# Αλγόριθμος Braid: Εξόρυξη ροών δεδομένων μέσω ομαδοποίησης λόγω συσχέτισης με καθυστέρηση

Πυργερής Γιώργος
pyrgeris@ceid.upatras.gr

19 Ιουλίου 2013

## Περίληψη

Η εργασία έχει ως σκοπό την παρακολούθηση πολλαπλών αριθμητικών ροών δεδομένων και τον καθορισμό των ζευγών που συσχετίζονται με κάποια καθυστέρηση, καθώς και ο υπολογισμός της τιμής της καθυστέρησης αυτής.

Προτείνεται η χρήση του αλγόριθμου Braid . Πρόκειται για μία μέθοδο υπολογισμού της συσχέτισης με καθυστέρηση (lag correlation), μεταξύ των ροών δεδομένων. Υποστηρίζει ροές δεδομένων με σχεδόν άπειρο μήκος, κλιμακωτά, γρήγορα και με μικρό υπολογιστικό κόστος .

Επίσης παρουσιάζεται μία θεωρητική ανάλυση βασισμένη στο θεώρημα Nyquist σύμφωνα με την οποία, αποδεικνύεται ότι ο Braid υπολογίζει το lag correlation με μικρό και συνήθως μηδενικό σφάλμα. Το μέγιστο σχετικό σφάλμα υπολογίζεται γύρω στο 1%. Η ταχύτητα επεξεργασίας του αλγόριθμου είναι γύρω στις 14.000 φορές πιο γρήγορη από την απλή υλοποίηση που θα μπορούσε κάποιος να εφαρμόσει για την επίλυση του προβλήματος.

## Εισαγωγή

Αρχικό Πρόβλημα: Δοθέντων δύο ακολουθιών που εξελίσσονται παράλληλα στο χρόνο, ίδιου μήκους $n$ , πρέπει να μπορεί να υπολογιστεί ανά πάσα χρονική στιγμή α) αν υπάρχει συσχέτιση μέσω καθυστέρησης (lag correlation), τιμής $l$, μεταξύ τους και β) αν ναι, ποια είναι η τιμή της καθυστέρησης αυτής.

Κύριο Πρόβλημα: Δοθέντων $k$ ακολουθιών που εξελίσσονται παράλληλα στο χρόνο, ίδιου μήκους $n$, πρέπει να μπορεί να υπολογιστεί ανά πάσα χρονική στιγμή α) ποιά ζεύγη έχουν συσχέτιση μέσω καθυστέρησης (lag correlation) τιμής $l$, μεταξύ τους και β) να μπορεί να επιστρέφεται τιμή της καθυστέρησης αυτής.

Διαισθητικά δύο ακολουθίες δεδομένων έχουν συσχέτιση με καθυστέρηση (lag correlation) τιμής $l$, αν δείχουν σχεδόν ίδιες αν η μία καθυστερήσει κατά $l$ χρονικές στιγμές. Αν οι δύο ακολουθίες ήταν στατικές το πρόβλημα θα ήταν τετριμένο. Απλά θα έπρεπε να υπολογιστεί ο συντελεστής συσχέτισης(correlation factor $R(l)$) μέσω της συνάρτησης CCF (cross-correlation function) και να επιστραφεί η τιμή της καθυστέρησης $l$, τη στιγμή που μεγιστοποιείται ο συντελεστής. Αλλά οι δυο ακολουθίες δεδομένων $X, Y$ συνεχώς μεγαλώνουν σε βάθος χρόνου. Χρειαζόμαστε μία μέθοδο με τα εξής χαρακτηριστικά :

1. Επεξεργασία ανα πάσα στιγμή και ταχύτατα. Ο χρόνος επεξεργασίας θα πρέπει να είναι υπό-γραμμικός (στη βέλτιστη περίπτωση ακαριαίος) σε ακολουθίες με μήκος $n$.

2. Ευελιξία. Οι απαιτήσεις μνήμης πρέπει να είναι υπό-γραμμικές σε σχέση με το μήκος $n$ των ακολουθιών.

3. Ακρίβεια. Δεδομένου ότι τα ακριβή αποτελέσματα απαιτούν πάρα πολύ χώρο και χρόνο, χρειαζόμαστε προσεγγίσεις. Τέτοιου είδους προσέγγιση εισάγει ένα μικρό σφάλμα.

Ο αλγόριθμος Braid είναι ο πρώτος αλγόριθμος που καλύπτει όλα τα παραπάνω χαρακτηριστικά.

# Προτεινόμενη Μέθοδος

Μια ροή δεδομένων X είναι μία διακριτή ακολουθία αριθμών $\{x_1, x_2, ..., x_n\}$ όπου $x_n$ είναι η πιο πρόσφατη τιμή. Το μέγεθος $n$ αυξάνεται σε κάθε χρονική στιγμή. Ο ορισμός του συντελεστή συσχέτισης $R(0)$ μεταξύ δύο χρονικών ακολουθιών $X$ και $Y$, ίδιου μήκους $n$ και μηδενικής καθυστέρησης $l = 0$ είναι ευρέως γνωστός σαν συντελεσής του Pearson($p$).

$$p = R(0) = \frac{\sum_t (x_t - \bar{x}) * (y_t - \bar{y})}{\sigma(x) * \sigma(y)}$$

όπου $\bar{x}$, $\bar{y}$ είναι οι μέσοι όροι των ακολουθιών $X$ και $Y$. Για τιμή $l \geq 0$ όπου $l$, η καθυστέρηση, το $R(l)$ μας δίνει το συντελεστή συσχέτισης, όταν η ακολουθία $X$ είναι καθυστερημένη κατά $l$. Το $R(l)$ δίνεται από τον τύπο :

$$R(L) = \frac{\sum_{t=l+1}^{n} (x_t - \bar{x}) * (y_{t-l} - \bar{y})}{\sqrt{\sum_{t=l+1}^{n} (x_t - \bar{x})^2)} * \sqrt{\sum_{t=1}^{n-l} (y_t - \bar{y})^2)}}, \ \ \bar{x} = \frac{1}{n-l} \sum_{t=l+1}^{n} x_t, \ \ \bar{y} = \frac{1}{n-l} \sum_{t=1}^{n-l} y_t$$

Τονίζεται ότι σημαντικές είναι μόνο οι απόλυτες τιμές του $R(l)$.

$$score(l) = \mid R(l) \mid$$

**ΟΡΙΣΜΟΣ - Lag Correlation:** Δύο ακολουθίες $X$, $Y$ έχουν συσχέτιση με καθυστέρηση(Lag Correlation) με τιμή $l$, και συγκεκριμένα το $X$ καθυστερεί το $Y$ κατά $l$, όταν :

1. Η απόλυτη τιμή του $R(l)$ μεταξύ του $x_t$ και του $y_{t-l}$ είναι πάνω από ένα φράγμα $\gamma$, έστω $\gamma = 0.4$ και η τιμή αυτή είναι τοπικό μέγιστο.

2. Και αυτό είναι το πρώτο τέτοιο μέγιστο, αν υπάρχουν και άλλα τοπικά μέγιστα.

Υπάρχει περιορισμός για τη μέγιστη τιμή του $l$ και καθορίζεται από τον τύπο $m = \frac{n}{2}$, δηλαδή η μέγιστη τιμή της καθυστέρησης πρέπει να είναι ίση ή μικρότερη από το μισό της ακολουθίας.