ΠΡΟΓΡΑΜΜΑ ΚΑΙ ΠΕΡΙΛΗΨΕΙΣ

5^{ο} Ελληνικό Συμπόσιο Διαχείρισης Δεδομένων
(ΕΣΔΔ'2006)
Ξενοδοχείο Φιλίππειο
Θεσσαλονίκη - Ελλάδα
7-8 Σεπτεμβρίου 2006

ΤΕΛΙΚΟ ΠΡΟΓΡΑΜΜΑ

7 Σεπτεμβρίου (Πέμπτη) **Αίθουσα Μακεδονία**

8:00 – 8:45: Εγγραφές

8:45 – 9:00: **Έναρξη**: Γ. Μανωλόπουλος (Γενικός Πρόεδρος), Π. Τριανταφύλλου (Πρόεδρος Επιστημονικής Επιτροπής)

9:00 – 10:00: **Προσκεκλημένη ομιλία:** "Energy and the Third Law of Compu-Dynamics: From Mobility to Pervasiveness", Πάνος Χρυσάνθης (U of Pittsburgh)

10:00 – 10:30: Διάλλειμμα - καφές

10:30 – 12:00: **Συνεδρία 1: Data Summaries**

"Συνόψεις Γράφου Πλειάδων για Σχεσιακά Δεδομένα", J. Spiegel, N. Πολυζώτης (U of California, Santa Cruz)

"Γρήγορη Προσεγγιστική Παρακολούθηση Συνόψεων Wavelet σε Ρεύματα Δεδομένων", G. Cormode (Lucent Bell Labs), M. Γαροφαλάκης (Intel Research Berkeley), Δ. Σαχαρίδης (ΕΜΠ)

"Κατανεμημένες Συνόψεις Κατακερματισμού: Αποδοτική Εκτίμηση του Πλήθους Εγγραφών Μεγάλων Πολυσυνόλων σε Δίκτυα Δεδομένων Διαδικτυακής Κλίμακας", Ν. Ντάρμος (Παν. Πατρών), G. Weikum (MPI, Saarbruecken)

12:00 – 13:30: Γεύμα

13:30 – 15:00: **Paper Session 2: Data in Overlays**

"Κατασκευή Δικτύων Επικάλυψης με Γνώση του Φορτίου Ερωτήσεων Βασισμένη στη Χρήση Ιστογραμμάτων", Γ. Κολωνιάρη, Γ. Πετράκης, Ε. Πιτουρά, Τ. Τσότσος (Παν. Ιωαννίνων)

"Επεξεργασία Ερωτήσεων Σύζευξης σε Συστήματα Ομοτίμων", Σ. Ιδρέος, Χ. Τρυφωνόπουλος (Πολυτ. Κρήτης), Μ. Κουμπαράκης (Παν. Αθηνών)

"PastryStrings - Σύστημα Δημοσίευσης/συνδρομής Περιεχομένου σε DHT Δίκτυο Ομοτίμων", Ι. Αικατερινίδης (Παν. Πατρών)

15:00 – 15:30: Διάλλειμμα - καφές

15:30 – 16:00: **Ομιλία ORACLE**: "Grid Computing", Γ. Μπούρμας (Oracle Greece)

16:00 – 17:00: **Paper Session 3: Data Mining**

"LOCUS: Χαλαρή και Βέλτιστη Ταξινόμηση με Απεριόριστη Κλιμάκωση", Κ. Μορφονιός, Γ. Ιωαννίδης (Παν. Αθηνών)

"Το Πλαίσιο MONIC για τον Εντοπισμό Μεταβολών σε Συστάδες", Μ. Σπηλιοπούλου (U of Magdeburg), Ε. Ντούτση, Γ. Θεοδωρίδης (Παν. Πειραιά), R. Schult (U of Magdeburg)

17:00 – 17:30: Διάλλειμμα - καφές

17:30 – 18:15: **Συνεδρία 4: Pot Pouri**

"Online Παρακολούθηση της Κατάστασης Δικτύου σε Περιβάλλον Δικτύων Μικροαισθητήρων", Μ. Χαλκίδη, Β. Καλογεράκη, Δ. Γουνόπουλος, Δ. Παπαδόπουλος, Δ. Ζεϊναλιπούρ-Γιαζτί (U of California, Riverside), Μ. Βλάχος (IBM T.J. Watson)

"Έλεγχος Περιεκτικότητας για Ερωτήσεις Δεντρικών Προτύπων Μερικού Προσδιορισμού", Δ. Θεοδωράτος, Θ. Δαλαμάγκας, P. Placek, Σ. Σουλδάτος, Τ. Σελλής (ΕΜΠ)

" PROTEAS: ένας Αλγόριθμος Εξόρυξης Δεδομένων Βασισμένος σε Πεπερασμένα Αυτόματα για την Εξαγωγή Κανόνων Ταξινόμησης Πρωτεϊνών", Φ. Ψωμόπουλος, Π. Μήτκας (Αριστοτέλειο Παν.)

