ISSN 0254-2757

QI QUÆSTIONES INFORMATICÆ

Volume 6 • Number 3 November 1988

T McDonald	**A Proposed Computer Network for Researchers**	**95**
T H C Smith	**Finding a Cheap Matching**	**100**
P J S Bruwer	**Ranking Information System Problems in a User Environment**	**104**
S W Postma N C K Phillips	**The Parallel Conditional**	**109**
D G Kourie R J van den Heever	**Experiences in CSP Trace Generation**	**113**
G de V de Kock	**Die Meting van Sukses van Naampassingsalgoritmes in 'n Genealogiese Databasis**	**119**
R Short	**Learning the First Step in Requirements Specification**	**123**
E C Anderssen S H von Solms	**Frame Clipping of Polygons**	**129**

The official journal of the Computer Society of South Africa and of the South African Institute of Computer Scientists

Die amptelike vaktydskrif van die Rekenaarvereniging van Suid-Afrika en van die Suid-Afrikaanse Instituut van Rekenaarwetenskaplikes

QUÆSTIONES INFORMATICÆ

The official journal of the Computer Society of South Africa and of the South African Institute of Computer Scientists

Die amptelike vaktydskrif van die Rekenaarvereniging van Suid-Afrika en van die Suid-Afrikaanse Instituut van Rekenaarwetenskaplikes

Editor

Professor J M Bishop
Department of Computer Science
University of the Witwatersrand
Johannesburg
Wits
2050

Editorial Advisory Board

Professor D W Barron
Department of Mathematics
The University
Southampton S09 5NH
UNITED KINGDOM

Professor G Wiechers
77 Christine Road
Lynwood Glen
Pretoria
0081

Professor K MacGregor
Department of Computer Science
University of Cape Town
Private Bag
Rondebosch
7700

Professor H J Messerschmidt
Die Universiteit van die Oranje-Vrystaat
Bloemfontein
9301

Dr P C Pirow
Graduate School of Business Admin.
University of the Witwatersrand
P O Box 31170
Braamfontein
2017

Professor S H van Solms
Departement van Rekenaarwetenskap
Randse Afrikaanse Universiteit
Auckland Park
Johannesburg
2001

Professor M H Williams
Department of Computer Science
Herriot-Watt University
Edinburgh
Scotland

Production

Mr Q H Gee
Department of Computer Science
University of the Witwatersrand
Johannesburg
Wits
2050

Subscriptions

The annual subscription is

	SA	US	UK
Individuals	R20	$7	£5
Institutions	R30	$14	£10

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

Quæstiones Informaticæ is prepared by the Computer Science Department of the University of the Witwatersrand and printed by Printed Matter, for the Computer Society of South Africa and the South African Institute of Computer Scientists.

Die Meting van Sukses van Naampassingsalgoritmes in 'n Genealogiese Databasis

G de V de Kock
Departement Rekenaarwetenskap, Universiteit Port Elizabeth, Posbus 1600, Port Elizabeth, 6000

Abstract

Norms to measure the success of surname matching algorithms for use in a South African Genealogical Database are proposed. Surnames in the database can be grouped in equivalence classes.
These algorithms are taken from approximate word or string matching algorithms based on equivalence or similarity relations as proposed in the literature.
Keywords: genealogical database, word matching, name matching

Ontvang Augustus 1988, Aanvaar Oktober 1988

1 Inleiding

Die Departement Rekenaarwetenskap het oor die afgelope jare 'n Genealogiese Inligtingstelsel (genoem UPEGIS), ontwikkel. Met behulp van navorsingstoekennings van die R.G.N. en U.P.E. is genealogiese navorsing gedoen en die resultate in die stelsel ingevoer. Gevolglik bevat hierdie inligtingstelsel 'n genealogiese databasis wat tans reeds uit die besonderhede van ongeveer 60 000 persone bestaan.

Een van die probleme wat dikwels opduik, is om te bepaal of 'n gegewe persoon waarvan soms net die van en 'n noemnaam bekend is, reeds in die databasis is en of hy nog ingevoeg moet word. Die eerste stap in hierdie besluit is om te bepaal welke vanne (of voorname) in die databasis ooreenkom (soortgelyk is aan) die gegewe.

Hierdie probleem kom natuurlik voor in verskeie stelsels soos byvoorbeeld plekbesprekingsstelsels en spellingkorrigeerders. Die Suid-Afrikaanse vanne, veral die variasies en spellings daarvan, is uniek.

