South African
Computer
Journal
Number 11
May 1994

Suid-Afrikaanse
Rekenaar-
tydskrif
Nommer 11
Mei 1994

Computer Science and Information Systems

Rekenaarwetenskap en Inligtingstelsels

The South African Computer Journal
An official publication of the Computer Society of South Africa and the South African Institute of Computer Scientists

Die Suid-Afrikaanse Rekenaartydskrif
'n Amptelike publikasie van die Rekenaarverenging van Suid-Afrika en die Suid-Afrikaanse Instituut vir Rekenaarwetenskaplikes

Editor
Professor Derrick G Kourie
Department of Computer Science
University of Pretoria
Hatfield 0083
Email: dkourie@dos-lan.cs.up.ac.za

Subeditor: Information Systems
Prof John Shochot
University of the Witwatersrand
Private Bag 3
WITS 2050
Email: 035ebrs@witsvma.wits.ac.za

Production Editor
Dr Riël Smit
Mosaic Software (Pty) Ltd
P.O.Box 16650
Vlaeberg 8018
Email: gds@cs.uct.ac.za

Subscriptions

	Annual	Single copy
Southern Africa:	R45,00	R15,00
Elsewhere:	$45,00	$15,00

to be sent to:
Computer Society of South Africa
Box 1714 Halfway House 1685

Phone: +27 (11) 315-1319 Fax: +27 (11) 315-2276

Technical Note

An Evaluation of Substring Algorithms that Determine Similarity Between Surnames

G de V de Kock C du Plessis
Department of Computer Science, University of Port Elizabeth, P.O. Box 1600, Port Elizabeth, 6000

Abstract

The problem investigated in this study is, given a surname, determine similar surnames in a genealogical database. There exist a number of algorithms to determine the similarity between two strings based on their common substrings. The surnames in an existing genealogical database were used in an evaluation process to determine the relative success of these algorithms. The methods used to evaluate the performance of the algorithms and the algorithms are discussed briefly.
Keywords: *genealogical database, word, string and surname matching*
Computing Review Categories: *H.5, I.1.7*

Received: June 1993, Accepted: August 1993, Final version: December 1993.

1 Introduction

The primary problem is, given a surname, determine similar surnames in a genealogical database. More information on the origin of the problem is given in [2]. Other algorithms considered are given in [4] and [3]. In this paper applicable substring algorithms are briefly discussed, in some cases adapted and then evaluated. The norms and statistics used for the evaluation are also stated.

All the surnames in the UPE genealogical database at the time of the study formed the test dataset. Prepositions such as "de", "van der", "le" etc. have been dropped from the surnames in the test dataset.

Surnames with the same origin, spelling variations and aliases have been grouped together in mutually exclusive equivalence classes, the so-called *ideal classes* or *partition*.

A mathematical formulation of the primary problem follows:
Given a set of surnames, V, (the test dataset)

$$V = \{v_1, v_2, \ldots v_n\}$$

and the *ideal partition*, $P = \{V_1, V_2, \ldots V_p\}$
with $V = \bigcup_{i=1}^{p} V_i$ and $V_i \cap V_j = \phi \forall i \neq j$.
Given a surname, v, determine:

- i such that v is similar to the surnames in V_i, or
- the set of surnames in V which has a similarity to v greater than a predetermined value using some or other norm (criterion).

The basic statistics of the test dataset are given in Table 1.

2 Success norms

The success norms discussed in [2] are stated briefly. Let A^* be the set of all possible strings over the alphabet A, then $V \subset A^*$. The similarity between two strings is usually a function, $G : A^* \times A^* \rightarrow E$, where E is normally the interval [0,1] on the real number line. Most of the similarity norms between two strings, u and v, are of the form:

$$G(u,v) = \frac{T(u,v)}{N(u,v)}$$

where $T(u,v)$ is a function of the common substrings in u and v, and $N(u,v)$ is a normalising function to ensure that $G(u,u) = 1$. Furthermore, it is required that $G(u,v) = G(v,u)$ and $G(u,v) = G(u^R, v^R)$, where u^R is the reverse of u.

