear that $\xi \Leftrightarrow \xi'$. Since ξ' is a CNF, we obtain an explicit reduction from MCP to SAT.

Using standard transformations (see e.g. [26]) we can obtain an explicit transformation ξ' into ξ'' such that $\xi' \Leftrightarrow \xi''$ and ξ'' is a 3-CNF. It is easy to see that ξ'' gives us an explicit reduction from MCP to 3SAT.

There is a well known site on which posted solvers for SAT [27]. They are divided into two main classes: stochastic local search algorithms and algorithms improved exhaustive search. All solvers allow the conventional format for recording DIMACS boolean function in conjunctive normal form and solve the corresponding problem [28]. In addition to the solvers the site also represented a large set of test problems in the format of DIMACS. This set includes a randomly generated problems of 3SAT.

We create a generator of natural instances for LCS. Also we use test problems from [27]. We use algorithms from [27]. Also we design our own genetic algorithm for SAT which based on algorithms from [27].

We use heterogeneous cluster based on three clusters (Cluster USU, Linux, 8 calculation nodes, Intel Pentium IV 2.40GHz processors; umt, Linux, 256 calculation nodes, Xeon 3.00GHz processors; um64, Linux, 124 calculation nodes, AMD Opteron 2.6GHz bi-processors) [29].

Each test was run on a cluster of at least 100 nodes. The maximum solution time was 6 hours. The average time to find a solution was 11.4 minutes. The best time was 7 seconds.

ACKNOWLEDGEMENTS. The work was partially supported by Analytical Departmental Program "Developing the scientific potential of high school" 8.1616.2011.

References

[1] V. Yu. Popov, Computational complexity of problems related to DNA sequencing by hybridization, *Doklady Mathematics*, 72 (2005), 642-644.

[2] V. Popov, The approximate period problem for DNA alphabet, *Theoretical Computer Science*, 304 (2003), 443-447.

[3] V. Popov, The Approximate Period Problem, *IAENG International Journal of Computer Science*, 36 (2009), 268-274.

[4] V. Popov, Approximate Periods of Strings for Absolute Distances, *Applied Mathematical Sciences*, 6 (2012), 6713-6717.

[5] V. Popov, Multiple genome rearrangement by swaps and by element duplications, *Theoretical Computer Science*, 385 (2007), 115-126.

[6] V. Popov, Sorting by prefix reversals, *IAENG International Journal of Applied Mathematics*, 40 (2010), 247-250.

[7] Iliopoulos, C.S., Smyth, W.F., "On-line algorithms for k-covering," *Proceedings of the Ninth Australasian Workshop on Combinatorial Algorithms*, Curtin University of Technology, pp. 107-116 7 /98.

[8] Cole, R., Iliopoulos, C.S., Mohamed, M., Smyth, W.F., Yang, L., "Computing the minimum k-cover of a string," *Proceedings of the Prague Stringology Conference*, Prague Stringology Club, pp. 51-62 9 /03.

[9] A. Gorbenko and V. Popov, The Minimum Test Collection Problem, *Applied Mathematical Sciences*, 7 (2013), 1191-1193.

[10] A. Gorbenko and V. Popov, The Farthest Substring Problem, *Applied Mathematical Sciences*, 7 (2013), 1209-1212.

[11] A. Gorbenko and V. Popov, Computational Experiments for the Problem of Hamiltonian Path with Fixed Number of Color Repetitions, *Advanced Studies in Theoretical Physics*, 7 (2013), 121-126.

[12] A. Gorbenko and V. Popov, On Hamilton Paths in Grid Graphs, *Advanced Studies in Theoretical Physics*, 7 (2013), 127-130.

[13] A. Gorbenko and V. Popov, Computational Experiments for the Problem of Sensor-Mission Assignment in Wireless Sensor Networks, *Advanced Studies in Theoretical Physics*, 7 (2013), 135-139.

[14] A. Gorbenko and V. Popov, The Swap Common Superstring Problem, *Applied Mathematical Sciences*, 7 (2013), 609-614.

[15] A. Gorbenko and V. Popov, The String Barcoding Problem, *Applied Mathematical Sciences*, 7 (2013), 615-622.

[16] A. Gorbenko and V. Popov, On Multiple Occurrences Shortest Common Superstring Problem, *Applied Mathematical Sciences*, 7 (2013), 641-644.

[17] A. Gorbenko and V. Popov, Longest Common Parameterized Subsequences with Fixed Common Substring, *Applied Mathematical Sciences*, 7 (2013), 645-650.

[18] A. Gorbenko and V. Popov, The Hamiltonian Alternating Path Problem, *IAENG International Journal of Applied Mathematics*, 42 (2012), 204-213.

[19] A. Gorbenko and V. Popov, The Problem of Finding Two Edge-Disjoint Hamiltonian Cycles, *Applied Mathematical Sciences*, 6 (2012), 6563-6566.

[20] A. Gorbenko and V. Popov, Multiple Occurrences Shortest Common Superstring Problem, *Applied Mathematical Sciences*, 6 (2012), 6573-6576.

[21] A. Gorbenko and V. Popov, The Far From Most String Problem, *Applied Mathematical Sciences*, 6 (2012), 6719-6724.

[22] A. Gorbenko and V. Popov, Hamiltonian Alternating Cycles with Fixed Number of Color Appearances, *Applied Mathematical Sciences*, 6 (2012), 6729-6731.

[23] A. Gorbenko and V. Popov, On the Longest Common Subsequence Problem, *Applied Mathematical Sciences*, 6 (2012), 5781-5787.

[24] A. Gorbenko and V. Popov, The Longest Common Parameterized Subsequence Problem, *Applied Mathematical Sciences* 6 (2012), 2851-2855.

[25] A. Gorbenko and V. Popov, The set of parameterized k-covers problem, *Theoretical Computer Science*, 423 (2012), 19-24.

[26] A. Gorbenko and V. Popov, The c-Fragment Longest Arc-Preserving Common Subsequence Problem, *IAENG International Journal of Computer Science*, 39 (2012), 231-238.

[27] http://people.cs.ubc.ca/∼hoos/SATLIB/index-ubc.html

[28] http://www.cs.ubc.ca/∼hoos/SATLIB/Benchmarks/SAT/satformat.ps

[29] http://parallel.imm.uran.ru/mvc_now/hardware/supercomp.htm

Received: February 12, 2013