Ons het byvoorbeeld van Duitse oorsprong die van, *Kürz*, wat vandag bestaan in die volgende variasies : Coertz, Coorts, Koorts, Koort, Coertse, Koortsen en Coertsen, om maar net 'n paar van die spelvariasies te noem. Franse en Duitse vanne se uitspraak en spelling is byvoorbeeld deur Hollandse, Skotse en Engelse amptenare en predikante "gewysig". Ons het byvoorbeeld die handgeskrewe inskrywing, *"Faroei"* , in die doopregister wat eers verkeerdelik as "Fourie" geneem is, en heelwat later as 'n verbuiging van "Van Rooyen", geidentifiseer is. Later is vanne nog verder deur die volksmond "verbuig". Die verbuigings is vererger deurdat die geletterdheidsvlak van byvoorbeeld die grensboere ens. relatief laag was.

Die versameling van vanne reeds in die databasis, kan in groepe of klasse verdeel word. Al die vanne in 'n bepaalde klas het byvoorbeeld uit 'n gemeenskaplike oorsprong ontstaan of is onskeidbaar van die res in die klas, en kan as *ekwivalent* beskou word. Hierdie groepering in onderling uitsluitende klasse stel 'n ekwivalens partisie van die versameling vanne daar.

Met die huidige stand van die databasis het ons ongeveer 6 000 vanne en sowat 4 000 verskillende klasse. Onder die 6 000 tel ons slegs daardie spelvariasies wat as vanne nog bestaan of bestaan het. Opsigtelike spelfoute wat in argiewe voorkom, word nie as 'n nuwe spelvariasies geneem nie.

Die primêre probleem is die volgende : gegee 'n van, bepaal

- die klas waarin dit voorkom,
- al die vanne waaraan dit soortgelyk (ekwivalent) is in volgorde van een of ander kriterium.

'n Wiskundige formulering van hierdie primêre probleem volg :
Gegee 'n versameling van vanne in die databasis,

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

en ekwivalensklasse of partisie, P, die sogenaamde *ideale partisie* :

$$P = \{V_1, V_2, \ldots, V_p\} \quad \text{waar}$$

$$V = \bigcup_{i=1}^{p} V_i \text{ en } V_i \cap V_j = \phi \ \forall\, i \neq j.$$

Gegee enige van, v ,

1. As $v \in V$ bepaal $i \ni v \in V_i$,
2. As $v \notin V$ bepaal

(a) in welke ekwivalensklas V_i, behoort v te val,

(b) wat is die "naaste" vanne of ekwivalensklasse aan v.

In hierdie referaat bespreek ons nie die individuele algoritmes nie, maar poog om norme daar te stel wat gebruik kan word om die sukses en die bruikbaarheid van die algoritmes spesifiek vir die genealogiese databasis, te evalueer. Volledige bewyse vir al die stelling word gegee in [5].

2 Ekwivalensalgoritmes

Daar bestaan verskeie algoritmes wat enige van of woord onmiddelik in 'n ekwivalensklas plaas. Ons noem die metodes wat enige woord reduseer na 'n string of kode van 'n vasgestelde aantal karakters, byvoorbeeld die "Soundex"-kode van Odell en Russel, sien [6], en verder [4] en [3], om 'n paar te noem. Alle vanne of stringe wat dieselfde kode het, is in dieselfde klas. Dit defineer duidelik 'n partisie van V.

Die primêre probleem soos hierbo gestel, word grotendeels maklik deur hierdie algoritmes opgelos. Daar word net bepaal in welke klas die gegewe van val, en hiervoor bestaan daar effektiewe algoritmes. Die kwessie van naaste vanne of ander ekwivalensklasse kan slegs bepaal word as daar een of ander afstandsmaat gedefineer is.

Die sekondêre probleem is nou om uitspraak te gee oor hoe goed so 'n partisie is, en wel spesifiek vir die gebruik in die genealogiese databasis. Ons kan dit slegs meet aan die *ideale partisie*, maar welke norm moet gebruik word? Dit is veral hierdie aspek wat in hierdie referaat aangespreek word.

Die *sukses* van 'n partisie,

$$Q = \{Q_1, Q_2, \ldots, Q_q\} \text{ , van } V$$

relatief tot die partisie P, word formeel gedefineer as :

$$N(P,Q) = \max_f \sum_{A \in P} | A \cap f(A) |$$

waar $f : P \rightarrow Q \cup \{\phi\}$ en f sodanig is dat

$$f(V_i) \cap f(V_j) \neq \phi \Rightarrow i = j,$$

m.a.w. geen V_i en V_j, $i \neq j$, beeld af op dieselfde element (klas) van Q nie.
Sonder enige verlies van algemeenheid kan ons die verdere beperking op f plaas dat :

$$V_i \cap f(V_i) = \phi \implies f(V_i) = \phi$$

Om die maksimum afbeelding te vind, kom neer op die oplossing van 'n gewone toewysingsprobleem waarvoor daar baie effektiewe algoritmes bestaan.