18:15-19:15: **Προσκεκλημένο Tutorial:** "Top-K Query Processing", Δ. Γουνόπουλος (U of California, Riverside)

19:00 – 21:00: Δεξίωση

8 Σεπτεμβρίου (Παρασκευή) Αίθουσα Μακεδονία

9:00 – 10:30: **Συνεδρία 5: Relational Back-Ends**

"CURE για Κύβους: Κατασκευή Κύβου με Χρήση Σχεσιακών Συστημάτων", Κ. Μορφονιός, Γ. Ιωαννίδης (Παν. Αθηνών)

"Νέες Τεχνικές για Αποδοτική Εκτέλεση XPath Επερωτήσεων", Χ. Γεωργιάδης, Β. Βασάλος, (Οικονομικό Παν. Αθηνών)

"Précis: Ένας Καινούριος Τρόπος Απαντήσεων σε Επερωτήσεις Πάνω σε Σχεσιακά Δεδομένα", Γ. Κουτρίκα, Α. Σιμιτσής, Γ. Ιωαννίδης (Παν. Αθηνών)

10:30 – 11:00 Διάλλειμμα - καφές

11:00 – 12:00: **Συνεδρία 6: Personalization**

"UPR: Κατάταξη Σελίδων Βασισμένη στα Δεδομένα Χρήσης για Εξατομίκευση Δικτυακών Τόπων", Μ. Ειρηνάκη, Μ. Βαζυργιάννης (Οικονομικό Παν. Αθηνών)

"Ερώτησεις Προτίμησης με βάση το Συμφραζόμενο Περιβάλλον: μια Επισκόπηση", Κ. Στεφανίδης, Ε. Πιτουρά, Π. Βασιλειάδης (Παν. Ιωαννίνων)

12:00 – 13:30: Γεύμα

13:30 – 15:00: **Συνεδρία 7: Advanced Queries and View Updates**

"Ένας Συνδυασμός Δένδρων Trie και Αντεστραμμένων Αρχείων για τη Δεικτοδότηση Χαρακτηριστικών Βασισμένων σε Σύνολα", Μ. Τερροβίτης, Σ. Πασσάς (ΕΜΠ), Π. Βασιλειάδης (Παν. Ιωαννίνων), Τ. Σελλής (ΕΜΠ)

"Συνεχής Ελεγχος Κορυφαίων-k Ερωτημάτων σε Συνεχείς Ροές Δεδομένων", Κ. Μουρατίδης, Σ. Μπακίρας, Δ. Παπαδιάς, (Hong Kong U of Science & Technology)

"Επιτυγχάνοντας την Ενημέρωση Όψεων χωρίς Παρενέργειες", Γ. Κωτίδης, D. Srivastava, Γ. Βελεγράκης (AT&T Labs)

15:00 – 16:00 **Συνεδρία 8: Web Data**

"Χρονοδρομολόγηση Ενημερώσεων και Ερωτήσεων με Συμβόλαια Ποιότητας για Βάσεις Δεδομένων του Παγκόσμιου Ιστού ", H. Qu, A. Λαμπρινίδης (U of Pittsburgh)

"Σχετικά με τα Γνωρίσματα των Γράφων των Σχημάτων του Σημασιολογικού Ιστού", Γ. Θεοχάρης, Β. Χριστοφίδης, Γ. Τζίτζικας (Παν. Κρήτης)

16:00 – 16:30: Διάλλειμμα - καφές

16:30 -- 17:15 **Συνεδρία 9: Distributed Data**

"Αποτελεσματικές Προσεγγιστικές Τεχνικές για Υπολογισμό Ερωτήσεων σε Κατανεμημένες Βάσεις Δεδομένων Υψηλής Κλίμακας", Β. Arai, G. Das, Δ. Γουνόπουλος, Β. Καλογεράκη (U of California, Riverside)

"Η Προσέγγιση SOWES για Αναζήτηση στον Παγκόσμιο Ιστό με χρήση Σημασιολογικών Δικτύων Επικάλυψης", K. Noervaag, Χ. Δουλκερίδης, Μ. Βαζυργιάννης (Οικονομικό Παν. Αθηνών)

"Αυτόματη Κατανομή Επερωτήσεων σε Αυτόνομες Κατανεμημένες Βάσεις Δεδομένων", Φ. Πεντάρης, Γ. Ιωαννίδης (Παν. Αθηνών)

17:15 – 17:45 Διάλλειμμα - καφές

17:45 – 19:15: **Στρογγυλή τράπεζα:** "Νέες Ερευνητικές Κατευθύνσεις στη Διαχείριση Δεδομένων", Τ. Σελλής (ΕΜΠ)

ΠΕΡΙΛΗΨΕΙΣ ΟΜΙΛΙΩΝ

Power-Aware Data Management and the Third Law of Compu-Dynamics
Πάνος Χρυσάνθης (University of Pittsburgh)

Abstract: In the last 50 years since the very first computer, we have witnessed many hardware and software evolutions but the shape of computing has always been defined by disruptive technologies. The best known example of such a disruptive technology from the 80s is the PC. The advent of mobile and tiny computing devices is the most recent disruptive technology that has impacted every aspect of life from communication, to business, to education, to health, to entertainment. It has also caused a paradigm shift in designing and developing small-sized computing systems, in general and data management protocols, in particular. In this new paradigm, power or energy consumption has become the new governing law for data management in addition to the two traditional laws of time and space complexities. In this talk we will focus on this new, third law of Compu-Dynamics and discuss algorithms, protocols and techniques for power-aware data management.