A distance $d = 1 - G$, can be defined between two strings. The function d is not necessarily a metric, since in most cases the triangular inequality is not satisfied.

For each algorithm evaluated, a centre (c_i) and a radius (r_i) are defined for each V_i:

$$r_i = \min_{v \in V_i} \max_{w \in V_i} d(v,w)$$

The centre, c_i, of V_i, is the element of V_i, such that

1. $r_i = \max_{w \in V_i} d(c_i, w)$
2. $|\{w : w \in V, d(c_i,w) \leq r_i\} \cap (V - V_i)|$, is a minimum.

Define the so-called "circle" with centre c_i, and radius $r_i + \gamma$, with $\gamma \geq 0$, as follows:

$$C_i(\gamma) = \{w : w \in V \text{ and } d(c_i,w) \leq r_i + \gamma\}$$

Thus, $V_i \subseteq C_i(\gamma) \forall i$.

Table 1. Test dataset – constants

Explanation	Symbol/ formula	Value
The number of different surnames in the test dataset	$n = \|V\|$	7 169
The number of *ideal classes* in the ideal partition	p	4 093
The average number of surnames per ideal class	$y_P = \frac{n}{p}$	1,75
The ideal classes with more than one element, i.e $B = \{V_i : V_i \in P \text{ and } \|V_i\| > 1\}$	$n_1 = \|B\|$	1 364
The average number of surnames per class in B	$y_B = \frac{\sum_{V_i \in B} \|V_i\|}{n_1}$	3,26
The number of surnames in the test dataset ($\rho(v)$ is the frequency of occurrence for a surname, v)	$N = \sum_{v \in V} \rho(v)$	92 327
The average length of a surname in characters	ℓ	6,93

Any set $U = \{U_1, U_2, \ldots, U_p\}$, such that $V_i \subseteq U_i$, $\forall V_i \in V$ is called a *class cover* or a *C-cover* of V.

It follows that $\bigcup_{i=1}^{p} U_i = V$.

For any $\gamma \geq 0$, $C_\gamma = \{C_1(\gamma), C_2(\gamma), \ldots, C_p(\gamma)\}$ is a C-cover of V. Here the discussion is restricted to $U = C_0$.

The success of the algorithm is defined as the percentage of the elements in V which appear in only one U_i. It can formally be calculated in two ways, viz:

$$S_1(U) = \frac{100}{n}\left(|V| - \left(\sum_{i=1}^{p} \left|\bigcup_{\substack{j=1 \\ j \neq i}}^{p} V_i \cap U_j\right|\right)\right) \qquad (1)$$

Secondly, the frequencies of the surnames are taken into account.

Let $D_i = \{v : v \in \bigcup_{\substack{j=1 \\ j \neq i}}^{p} V_i \cap U_j \text{ and } v \in V_i\}$.

D_i is the set of elements of V_i which appear in other *ideal classes'* "circles".

$$S_2(U) = \frac{100}{N}\left(N - \left(\sum_{i=1}^{p} \sum_{v \in D_i} \rho(v)\right)\right) \qquad (2)$$

Practical problems

In applying these success norms a number of practical problems have been experienced resulting in the following adjustments:

a. A surname, $u \in V_i$, may have such a small similarity to the other surnames in V_i, that the radius r_i is very large. If it is close to one then U_i includes most of the surnames in V.

Therefore, it has been decided that in the event of surname $u \in V_i$ having a smaller similarity to each of the other surnames $v \in V_i$ than a predetermined cut-off value, l_1, then such a surname is considered an *outlier*. The "covering" circle of $V_i := V_i - \{u\}$, is then determined instead, i.e. u is excluded from V_i.

b. For some *ideal class*, V_k, it may happen that each surname in V_k has such a small similarity to all the other surnames in V_k, that the radius is close to one. Thus, all the surnames of V_k can be considered as outliers.