Die volgende stelling geld :

Stelling 2.1
Vir enige twee partisies R en S, van V geld :

1. $0 \leq N(R,S) \leq |V|$
2. $N(R,S) = N(S,R)$
3. $N(R,S) = |V| \iff R = S$

Soortgelyk kan ons 'n afstandsmaat, $D(R,S)$, tussen twee partisies R en S, van V soos volg defineer :

$$D(R,S) = \min_f \sum_{A \in R} | A - f(A) |$$

waar $f : P \rightarrow Q \cup \{\phi\}$ en

f sodanig is dat $f(V_i) \cap f(V_j) \neq \phi \Rightarrow i = j$.

Die volgende stelling kan nou bewys word :

Stelling 2.2

1. $D(R,S) = |V| - N(R,S)$
2. $D(R,S) \geq 0$
3. $D(R,S) = 0 \iff R = S$
4. $D(R,S) = D(S,R)$
5. $D(R,S) \leq D(R,T) + D(T,S)$ *(Driehoeksongelykheid)*
6. $D(R,S)$ *is 'n metriek*

3 Gelyksoortigheidsnorme

Daar is verskei algoritmes wat 'n afstand of metriek tussen woorde (name) defineer. Bickel, [2], defineer byvoorbeeld 'n afstand wat gebaseer is op 'n geweegde som van gemeenskaplike letters. 'n Geweegde som van gemeenskaplike lettergrepe word ook gebruik in [1] en [8]. Wagner en medeouteurs, [10] en [9], baseer die afstand op 'n geweegde som van die aantal operasies (transformasies) wat nodig is om die een woord na die ander te transformeer.

Die afstand is gewoonlik 'n funksie,

$$d : A^* \times A^* \rightarrow E,$$

waar E 'n interval op die reële of integer getallelyn is, en waar A^* die versameling van alle moontlike woorde is, d.w.s.

$$A^* \supset V.$$

Hierdie afstandsmaat voldoen normaalweg aan die volgende vereistes van 'n metriek :

1. $d(v,w) \geq 0 \ \forall\, v,w \in A^*$,
2. $d(v,w) = d(w,v) \ \forall\, v,w \in A^*$,
3. $d(v,w) = 0 \iff v = w$,
4. $d(u,w) \leq d(u,v) + d(v,w)$ $\forall\, u,v,w \in A^*$.

Twee vanne (of woorde) word as 'naby aan mekaar' of *gelyksoortig*, beskou as hul afstand van mekaar kleiner is as 'n voorafgestelde drumpelwaarde, β.
Hierdie benadering kan beskou word as 'n gelyksoortigheidsrelasie, $\overset{\beta}{\sim}$.

$$\forall v,w \in A^* \quad v \overset{\beta}{\sim} w \iff d(v,w) \leq \beta$$

Defineer nou die afstand tussen V_i en die naaste element uit die ander V_j's as :

$$\delta_i = \min_{w \in V - V_i, v \in V_i} d(v,w) \qquad \text{vir } i = 1 \text{ tot } p$$

Laat

$$\delta = \min_i \delta_i$$

Defineer nou $\forall\, V_i \in V$

$$U_i(\beta) = \{a : a \in V,\ d(a,b) \leq \beta,\ b \in V_i\}$$

$$\text{en } Q_\beta = \{U_1(\beta), U_2(\beta), \ldots, U_p(\beta)\}$$

$U_i(\beta)$ bevat dus al die vanne in V wat gelyksoortig is aan V_i.
Die bewerings van die volgende stelling volg direk of kan maklik bewys word :

Stelling 3.1

1. $U_i(\beta) \supseteq V_i$,
2. $\sum_{i=1}^{p} U_i(\beta) = V$
3. $\delta_i > \beta \implies U_i(\beta) = V_i$.
4. $\delta > \beta \implies Q_\beta = P$, *en verdeel die metode met die drumpelwaarde β, die versameling V in die ideale partisie, d.w.s.* $D(P, Q_\beta) = 0$,
5. *As* $U_i \cap U_j = \phi \ \forall\ U_i \neq U_j$, *dan vorm* Q_β *'n partisie van* V, *en kan behandel word soos in afdeling 2 beskryf.*

In die meeste gevalle is Q_β egter nie 'n partisie van V nie, maar oordek V wel klasgewys. Ons kan so 'n klasgewyse oordekking ook op ander maniere defineer.
Laat

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

dan defineer ons die *straal* van V_i soos volg :

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

$$\text{d.w.s } r_i = \min_{v \in V_i} R_i(v)$$

Laat

$$\rho = \max_i r_i$$

Kies die *middelpunt*, c_i, van V_i, sodanig dat

1. $r_i = R_i(c_i)$ met $c_i \in V_i$ en
2. $|\{w : w \in V,\ d(c_i,w) \leq r_i\} \cap (V - V_i)|$, 'n minimum is.