Top-K Query Processing
Δημήτρης Γουνόπουλος (U of California, Riverside)

Περίληψη: Τα τελευταία χρόνια υπάρχει μεγάλο ενδιαφέρον στην επιστημονική κοινότητα για την σχεδίαση και υπολοποίηση αποτελεσματικών τεχνικών και αλγόριθμων ανεύρεσης σε σχεσιακές βάσεις δεδομένων, σε κατανεμημένες βάσεις δεδομένων, σε μεγάλα συστήματα αισθητήρων, σε συστήματα αποθήκευσης πολυμέσων. Πιό συγκεκριμένα, σε πολλές νέες εφαρμογές το πρόβλημα είναι να δώσουμε τεχνικές που επιτρέπουν στο χρήστη να ψάξει τα δεδομένα με ad-hoc ερωτήσεις. Είναι σημαντικό, για να είναι η αναζήτηση πρακτική, οι ερωτήσεις αυτές να μπορούν να απαντηθούν γρήγορα. Για παράδειγμα, ο χρήστης μπορεί να ψάχνει, σε πολλές βάσεις δεδομένων ταυτόχρονα, για προιόντα όπως αυτοκίνητα, ηλεκρονικά, εστιατόρια, κτλ, που πληρούν κάποιες προυποθέσεις. Μια βασική μέθοδος για τέτοιου είδους αναζήτηση, είναι η χρήση Top-K ερωτήσεων. Σε αυτό το μοντέλο, η μέθοδος ενεύρεσης επιστρέφει τα K αποτελέσματα που καλύτερα ικανοποιούν τις προυποθέσεις της ερώτησης. Τεχνικές και αλγόριθμοι για την απάντηση Top-K ερωτήσεων είναι το αντικείμενο εκτεταμένης ερευνητικής δουλειάς τα τελευταία χρόνια. Σε αυτή την ομιλία θα εξετάσουμε το γενικό Top-K πρόβλημα, ξεκινώντας από τους βασικούς αλγόριθμους TA και FA, και προχωρόντας σε σημαντικές παραλαγές των αλγοριθμων αυτών που είτε υποθέτουν περιορισμένες δυνατότητες πρόσβασης, ή στοχαστικές προσεγγίσεις. Ενα σημαντικό κομμάτι της ομιλίας θα επικεντρωθεί σε κατανεμημένες τεχνικές για τον υπολογισμό των Top-K αποτελεσμάτων, στην περίπτωση που έχουμε κάθετη κατανομή των δεδομένων (στην περίπτωση αυτή κάθε κόμβος, ή βάση δεδομένων, έχει ένα υποσύνολο των χαρακτηριστικών κάθε αντικειμένου). Θα εξετάσουμε επίσης χρήσεις αυτών των αλγορίθμων σε εφαρμογές στο χώρο του διαδικτύου, σε συστήματα peer-to-peer, και δίκτυα αισθητήρων. Σε αυτή την ομιλία θα επικεντρωθούμε στο να καλύψουμε τις βασικές εξελιξεις στο χώρο, και να προβάλουμε τα νέα προβλήματα στην περιοχή, τα οποία μπορούν να γίνουν αντικείμενο νέας έρευνας.

Συνόψεις Γράφου Πλειάδων για Σχεσιακά Δεδομένα
J. Spiegel, N. Πολυζώτης (U of California, Santa Cruz)

Περίληψη: Η παρούσα εργασία εισάγει τις Συνόψεις Γράφου Πλειάδων, μια νέα ομάδα συνόψεων που επιτρέπει την εκτίμηση, με χαμηλό λάθος, της επιλεκτικότητας ερωτημάτων ζεύξεων πάνω από σχεσιακές βάσεις. Το προτεινόμενο πλαίσιο συνόψισης υιοθετεί μια ημι-δομημένη θεώρηση μιας σχεσιακής βάσης, όπου οι πλειάδες απεικονίζονται σε έναν γράφο και τα ερωτήματα σε διασχίσεις του γράφου. Η κύρια ιδέα είναι η περίληψη της δομής του γράφου σε μια μικρή σύνοψη, ώστε η διάσχιση του αρχικού γράφου δεδομένων για ένα ερώτημα να μεταφράζεται σε μια αντίστοιχη διάσχιση πάνω από την σύνοψη. Η εργασία μας περιγράφει τις λεπτομέρειες του μοντέλου συνόψισης που προτείνουμε, και παρουσιάζει έναν αποδοτικό αλγόριθμο για την κατασκευή σύνοψης για συγκεκριμένο μέγεθος χώρου αποθήκευσης. Τέλος, περιγράφουμε τα αποτελέσματα μιας πειραματικής μελέτης για την απόδοση των προτεινόμενων συνόψεων σε πραγματικά και συνθετικά δεδομένα. Τα αποτελέσματα δείχνουν ότι οι Συνόψεις Γράφου Πλειάδων παράγουν ακριβείς εκτιμήσεις για περίπλοκα ερωτήματα, παραμένοντας πιο αποδοτικές από παρόμοιες τεχνικές που έχουνε προταθεί στο παρελθόν.