This case has been handled as follows: whenever the radius $r_k > (1 - l_2)$, a predetermined cut-off value, then all the surnames in V_k, excluding the one with the highest frequency of occurrence are taken as outliers. The result is a circle with a very small radius, ϵ say, containing a group consisting of one surname. (To simplify further discussion such a circle will be referred to as a "circle with a zero radius".)

c. For some ideal classes more than one surname may be a candidate for the centre of the circle (e.g. a class containing only two elements).

Let K_m be the set of candidates for the centre of V_m.

Let $X = \max_{v \in K_m}(\min_{w \in V - U_m} d(v, w))$.

The element $v_y \in K_m$, such that $\min_{w \in V - U_m} d(v_y, w) = X$, is chosen as the centre.

After the sets U_i's have been determined, each surname $v \in V$, is an element of one and only one of the following sets:

W_U : The set of outliers.

W_D : The set of surnames appearing in more than one circle, U_i, which do not belong to W_U.

W_G : The set of surnames appearing in only one circle which do not belong to W_U.

W_U, W_D and W_G form a partition of V.

In order to cater for the outliers the success norms of an algorithm as given in equations 1 and 2, are adapted as follows:

$$\begin{aligned} S_3(U) &= \frac{100}{n}(|V| - |W_D| - |W_U|) \\ &= \frac{100}{n}|W_G| \end{aligned} \qquad (3)$$

$$S_4(U) = \frac{100}{N}\left(\sum_{v \in W_G} \rho(v)\right) \quad (4)$$

Statistics determined

The adjustments made may impact the functionality of the success norms. To guard against this and to gain a better understanding of the algorithms and the effects of the parameters l_1 and l_2, the following statistics have been determined.

Let s_v be the number of circles containing the surname, v. Note that the radius of a circle is zero, not only under the circumstances previously explained, but also when an ideal class contains only one element. Let:

$$G = \{U_i : r_i > 0 \text{ and } U_i \in U\} \quad (5)$$
$$I_G = \{i : U_i \in G\} \quad (6)$$

1. The percentage of surnames, pv_u, considered as outliers:
$$pv_u = \frac{|W_U|}{|V|} \times 100 \quad (7)$$
2. The average number of *foreign* circles [1] in which an outlier occur:
$$gs_u = \frac{\sum_{v \in W_U} s_v}{|W_U|}$$
Note that if a surname, $v_k \in V_i$, has been taken as an outlier then $v_k \notin U_i$.
3. The percentage of surnames, excluding outliers, appearing in more than one circle:
$$pv_d = \frac{|W_D|}{|V - W_U|} \times 100 \quad (8)$$
4. The average number of circles in which each of the elements of W_D appears:
$$gs_d = \frac{\sum_{v \in W_D} s_v}{|W_D|} \quad (9)$$
Note that $gs_d > 1$.
5. To get an indication of how many of these surnames appear in 2, 3, 4, ... circles, a frequency distribution has been determined. The frequencies are expressed as percentages and denoted by pv_{di} with $i = 2, 3, 4, \ldots$
6. Percentage of ideal classes left out pg_u. See the practical problem (b) mentioned above.
7. The average number of surnames used to determine a circle:
$$gv_s = \frac{\sum_{i \in I_G} |V_i - W_U|}{|G|} \quad (10)$$
8. The average number of foreign surnames in each of the elements of G, gn_s (excluding outliers).
$$T_G = \sum_{U_i \in G} |U_i \cap (V - V_i) - W_U| \quad (11)$$
$$gn_s = \frac{T_G}{|G|} \quad (12)$$
9. The average number of outliers in each element of G:
$$T_2 = \sum_{U_i \in G} |U_i \cap W_U| \quad (13)$$
$$gu_s = \frac{T_2}{|G|} \quad (14)$$
10. The average number of elements of foreign groups in each element of G, gg_s.
$$X_i = \{V_j : i \neq j, U_i \cap V_j \neq \phi\}$$
$$T_3 = \sum_{i \in G} |X_i| \quad (15)$$
$$gg_s = \frac{T_3}{|G|} \quad (16)$$

[1] A surname appears in a *foreign* circle, if it lies in the circle of another *ideal class*.