Let op dat

$$d(c_i, v) \leq r_i \quad \forall\, v \in V_i.$$

Nou defineer ons die "sirkel" met middelpunt, c_i, en straal, $r_i + \gamma$, met $\gamma \geq 0$, soos volg :

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

Dit volg direk dat

$$C_\gamma = \{C_1(\gamma), C_2(\gamma), \ldots, C_p(\gamma)\},$$

ook 'n klasgewyse oordekking van V is.
Die voordeel van hierdie oordekking is dat γ en die versameling pare bestaande uit middelpunte en strale, $\{(c_i, r_i) : i = 1 \text{ tot } p\}$, dit ten volle beskryf en as 'n *kanoniese vorm* van die oordekking beskou kan word.

Stelling 3.2

1. $\gamma < \beta \implies C_i(\gamma) \subseteq C_i(\beta)$
2. $V_i \subseteq U_i(\beta) \subseteq C_i(\beta) \quad \forall\, i = 1 \text{ tot } p$
3. $\delta_i > r_i \implies C_i(0) = V_i$
4. $\delta_i > r_i \ \forall\, i = 1 \text{ tot } p \implies C_0 = P.$
5. $\delta > \rho \implies C_0 = P.$

Die primêre probleem word nou soos volg opgelos. Vir 'n gegewe $w \in V$ of A^*, bepaal die versameling indekse

$$I_\beta(w) = \{i : d(c_i, w) \leq r_i + \beta\}$$

Let op dat alle elemente uit V wat gelyksoortig is aan w moet in een van die V_i met

naaste vanne aan v kan nou met die afstandsmaat in hierdie ekwivalensklasse bepaal word.

Laat ons nou in die algemeen so 'n *klasgewyse oordekking*, of *K-oordekking* van V soos volg defineer :

$$Q = \{U_1, U_2, \ldots, U_p\}, \text{ met}$$

$$U_i \supseteq V_i, \quad \forall V_i \in V,$$

$$\text{dan volg } \sum_{i=1}^{p} U_i = V.$$

As maatstaf van hoe bruikbaar die betrokke algoritme wat so 'n K-oordekking defineer, vir ons doeleindes is, defineer ons nou die volgende maatstawwe :

$$\begin{aligned} M_1(Q) &= \textstyle\sum_{i=1}^{p}(\mid V_i - \sum_{\substack{j=1 \\ j\neq i}}^{p} V_i \cap U_j) \\ &= \textstyle\sum_{i=1}^{p}(\mid V_i \mid - \sum_{\substack{j=1 \\ j\neq i}}^{p} \mid V_i \cap U_j \mid) \\ &= \mid V \mid - \textstyle\sum_{\substack{j=1 \\ j\neq i}}^{p} \mid V_i \cap U_j \mid) \end{aligned}$$

$$M_2(Q) = \sum_{i=1}^{p} \frac{|V_i|^2}{|U_i|}$$

Beide hierdie norme kan beskou word as 'n maatstaf van *noukeurigheid* wat algemeen gedefineer word as die verhouding van die aantal bruikbare rekords (woorde) tot die aantal rekords (woorde) ontsluit. Vergelyk [6].
Ons neem ook hiermee nie die kompleksiteit van die betrokke algoritme in ag nie.

Vir albei gevalle volg direk van die definisies :

Stelling 3.3 *Vir* $k = 1$ *en* 2 :

$$0 \leq M_k(Q) \leq \mid V \mid \text{ en}$$

$$M_k(Q) = \mid V \mid \iff Q = P$$

4 Slot

Die maatstawwe hierin voorgestel, word tans in 'n studie gebruik om die beste algoritme vir die oplossing van die primêre probleem te bepaal en te ontwikkel. Ons is van mening dat 'n gelyksoortigheidsmaat waarvoor

$$C_\beta = P,$$

die ideale partisie, die beste oplossing van die primêre probleem sal wees.