Γρήγορη Προσεγγιστική Παρακολούθηση Συνόψεων Wavelet σε Ρεύματα Δεδομένων

G. Cormode (Lucent Bell Labs), Μ. Γαροφαλάκης (Intel Research Berkeley), Δ. Σαχαρίδης (ΕΜΠ)

Περίληψη: Τα πρόσφατα χρόνια έχει παρατηρηθεί ένα αυξανόμενο ενδιαφέρον για αποτελεσματικούς αλγορίθμους περίληψης και επερώτησης μεγάλων και υψηλής ταχύτητας ρευμάτων δεδομένων. Οι πιθανοτικές συνόψεις σκίτσων παρέχουν ακριβείς προσεγγίσεις για γενικής χρήσης περιλήψεις της κατανομής ρευμάτων δεδομένων (π.χ., wavelets). Η εστίαση των προηγούμενων δουλειών ήταν κυρίως στην ελαχιστοποίηση των απαιτήσεων χώρου της διατηρούμενης σύνοψης - εντούτοις, για την αποτελεσματική ανάλυση μεγάλων ρευμάτων, μια κρίσιμη πρακτική απαίτηση είναι η μέθοδος να βελτιστοποιεί επίσης: (1) τον χρόνο ανανέωσης για την ενσωμάτωση ενός στοιχείου του ρεύματος στο σκίτσο, και (2) τον χρόνο ερώτησης για την παραγωγή μιας κατά προσέγγιση περίληψης (π.χ., οι μεγαλύτεροι συντελεστές wavelet) από το σκίτσο. Τέτοια χρονικά κόστη πρέπει να είναι αρκετά μικρά ώστε να αντιμετωπίσουν τους γρήγορους ρυθμούς άφιξης δεδομένων και την απαίτηση για ερωτήσεις σε πραγματικό χρόνο των τυπικών εφαρμογών ρευμάτων δεδομένων (π.χ., η παρακολούθηση του δικτύου ενός παροχέα Internet). Καθώς η μνήμη σε υπολογιστικά συστήματα είναι φτηνή και άφθονη, ο χώρος είναι συχνά μόνο μια δευτερεύουσα ανησυχία πολύ μετά από το χρονικό κόστος ερώτησης και ανανέωσης. Σε αυτή τη δουλειά, προτείνουμε την πρώτη γρήγορη λύση στο πρόβλημα της παρακολούθησης προσεγγίσεων wavelets μονοδιάστατων και πολυδιάστατων ρευμάτων δεδομένων, βασισμένες σε μια νέα σύνοψη ρευμάτων, του Group-Count σκίτσου (GCS). Με την επιβολή μιας ιεραρχικής δομής των ομάδων πάνω από τα δεδομένα και την εφαρμογή του GCS, οι αλγόριθμοί μας μπορούν γρήγορα να ανακτήσουν τους σημαντικότερους συντελεστές wavelet με εγγυημένη ακρίβεια. Εισάγουμε την δυνατότητα ανταλλαγής μεταξύ του χρόνου ερώτησης και του χρόνου ανανέωσης, αλλάζοντας την ιεραρχική δομή των ομάδων, κάτι που επιτρέπει να βρούμε τη σωστή ισορροπία για το συγκεκριμένο ρεύμα δεδομένων. Η πειραματική ανάλυση επιβεβαιώνει αυτήν την δυνατότητα, και δείχνει ότι όλες οι μέθοδοί μας ξεπερνούν σημαντικά τις προηγουμένως γνωστές μεθόδους και σε ότι αφορά το χρόνο ανανέωσης και το χρόνο ερώτησης, διατηρώντας παράλληλα ένα υψηλό επίπεδο ακρίβειας.

Κατανεμημένες Συνόψεις Κατακερματισμού: Αποδοτική Εκτίμηση του Πλήθους Εγγραφών Μεγάλων Πολυσυνόλων σε Δίκτυα Δεδομένων Διαδικτυακής Κλίμακας

Ν. Ντάρμος (Παν. Πατρών), G. Weikum (MPI, Saarbruecken)