3 Algorithms and evaluation

For each algorithm the results have been determined for three different combinations of values for (l_1; l_2), viz (0,1; 0,1), (0,2; 0,1) and (0,2; 0,2). A detailed discussion of the evaluation is given in [4]. Here the main results are summarised.

Position-independent algorithms

Findler and Van Leeuwen [5] investigated a class of similarity norms and proposed the following:

$$G_1(u,v) = \frac{T_1(u,v)}{N_1(u,v)} \text{ where}$$

$$T_1(u,v) = \sum_{\alpha \in (u^+ \cap v^+)} \min\{p(u:\alpha), p(v:\alpha)\} \cdot |\alpha|$$

$$N_1(u,v) = \sum_{\alpha \in (u^+ \cup v^+)} \max\{p(u:\alpha), p(v:\alpha)\} \cdot |\alpha|$$

and

$$G_2(u,v) = \frac{T_2(u,v)}{N_2(u,v)} \text{ where}$$

$$T_2(u,v) = T_1(u,v)$$

$$N_2(u,v) = \left[\sum_{\alpha \in u^+} p(u:\alpha) \cdot |\alpha|\right]^{1/2} \cdot \left[\sum_{\alpha \in v^+} p(v:\alpha) \cdot |\alpha|\right]^{1/2}$$

Where

s^+ = $\{w : w$ is a substring of $s\}$,
$p(s : \alpha)$ is the number of times that the substring α appears in the string s, and
$|\beta|$ is the length of substring β.

E.g. consider the surnames u ="clerq" and v = "klerk":

u^+ = {c, l, e, r, q, cl, le, er, rq, cle, ler, erq, cler, lerq, clerq }
v^+ = {k, l, e, r, kl, le, er, rk, kle, ler, erk, kler, lerk, klerk }
$u^+ \cap v^+$ = {l, e, r, le, er, ler }
$u^+ \cup v^+$ = {c, l, e, r, q, cl, le, er, rq, cle, ler, erq, cler, lerq, clerq, k, kl, rk, kle, erk, kler, lerk, klerk }
$T_1(u, v)$ = $T_2(u, v) = 10$
$N_1(u, v)$ = 60 (Note that "k" appears twice in u).
$G_1(u, v)$ = $\frac{10}{60} = 0{,}17$
$N_2(u, v)$ = $35^{1/2} \cdot 35^{1/2} = 35$
$G_2(u, v)$ = $\frac{10}{35} = 0{,}29$

Table 2. Results for Methods 1 and 2

	Method					
	1a	1b	1c	2a	2b	2c
l_1	0,1	0,2	0,2	0,1	0,2	0,2
l_2	0,1	0,1	0,2	0,1	0,1	0,2
pv_u	16,4	24,0	28,7	5,6	9,2	18,5
gs_u	2,00	0,75	0,34	11,1	5,76	1,74
pg_u	6,60	13,9	15,7	1,78	4,86	7,84
pv_d	55,4	23,5	8,5	95,5	85,8	52,2
gs_d	3,21	2,30	2,08	8,87	5,22	2,90
pv_{d2}	45,7	76,0	92,6	7,40	20,0	53,2
pv_{d3}	24,0	19,4	7,14	8,99	18,3	24,1
S_3	37,3	58,2	65,2	4,28	12,9	39,0
S_4	33,4	61,4	80,7	2,34	9,45	35,2
ps_{r0}	73,3	80,6	82,4	68,5	71,6	74,5
gv_s	2,73	2,71	2,41	3,07	3,07	2,68
gn_s	8,16	2,83	0,90	41,9	22,4	7,03
gu_s	1,29	0,62	0,22	2,23	1,92	1,33
gg_s	6,72	2,30	0,81	31,7	17,7	5,87

The results for G_1 and G_2 for the three combinations of values for l_1 and l_2 are given in Table 2 under the columns Method 1a–c and 2a–c respectively.
Although Method 1 seems to be good, note that

a) too many surnames have been taken as outliers (cf. pv_u);
b) for too many *ideal classes* a too small similarity between the elements of the class has been determined (cf. pg_u).