References

[1] Angell,R.C. Freund,G.E. and Willet,P., [1983], Automatic spelling correction using a trigram similarity measure. *Inf. Proc. & Management* **19(4)**, *1983, pp. 255–261.*

[2] Bickel,M.A. [1987], Automatic correction to misspelled names : a Fourth-generation approach. *Comms. ACM* **30(3)**, *1987, pp 224–228.*

[3] Blair,C.R. [1960], A Program for Correcting Spelling Errors. *Inf.Control* **3**, *1960, pp. 60–67.*

[4] Davidson,L. [1962], Retrieval of Misspelled Names in an Airline's Passenger Record System. *Comms. ACM* **5(3)**, *1962, pp169–171.*

[5] De Kock, G.de V. [1988] Die meting van sukses van naampassingsalgoritmes in 'n genealogiese databasis. *Dept. Rekenaarwetenskap, U.P.E., Publikasiereeks Nr. 88/04, 1988, pp 11.*

[6] Hall, P.A.V. and Dowling, G.R. [1980] Approximate String Matching. *Comp.Surveys ACM* **12(4)**, *1980, pp. 381–402.*

[7] Ito,T. and Kizawa,M. [1983], Hierarchical File Organization and Its Application to Similar-String Matching. *Trans. on Database Sys.* **8(3)**, *1983, pp. 410–433.*

[8] Owolabi,O. and McGreggor,D.R. [1988], Fast Approximate String Matching. *Software Practice & Experience* **18(4)**, *1988, pp. 387–393.*

[9] Wagner,R.A. and Fischer,M.J. [1974], The String-to-String Correction Problem. *Journal ACM* **21(1)**, *1974, pp 168–173.*

[10] Lowrance,R. and Wagner,R.A. [1975], An Extension of the String-to-String Correction problem. *Journal ACM* **22(2)**, *1975, pp. 177–183.*

NOTES FOR CONTRIBUTORS

The purpose of the journal will be to publish original papers in any field of computing. Papers submitted may be research articles, review articles and exploratory articles of general interest to readers of the journal. The preferred languages of the journal will be the congress languages of IFIP although papers in other languages will not be precluded.

Manuscripts should be submitted in triplicate to:

Professor ~~J M Bishop~~ D.C. Kaire
Department of Computer Science
University of the Witwatersrand
Johannesburg
Wits
2050

Form of manuscript

Manuscripts should be in double-space typing on one side only of sheets of A4 size with wide margins.

The first page should include the article title (which should be brief), the author's name and affiliation and address. Each paper must be accompanied by an abstract less than 200 words which will be printed at the beginning of the paper, together with an appropriate key word list and a list of relevant Computing Review Categories.

Manuscripts may be provided on disc ~~using any Apple Macintosh package or~~ in ASCII format, once that a submitted paper has been accepted.

For authors wishing to provide camera-ready copy, a page specification is freely available on request from the Editor.

Tables and figures

Tables and figures should not be included in the text, although tables and figures should be referred to in the printed text. Tables should be typed on separate sheets and should be numbered consecutively and titled.

Figures should also be supplied on separate sheets, and each should be clearly identified on the back in pencil with the authors name and figure number. Original line drawings (not photocopies) should be submitted and should include all the relevant details. Photographs as illustrations should be avoided if possible. If this cannot be avoided, glossy bromide prints are required.

Charges

Symbols

Mathematical and other symbols may be either handwritten or typewritten. Greek letters and unusual symbols should be identified in the margin. Distinction should be made between capital and lower case letters; between the letter O and zero; between the letter I, the number one and prime; between K and kappa.

References

References should be listed at the end of the manuscript in alphabetic order of the author's name, and cited in the text in square brackets. Journal references should be arranged thus:

[1] E. Ashcroft and Z. Manna, [1972], The Translation of 'GOTO' Programs to 'WHILE' programs, *Proceedings of IFIP Congress 71,* North-Holland, Amsterdam, 250-255.

[2] C. Bohm and G. Jacopini, [1966], Flow Diagrams, Turing Machines and Languages with only Two Formation Rules, *Comm. ACM,* **9**, 366-371.

[3] S. Ginsburg, [1966], *Mathematical Theory of Context-free Languages*, McGraw Hill, New York.

Proofs

Proofs will be sent to the author to ensure that the papers have been correctly typeset and *not* for the addition of new material or major amendment to the texts. Excessive alterations may be disallowed. Corrected proofs must be returned to the production manager within three days to minimise the risk of the author's contribution having to be held over to a later issue.

Only original papers will be accepted, and copyright in published papers will be vested in the publisher.

Letters

A section of "Letters to the Editor" (each limited to about 500 words) will provide a forum for discussion ~~of recent problems~~.