Περίληψη: Εν γένει η καταμέτρηση και πιο συγκεκριμένα η εκτίμηση του αριθμού των εγγραφών (πολυ-)συνόλων είναι ιδιαίτερα επιθυμητή για μεγάλο αριθμό εφαρμογών, αποτελώντας ένα βασικό δομικό στοιχείο για την αποδοτική ανάπτυξη και χρήση των σύγχρονων συστημάτων διαχείρισης πληροφορίας διαδικτυακής κλίμακας. Παραδείγματα τέτοιων εφαρμογών βρίσκει κανείς από την βελτιστοποίηση επεξεργασίας ερωτημάτων σε διαδικτυακές βάσεις δεδομένων, ως την εκτίμηση της σημαντικότητας διαφόρων δεδομένων σε εφαρμογές ανάκτησης δεδομένων. Οι κύριοι περιορισμοί τους οποίους κάθε αποδεκτή λύση θα πρέπει να ικανοποιεί είναι: (α) αποδοτικότητα: ο αριθμός των κόμβων που πρέπει να προσπελάσουμε κατά την καταμέτρηση πρέπει να είναι μικρός, ώστε η λύση να έχει χαμηλές απαιτήσεις εύρους ζώνης και καθυστέρησης, (β) κλιμάκωση, αντιτιθέμενη τον στόχο της αποδοτικότητας: ένας πολύ μεγάλος αριθμός κόμβων μπορεί να χρειαστεί να προσθέσει στοιχεία σε ένα (πολυ-)σύνολο, κάτι που απαιτεί ευρέως κατανεμημένη λύση, αποφεύγοντας τα προβλήματα κλιμάκωσης, καθυστέρησης και διαθεσιμότητας των λύσεων που βασίζονται σε κεντρικούς εξυπηρετητές, (γ) εξισορρόπηση φόρτου κατά την αποθήκευση και ανάκτηση: ο φόρτος της καταμέτρησης (και όχι μόνο) θα πρέπει να μοιράζεται εξίσου στους κόμβους του δικτύου, (δ) ακρίβεια: οι εκτιμήσεις που παίρνουμε θα πρέπει να είναι (ρυθμίσιμα) υψηλής ακρίβειας και σταθερότητας δεδομένης της δυναμικής των δικτύων ομότιμων και της ύπαρξης σφαλμάτων), και (ε) απλότητα και ευκολία χρησιμοποίησης: θα πρέπει να αποφευχθεί η χρήση δομών δεικτοδότησης ειδικά για τις εφαρμογές καταμέτρησης, πέρα από τις ήδη υπάρχουσες στο σύστημα. Στην εργασία αυτή πρώτα συνεισφέρουμε μία μέθοδο εκτίμησης πλήθους στοιχείων (πολυ-)συνόλων ευρέως κατανεμημένη, κλιμακώσιμη, αποδοτική και ακριβή. Εν συνεχεία, δείχνουμε πως μπορούμε να χρησιμοποιήσουμε την λύση αυτή για να κατασκευάσουμε και να διατηρήσουμε ιστογράμματα, τα οποία αποτελούν βασικό δομικό στοιχείο για την βελτιστοποίηση ερωτημάτων σε κεντρικοποιημένες βάσεις δεδομένων, διευκολύνοντας έτσι τη μεταφορά τους στον χώρο τον δικτύων δεδομένων διαδικτυακής κλίμακας.

Κατασκευή Δικτύων Επικάλυψης με Γνώση του Φορτίου Ερωτήσεων Βασισμένη στη Χρήση Ιστογραμμάτων

Γ. Κολωνιάρη, Γ. Πετράκης, Ε. Πιτουρά, Θ. Τσότσος (Παν. Ιωαννίνων)

Περίληψη: Τα συστήματα ομότιμων προσφέρουν ένα αποδοτικό μέσο για τον διαμοιρασμό δεδομένων μεταξύ ενός δυναμικού μεταβαλλόμενου συνόλου, αποτελούμενο από έναν μεγάλο αριθμό αυτόνομων κόμβων. Κάθε κόμβος σε ένα σύστημα ομότιμων συνδέεται με έναν μικρό αριθμό άλλων κόμβων σχηματίζοντας με αυτόν τον τρόπο ένα δίκτυο επικάλυψης από κόμβους. Μια ερώτηση που γίνεται σε έναν κόμβο δρομολογείται μέσω του δικτύου επικάλυψης προς τους κόμβους που διαθέτουν δεδομένα που την ικανοποιούν. Σε αυτήν την εργασία, ασχολούμαστε με τη δημιουργία δικτύων επικάλυψης που εκμεταλλεύονται το φορτίο ερωτήσεων του συστήματος, έτσι ώστε οι κόμβοι να ομαδοποιούνται με βάση τα αποτελέσματα που δίνουν για ένα δεδομένο φορτίο ερωτήσεων. Το κίνητρο μας είναι η δημιουργία τέτοιων δικτύων επικάλυψης στα οποία οι κόμβοι που ικανοποιούν μεγάλο πλήθος παρόμοιων ερωτήσεων να βρίσκονται μόνο ελάχιστα βήματα μακριά. Η συχνότητα των ερωτήσεων λαμβάνεται επίσης υπ'όψιν έτσι ώστε οι δημοφιλείς ερωτήσεις να επηρεάζουν τη δημιουργία του δικτύου περισσότερο από τις μη δημοφιλείς. Επικεντρωνόμαστε σε ερωτήσεις επιλογής εύρους και χρησιμοποιούμε ιστογράμματα για να εκτιμήσουμε τα αποτελέσματα των ερωτήσεων κάθε κόμβου. Στη συνέχεια, οι κόμβοι ομαδοποιούνται σύμφωνα με την ομοιότητα των ιστογραμμάτων τους. Γι' αυτό το σκοπό, εισάγουμε ένα μέτρο συντακτικής απόστασης με γνώση του φορτίου μεταξύ ιστογραμμάτων το οποίο λαμβάνει υπ'όψιν του το φορτίο ερωτήσεων του συστήματος. Τα πειραματικά μας αποτελέσματα δείχνουν ότι τα ενήμερα με βάση του φορτιού δίκτυα επικάλυψης αυξάνουν το ποσοστό των αποτελεσμάτων που επιστρέφονται για μια ερώτηση για ένα δεδομένο αριθμό επισκεπτόμενων κόμβων, σε σύγκριση με τυχαία δίκτυα (χωρίς ομαδοποίηση) και με μη ενήμερα με βάση του φορτίου δίκτυα (δίκτυα στα οποία η ομαδοποίηση των κόμβων βασίζεται αποκλειστικά στο περιεχόμενο τους).