The main drawback of Method 2 is the large values for gn_s and gg_s.

In both cases substrings of all lengths are used which require an unacceptably large number of processing steps to determine the similarity. The main objection against these methods is that common substrings are counted independently of their relative position.

Position-dependent algorithms

Whenever a common substring appears in two strings, but the position where it appears in the two strings differs too much, then it is not used in the calculation of $T(u, v)$. Ito [7] adapted Findler's similarity measure, G_1, as follows: If

k is the maximum length of the substrings used,
s is the maximum number of places by which the starting position of two common substrings may differ, and
$u^+_{1,k}$ = $\{\alpha(r) : \alpha$ is a substring of u starting at position r, and $1 \leq |\alpha(r)| \leq k\}$,

Ito's measure $T_3(u, v)$, is determined as follows:

1. Let $A = \phi$.
2. Let $B = v^+_{1,k}$.
3. For each substring $\alpha_u(r_u) \in u^+_{1,k}$, in increasing order of its position:
   (a) Let $C = \{\alpha_v(r_v) : \alpha_v(r_v) \in B$, $\alpha_v = \alpha_u$ and $r_v \in [r_u - s, r_u + s]\}$.
   (b) If $|C| = 1$
   Let $A = A \cup \{\alpha_u(r_u)\}$ and $B = B - C$
   (c) If $|C| > 1$
   Let $\alpha'_v(r'_v)$ be the substring such that $\alpha'_v \in C$ and $r'_g = \min_{\alpha_v(r_v) \in C} r_v$.
   Let $A = A \cup \{\alpha_u(r_u)\}$ and $B = B - \{\alpha'_v(r'_v)\}$
4. $T_3(u, v) = \sum_{\alpha(r) \in A} |\alpha(r)|$

Determine $N_3(u, v)$ as follows:

1. Let $D = u^+_{1,k} \cup B$ (B as after step 3.)
2. Now $N_3(u, v) = \sum_{\alpha(r) \in D} |\alpha(r)|$

Thus, $G_3(u, v) = \dfrac{T_3(u, v)}{N_3(u, v)}$

For the surnames u ="clerq" and v = "klerk", $G_3(u, v)$ will be determined as follows for $s = 1$ and $k = 2$:

$u^+_{1,2}$ = { c (1), l (2), e (3), r (4), q (5), cl (1), le (2), er (3), rq (4)}
$v^+_{1,2}$ = { k (1), l (2), e (3), r (4), k (5), kl (1), le (2), er (3), rk (4)}

After step 3 of Ito's algorithm it follows:

A = { l (2), e (3), r (4), le (2), er (3)}
B = { k (1), k (5), kl (1), rk (4)}
$T_3(u, v)$ = 7
$N_3(u, v)$ = 19

Thus, G_3("clerq","klerk") = $\frac{7}{19} = 0{,}37$.

An evalution has been made using five different values of the parameters k and s in the range 1 to 3, and the combinations of l_1 and l_2 mentioned earlier. This report is restricted to the best two combinations which are called Method 3.1 and 3.2. The statistical results are given in Table 3. Method 3.1 yields the smallest value for pv_u and pg_u. Although Method 3.2 also yields small values for these statistics, they are twice as large as those for Method 3.1. Method 3.2 shows an improvement with respect to S_3, S_4 and pv_d, but the sum $pv_{d2} + pv_{d3}$, i.e. the percentage of elements of W_D appearing in two or less foreign circles is relatively large, viz 35,9; 59,6 and 82,3. Although more

surnames are classified as outliers (cf. pv_u), the values for pg_u are not that much larger. The larger values for S_3 and smaller values for gn_s and gg_s indicate that Method 3.2 resulted in a better grouping of the surnames.

