

Algebraic method for solution of some best matching problems

Michail SCHLESINGER

Institute of Cybernetics,
Ukrainian Academy of Sciences

40, Prospect Akademika Glusckova
252207, Kiev, Ukraine
Tel.: (044) 266-25-69 Fax: (044) 266-1570
schles%image.kiev.ua@ts.kiev.ua

Michail Schlesinger. Algebraic method for solution of some best matching problems

A problem of calculation of distance by Levenstein between string and regular language is investigated. A new approach for solution of such sort of problems is proposed. The new approach is based on the concept of generalized convolutions of functions and their equivalent transformations. On this basis both problem formulation and problem solution can be represented within the framework of a same mathematical formalism, and it becomes possible to obtain an expression for the problem solution as a result of equivalent transformation of the problem formulation expression.

1. Formulation and discussion of the problem

1.1. Introductory notions

Let V be a finite alphabet of signals, V^* be the set of strings, which are composed of signals from V , L be a subset in V^* and, finally, $d:V^* \times V^* \rightarrow R$ be a distinction function.

The best matching problem is meant as a calculation of the value

$$D(\bar{u}) = \min_{\bar{v} \in L} d(\bar{v}, \bar{u}) \quad (1)$$

for the given string \bar{u} and the given subset L of strings.

In this article the special case of the problem is discussed and solved, when L is a regular language and d is a distinction function by Levenstein [1]. Let us turn into consideration the notions, which the regular languages and the distinctions by Levenstein will be defined by.

1.2. Regular languages

Let S be a finite set of states; V , as before, is a finite set of signal values. Let $B:S \rightarrow \{0, \infty\}$, $T:S \times V \times S \rightarrow \{0, \infty\}$ and $E:S \rightarrow \{0, \infty\}$ be three functions.

Definition 1. A string (\bar{v}, s) , $\bar{v} \in V^*$, $s \in S$ is allowable by functions B and T , if:

1. \bar{v} is an empty string and $B(s) = 0$;

or

2. $\bar{v} = (\bar{v}', v)$, $\bar{v}' \in V^*$, $v \in V$, and there exists a state $s' \in S$,

such that the string (\bar{v}', s') is allowable and $T(s', v, s) = 0$.

Definition 2. A string \bar{v} is allowable by the functions B , T , E , if there exists a state $s \in S$, such that the string (\bar{v}, s) is allowable by functions B , T and $E(s) = 0$.

The set of allowable strings will be designated by $L_{B,T,E}$. Evidently, $L_{B,T,E}$ is a regular language [2].

1.3. Distinction by Levenstein

Let us consider the operations of the following three types over the strings and define their costs.

An operation “to change” transforms a string of type (\bar{v}, v, \bar{v}') , $\bar{v} \in V^*$, $v \in V$, $\bar{v}' \in V^*$, into another one (\bar{v}, u, \bar{v}') , $u \in V$. Costs of such operations are defined by some function $CH: V \times V \rightarrow R$.

An operation “to delete” transforms a string of type (\bar{v}, v, \bar{v}') into new one (\bar{v}, \bar{v}') . Costs of such type operations are defined by some function $DE: V \rightarrow R$.

An operation “to insert” transforms (\bar{v}, \bar{v}') into (\bar{v}, v, \bar{v}') . Its cost is defined by a function $IN: V \rightarrow R$.

Let some chain $\bar{v}_1, \bar{v}_2, \dots, \bar{v}_n$ of strings be such that every string $\bar{v}_i, i \neq 1$ can be obtained from the string \bar{v}_{i-1} by an operation c_i , every c_i being an operation of above mentioned type. Then we will say that this chain is a path from \bar{v}_1 to \bar{v}_n and that this path costs $\sum_{i=2}^n \varphi(c_i)$, where $\varphi(c_i)$ is a cost of the operation c_i .

Definition 3. A distinction by Levenstein of a string \bar{v} from a string \bar{u} is the cost of the cheapest path from \bar{v} to \bar{u} .

So defined distinction function will be designated by $d_{IN,CH,DE}$.

1.4. Formulation of the problem

It is necessary to construct the algorithm, that for every six-tuple B, T, E, CH, DE, IN of functions and for every string \bar{u} calculates a value

$$D(\bar{u}) = \min_{\bar{v} \in L_{B,T,E}} d_{IN,CH,DE}(\bar{v}, \bar{u}). \quad (2)$$

The relevant results on this izem are published in [3,4,5].

2. Formulation of the main result.

Let the functions B, T, E define the language $L_{B,T,E}$ and let the functions IN, CH, DE define the Levenstein's distinction $d_{IN,CH,DE}$, as it was defined above. The following theorem is valid [6].

Theorem. For any six-tuple B, T, E, CH, IN, DE of functions there exist such three functions $b: S \rightarrow R$, $f: S \times V \times S \rightarrow R$ and $e: S \rightarrow R$, that the equality

$$\begin{aligned}
D(u) &= \min_{\bar{v} \in L_{B,T,E}} d_{IN,CH,DE}(\bar{v}, \bar{u}) = \\
&= \min_{s_0, s_1, \dots, s_n} \left[b(s_0) + \sum_{i=1}^n f(s_{i-1}, u_i, s_i) + e(s_n) \right]
\end{aligned} \tag{3}$$

is valid for every string $\bar{u} = (u_1, u_2, \dots, u_n)$, $u_i \in V$.

Through the theorem the problem solution for every given language and every given distinction function consists of two steps. On the first step the functions B, T, E, IN, CH, DE are being converted into the functions b, f, e , whose properties and existence are stated by the theorem. The string \bar{u} under analysis is not used on this step. On the second step value $D(\bar{u})$ is calculated, using the expression in the right side of (3). These calculations are to be fulfilled in the following way.

Let f_0, f_1, \dots, f_n be some functions of type $S \rightarrow R$, n being length of the string \bar{u} under analysis. These functions are defined by the following expressions:

$$\begin{aligned}
f_0(s) &= b(s), \quad s \in S; \\
f_i(s) &= \min_{s' \in S} [f_{i-1}(s') + f(s', u_i, s)], \quad s \in S, \quad i = 1, \dots, n.
\end{aligned} \tag{4}$$

Then the value $\min_{s \in S} [f_n(s) + e(s)]$ is the solution of the problem, i.e. $D(\bar{u})$. One can see, that this way of computation has a complexity of order $k^2 \times n$, k being the amount of states in S .

The necessary proofs and explanations of the main result are described in [6,7].

References:

1. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. АН СССР.- 1965.-163, № 4. - С. 840-850.
2. Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции.- М. : Мир, 1978.- Т. 1.
3. Wagner, R. A., and Fischer, M. J. The string-to-string correction problem. J. ACM 21, 1 (Jan. 1974), 168-173.
4. Wagner, R. A., and Seiferas, J. I. Correcting counter-automaton-recognizable languages. SIAM. J. Comput. 7, 3 (1978), 357-375.
5. Aho, A. V., Hopcroft, J. E., and Ullman, J. D. The Design and Analysis of Computer Algorithms. Addison-Wesley, Reading, Mass., 1975.
6. Schlesinger, M. I. Systeme von Funktionsoperationen angewendet auf eine Aufgabe der besten Uebereinstimmung. //ISSN 0863-0798/Wissenschaftliche Beitrage zur Informatik - Fakultat Informatik TU Dresden/7(1994) Heft 3.- S. 62-79.
7. Шлезингер М.И. Обобщенные свертки функций и их применение для синтаксического анализа искаженных последовательностей //Проблемы управления и информатики, Киев, 1995 г, № 2.- С. 67-81.