Επεξεργασία Ερωτήσεων Σύζευξης σε Συστήματα Ομοτίμων

Σ. Ιδρέος, Χ. Τρυφωνόπουλος (Πολυτ. Κρήτης), Μ. Κουμπαράκης (Παν. Αθηνών)

Περίληψη: Σε αυτήν την εργασία μελετάμε το πρόβλημα της κατανεμημένης επεξεργασίας ερωτήσεων σύζευξης σε μεγάλα συστήματα ομότιμων που είναι οργανωμένα ως κατανεμημένοι πίνακες κατακερματισμού. Οι ερωτήσεις σύζευξης είναι περίπλοκο να απαντηθούν σε τέτοια περιβάλλοντα αφού πρέπει να συνδυάσουμε δεδομένα από διαφορετικά μέρη του δικτύου. Επιπλέον τα δεδομένα που αποτελούν μια απάντηση μπορεί να φθάνουν ασύγχρονα. Παρουσιάσουμε μια σειρά αλγορίθμων οι οποίοι προσπαθούν να κρατήσουν χαμηλά το κόστος των μηνυμάτων που δημιουργούνται στο δίκτυο αλλά και του φορτίου ερωτήσεων που έχει ο κάθε κόμβος.

PastryStrings - Σύστημα Δημοσίευσης/συνδρομής Περιεχομένου σε DHT Δίκτυο Ομοτίμων

Ι. Αικατερινίδης (Παν. Πατρών)

Περίληψη: Σε αυτήν την εργασία προτείνουμε και αναπτύσσουμε μια περιεκτική υποδομή (PastryStrings) που λειτουργεί πάνω από δίκτυο ομότιμων βασισμένο σε DHT (Distributed Hash Tables), για την υποστήριξη και την επεξεργασία συνδρομών που περιέχουν αριθμητικού τύπου πεδία (με δηλώσεις διαστημάτων τιμών και ισοτήτων) και αλφαριθμητικού τύπου πεδία (με δηλώσεις ισότητας, προθέματος, επιθέματος). Μιας και τα συστήματα δημοσίευσης συνδρομής βασισμένα στο περιεχόμενο (Content-based publish/subsrcibe) αποτελούν μία πρωτοπόρα κατηγορία εφαρμογών, διατυπώνουμε τους προτεινόμενους αλγόριθμους υπό την σκοπιά αυτού του περιβάλλοντος.

LOCUS: Χαλαρή και Βέλτιστη Ταξινόμηση με Απεριόριστη Κλιμάκωση

Κ. Μορφονιός, Γ. Ιωαννίδης (Παν. Αθηνών)

Περίληψη: Ένα από τα πιο ενδιαφέροντα σενάρια στην εξόρυξη γνώσης πάνω από πολυδιάστατα δεδομένα περιλαμβάνει ένα στάδιο επιλογής χαρακτηριστικών ακολουθούμενο από ένα στάδιο ταξινόμησης. Οι περισσότερες μέθοδοι που έχουν προταθεί μέχρι σήμερα για καθένα από τα δύο αυτά στάδια είναι βασισμένες στη μνήμη. Επιπλέον, οι περισσότερες μέθοδοι βασίζονται σε 'ανυπόμονη' (eager) ταξινόμηση, όπου πρώτα κατασκευάζεται (μέσω εκπαίδευσης) ένα γενικό μοντέλο περιγραφής των κλάσεων των δεδομένων στο δεδομένο χώρο του προβλήματος. Τα χαρακτηριστικά αυτά καθιστούν τέτοιες μεθόδους αναποτελεσματικές σε