Table 3. Results for Methods 3.1 and 3.2

	Method 3					
	1a	1b	1c	2a	2b	2c
l_1	0,1	0,2	0,2	0,1	0,2	0,2
l_2	0,1	0,1	0,2	0,1	0,1	0,2
s	1	1	1	1	1	1
k	2	2	2	3	3	3
pv_u	3,08	4,19	7,20	6,10	8,72	14,5
gs_u	9,83	5,43	3,24	5,37	2,92	1,42
pg_u	0,93	1,71	2,61	2,22	4,08	5,89
pv_d	98,2	91,7	75,6	88,4	70,5	47,4
gs_d	7,83	5,13	3,93	4,99	3,60	2,72
pv_{d2}	4,21	15,6	29,0	17,8	34,7	57,9
pv_{d3}	7,64	17,6	23,9	18,1	24,9	24,4
S_3	1,74	7,96	22,6	10,9	26,9	45,0
S_4	1,09	5,55	18,4	8,61	23,9	40,8
ps_{r0}	67,6	68,4	69,3	68,9	70,8	72,6
gv_s	3,15	3,15	3,04	3,07	3,05	2,81
gn_s	36,3	21,0	12,7	20,0	11,0	5,30
gu_s	1,05	0,75	0,86	1,22	0,92	0,78
gg_s	26,8	15,8	9,94	15,2	8,44	4,28

Other algorithms

Sidorov [8] investigated algorithms to determine the similarity between words in order to correct typing errors. His requirements were small memory usage and speed. Common substrings were only used if they appeared in the same order. This method proved to be unsuitable for solving the primary problem and will not be discussed. For details see [4].

Some algorithms are based on n-grams, an *n-gram* being a substring of length n. The most common values for n are two and three, in which case the n-grams are called di- and tri-grams respectively. Freund, Angell and Willet [6] and [1] investigated similarity measures, the so-called "Dice" and "Overlap" coefficients. The author's evaluation reported in [4] showed that tri-grams yielded better results for the test dataset. All the statistics however, indicate that these methods are not suitable for solving the primary problem.

In conclusion, then, Method 3.2 developed by Ito – although not ideal – appears to be the best of the various methods considered in this study.

References

1. R C Angell, G E Freund, and P Willet. 'Automatic spelling correction using a trigram similarity measure'. *Information Processing and Management*, **19**(4):255–261, (1983).
2. G d V De Kock. 'Die meting van sukses van naampassingsalgoritmes in 'n genealogiese databasis'. *Q.I.*, **6**(3):119–122, (1988).
3. G d V De Kock and C Du Plessis. 'Die empiriese evaluering van variasies van woordpassingsalgoritmes vir die bepaling van ekwivalente vanne in 'n genealogiese databasis'. *SART*, **10**:48–53, (1993).
4. C Du Plessis. Woordpassingalgorimes vir die bepaling van gelyksoortige vanne in 'n genealogiese inligtingstelsel. M.Sc. Verhandeling, Departement Rekenaarwetenskap, U.P.E., Februarie 1991.
5. N V Findler and J Van Leeuwen. 'A family of similarity measures between two strings'. *IEEE Trans on Pattern Analysis and Machine Intelligence*, **1**(1):116–118, (1979).
6. G E Freund and P Willet. 'Online identification of word variants and arbitrary truncation searching using a string similarity measure'. *Information Technology : Research and Development*, **1**(1):177–187, (1982).
7. T Ito and M Kizawa. 'Hierarchical file organization and its application to similar-string matching'. *ACM Transactions on Database Systems*, **8**(3):410–433, (1983).
8. A A Sidorov. 'Analysis of word similarity in spelling correction systems'. *Prog. Comput. Software*, **5**:274–277, (1979).