(α) πολύ μεγάλα σύνολα δεδομένων αποθηκευμένα σε βάσεις ή αποθήκες δεδομένων, (β) δεδομένα των οποίων η διαμέριση σε κλάσεις δεν μπορεί να συλληφθεί από ένα γενικό μοντέλο και είναι ευαίσθητη σε τοπικά χαρακτηριστικά, και (γ) δεδομένα που εισέρχονται στο σύστημα συνεχώς και περιέχουν παρεμβαλλόμενα προ-ταξινομημένα και αταξινόμητα στιγμιότυπα και των οποίων η επιτυχής ταξινόμηση είναι ευαίσθητη στη χρήση των πιο ολοκληρωμένων ή/και πρόσφατων πληροφοριών. Στο παρόν άρθρο, προτείνουμε το LOCUS, έναν κλιμακούμενο αλγόριθμο επιλογής χαρακτηριστικών και 'χαλαρής' (lazy) ταξινόμησης που αντιμετωπίζει τα προαναφερθέντα προβλήματα. Ο LOCUS βασίζεται σε ιδέες από την αναγνώριση προτύπων και αποδεικνύεται ότι συγκλίνει στο βέλτιστο κατά Bayes ταξινομητή καθώς αυξάνεται το μέγεθος του υποκειμένου συνόλου δεδομένων. Επιπλέον, ο LOCUS είναι κλιμακούμενος ως προς το μέγεθος του συνόλου δεδομένων και υλοποιήσιμος με χρήση πρότυπης SQL πάνω από αυθαίρετες σχέσεις αποθηκευμένες σε βάσεις δεδομένων. Σύμφωνα με όσα γνωρίζουμε, ο LOCUS είναι ο πρώτος ταξινομητής που συνδυάζει όλα τα παραπάνω χαρακτηριστικά. Δείχνουμε την αποτελεσματικότητα του LOCUS με τη βοήθεια πειραμάτων τόσο πάνω από πραγματικά όσο και πάνω από συνθετικά σύνολα δεδομένων, συγκρίνοντάς τον με δένδρα απόφασης βασισμένα στη μνήμη. Σε αρκετά σύνολα δεδομένων, ο LOCUS επιτυγχάνει μεγαλύτερη ακρίβεια από τα δένδρα απόφασης. Στις περιπτώσεις που χάνει, είναι σχετικά κοντά. Επιπλέον, ο LOCUS μπορεί να αναγνωρίσει πολλαπλές τάσεις στα δεδομένα και να βελτιώσει την απόδοσή του στην περίπτωση αυτή. Επιπροσθέτως, μπορεί να διαχειριστεί πραγματικά μεγάλα σύνολα δεδομένων, πέρα από τα όρια μεγέθους των δένδρων απόφασης, και επιδεικνύει σημαντική βελτίωση της ακρίβειας των αποτελεσμάτων του καθώς το μέγεθος των συνόλων δεδομένων που χειρίζεται μεγαλώνει. Τέλος, ο LOCUS μπορεί να παραλληλιστεί με τη χρήση απλών μηχανισμών, αφαιρώντας επομένως οποιαδήποτε ενδεχόμενα όρια στην κλιμάκωσή του. Συνολικά, τα αποτελέσματα είναι υποσχόμενα ως προς τη χρήση των δυνατοτήτων του LOCUS σαν βάση για επιλογή χαρακτηριστικών και ταξινόμηση.

Το Πλαίσιο MONIC για τον Εντοπισμό Μεταβολών σε Συστάδες

Μ. Σπηλιοπούλου (U of Magdeburg), Ε. Ντούτση, Γ. Θεοδωρίδης (Παν. Πειραιά), R. Schult (U of Magdeburg)

Περίληψη: Τελευταία έχουν προταθεί αρκετές επιστημονικές εργασίες σχετικά με τον εντοπισμό και την παρακολούθηση των αλλαγών στις συστάδες΄ η μελέτη στηρίζεται συνήθως στις "χωροχρονικές" ιδιότητες των συστάδων. Για τις πολλές εφαρμογές όπου παρουσιάζεται το συγκεκριμένο πρόβλημα, μεταξύ αυτών η διαχείριση πελατών, ο εντοπισμός απάτης και το μάρκετιγκ, είναι επίσης απαραίτητη και η βαθιά γνώση της φύσης αυτών των αλλαγών: Μία συστάδα, η οποία αντιστοιχεί σε μία ομάδα πελατών, απλά εξαφανίστηκε ή τα μέλη της "μετανάστευσαν" σε άλλες συστάδες; Μία ανερχόμενη συστάδα απεικονίζει μία νέα ομάδα πελατών ή αποτελείται από υπάρχοντες πελάτες των οποίων οι προτιμήσεις άλλαξαν; Προκειμένου να μπορούμε να απαντάμε σε τέτοιου είδους ερωτήσεις, προτείνουμε το MONIC, ένα πλαίσιο για τη μοντελοποίηση και την παρακολούθηση μεταβολών στις συστάδες. Το μοντέλο μας δεν περιορίζεται στις αλλαγές μίας συστάδας, αλλά συμπεριλαμβάνει και αλλαγές που εμπλέκουν παραπάνω από μία συστάδες, επιτρέποντας έτσι την κατανόηση των αλλαγών σε ολόκληρη τη συσταδοποίηση. Ο μηχανισμός εντοπισμού μεταβολών που προτείνουμε δεν στηρίζεται στις τοπολογικές ιδιότητες των συστάδων, οι οποίες διατίθενται μόνο για μερικούς τύπους συσταδοποίησης, αλλά στα περιεχόμενα του ρεύματος δεδομένων. Παρουσιάζουμε τα πρώτα αποτελέσματα από την εφαρμογή του MONIC στην ψηφιακή βιβλιοθήκη της ACM΄ τα αποτελέσματα δείχνουν ότι μελετώντας τις αλλαγές των συστάδων μπορούμε να αποκτήσουμε βαθιά γνώση των αλλαγών που υπέστησαν τα δεδομένα.