Notes for Contributors

The prime purpose of the journal is to publish original research papers in the fields of Computer Science and Information Systems, as well as shorter technical research papers. However, non-refereed review and exploratory articles of interest to the journal's readers will be considered for publication under sections marked as Communications or Viewpoints. While English is the preferred language of the journal, papers in Afrikaans will also be accepted. Typed manuscripts for review should be submitted in triplicate to the editor.

Form of Manuscript

Manuscripts for *review* should be prepared according to the following guidelines.

- Use wide margins and $1\frac{1}{2}$ or double spacing.
- The first page should include:
  - title (as brief as possible);
  - author's initials and surname;
  - author's affiliation and address;
  - an abstract of less than 200 words;
  - an appropriate keyword list;
  - a list of relevant Computing Review Categories.
- Tables and figures should be numbered and titled. Figures should be submitted as original line drawings/printouts, and not photocopies.
- References should be listed at the end of the text in alphabetic order of the (first) author's surname, and should be cited in the text in square brackets [1–3]. References should take the form shown at the end of these notes.

Manuscripts accepted for publication should comply with the above guidelines (except for the spacing requirements), and may be provided in one of the following formats (listed in order of preference):

1. As (a) LaTeX file(s), either on a diskette, or via e-mail/ftp – a LaTeX style file is available from the production editor;
2. As an ASCII file accompanied by a hard-copy showing formatting intentions:
   - Tables and figures should be on separate sheets of paper, clearly numbered on the back and ready for cutting and pasting. Figure titles should appear in the text where the figures are to be placed.
   - Mathematical and other symbols may be either handwritten or typed. Greek letters and unusual symbols should be identified in the margin, if they are not clear in the text.

   Further instructions on how to reduce page charges can be obtained from the production editor.
3. In camera-ready format – a detailed page specification is available from the production editor;
4. In a typed form, suitable for scanning.

Charges

Charges per final page will be levied on papers accepted for publication. They will be scaled to reflect scanning, typesetting, reproduction and other costs. Currently, the minimum rate is R30-00 per final page for LaTeX or camera-ready contributions and the maximum is R120-00 per page for contributions in typed format (charges include VAT).

These charges may be waived upon request of the author and at the discretion of the editor.

Proofs

Proofs of accepted papers in categories 2 and 4 above will be sent to the author to ensure that typesetting is correct, and not for addition of new material or major amendments to the text. Corrected proofs should be returned to the production editor within three days.

Note that, in the case of camera-ready submissions, it is the author's responsibility to ensure that such submissions are error-free. However, the editor may recommend minor typesetting changes to be made before publication.

Letters and Communications

Letters to the editor are welcomed. They should be signed, and should be limited to less than about 500 words.

Announcements and communications of interest to the readership will be considered for publication in a separate section of the journal. Communications may also reflect minor research contributions. However, such communications will not be refereed and will not be deemed as fully-fledged publications for state subsidy purposes.

Book reviews

Contributions in this regard will be welcomed. Views and opinions expressed in such reviews should, however, be regarded as those of the reviewer alone.

Advertisement

Placement of advertisements at R1000-00 per full page per issue and R500-00 per half page per issue will be considered. These charges exclude specialized production costs which will be borne by the advertiser. Enquiries should be directed to the editor.

References

1. E Ashcroft and Z Manna. 'The translation of 'goto' programs to 'while' programs'. In *Proceedings of IFIP Congress 71*, pp. 250–255, Amsterdam, (1972). North-Holland.
2. C Bohm and G Jacopini. 'Flow diagrams, turing machines and languages with only two formation rules'. *Communications of the ACM*, **9**:366–371, (1966).
3. S Ginsburg. *Mathematical theory of context free languages*. McGraw Hill, New York, 1966.

South African Computer Journal

Number 11, May 1994
ISSN 1015-7999

Suid-Afrikaanse Rekenaar-tydskrif

Nommer 11, Mei 1994
ISSN 1015-7999

Contents