Online Παρακολούθηση της Κατάστασης Δικτύου σε Περιβάλλον Δικτύων Μικροαισθητήρων

Μ. Χαλκίδη, Β. Καλογεράκη, Δ. Γουνόπουλος, Δ. Παπαδόπουλος, Δ. Ζεϊναλιπούρ-Γιαζτί (U of California, Riverside), Μ. Βλάχος (IBM T.J. Watson)

Περίληψη: Τα δίκτυα μικροαισθητήρων έχουν χρησιμοποιηθεί για ιχνηλάτιση γεγονότων που παρουσιάζουν ενδιαφέρον σε πολλές περιβαλλοντικές εφαρμογές ή άλλες εφαρμογές παρακολούθησης (monitoring). Ένα θέμα που παρουσιάζει ενδιαφέρον στα δίκτυα μικροαισθητήρων είναι πως να αναγνωρίσουμε με ακρίβεια τη συνολική κατάσταση του φαινομένου που παρατηρούμε. Στη συγκεκριμένη εργασία, προτείνουμε έναν online μηχανισμό για να καθορίσουμε αποδοτικά τη κατάσταση του δικτύου, εφαρμόζοντας κατανεμημένες λειτουργίες που ελαχιστοποιούν το κόστος επικοινωνίας. Στην προσέγγισή μας εφαρμόζουμε μία φάση μάθησης, κατά την οποία συλλέγουμε πληροφορία από τους μικροαισθητήρες και αναλύουμε αυτή τη

πληροφορία κατά το χρόνο εκτέλεσης για να βρούμε ένα σύνολο από δυνατές καταστάσεις δικτύου. Οι καταστάσεις ενημερώνονται on-line καθώς φτάνουν νέες μετρήσεις μικροαισθητήρων. Στη συνέχεια εξάγουμε τους κατάλληλους κανόνες και περιγράφουμε τις διάφορες καταστάσεις. Οι κανόνες αυτοί εφαρμόζονται στους μεμονωμένους μικροαισθητήρες και χρησιμοποιούνται για εύρεση της on-line κατάστασης του δικτύου κατά τη λειτουργία του. Τα πειράματα σε πραγματικά δεδομένα, δείχνουν ότι η προτεινόμενη μεθοδολογία μπορεί να είναι βιώσιμη λύση για πραγματικά συστήματα.

Έλεγχος Περιεκτικότητας για Ερωτήσεις Δεντρικών Προτύπων Μερικού Προσδιορισμού

Δ. Θεοδοράτος, Θ. Δαλαμάγκας, P. Placek, Σ. Σουλδάτος, Τ. Σελλής (ΕΜΠ)

Περίληψη: Στην εποχή μας, τεράστιος όγκος δεδομένων, μεταξύ άλλων και επιστημονικών, οργανώνονται ή διακινούνται σε δεντρική μορφή. Οι γλώσσες ερωτήσεων για δεδομένα δεντρικής μορφής, βασίζονται σε ερωτήσεις δεντρικών προτύπων. Όμως, η ανάγκη για ολοκλήρωση πηγών δεδομένων με διαφορετική δεντρική οργάνωση έχει οδηγήσει στο σχεδιασμό γλωσσών που επιτρέπουν ερωτήσεις δεντρικών προτύπων μερικού προσδιορισμού. Η εργασία αυτή ασχολείται με μια τέτοιας μορφής γλώσσα ερωτήσεων. Το κύριο χαρακτηριστικό της είναι ότι στα δεντρικά πρότυπα η δομή μπορεί να καθοριστεί πλήρως, μερικώς ή και καθόλου. Βασική προϋπόθεση για την βελτιστοποίηση ερωτήσεων αποτελεί η επίλυση του προβλήματος του ελέγχου περιεκτικότητας των ερωτήσεων. Στην εργασία αυτή μελετούμε το πρόβλημα αυτό για ερωτήσεις δεντρικών προτύπων μερικού προσδιορισμού. Ειδικά για την αποτίμηση τέτοιων ερωτήσεων χρησιμοποιούμε τους γράφους διαστάσεων, οι οποίοι και αποτελούν δομικές περιλήψεις δεντρικών δεδομένων. Εξετάζουμε το πρόβλημα του ελέγχου περιεκτικότητας ερωτήσεων είτε με την παρουσία είτε χωρίς την παρουσία γράφου διαστάσεων. Επίσης, προτείνουμε ικανές και αναγκαίες συνθήκες για να ισχύει η περιεκτικότητα σε κάθε μια από τις παραπάνω δύο περιπτώσεις. Τέλος, παρουσιάζουμε μια τεχνική ελέγχου περιεκτικότητας ερωτήσεων, με παρουσία γράφου διαστάσεων, που εκμεταλλεύεται την δομική πληροφορία του γράφου για να επιταχύνει τον έλεγχο. Η τεχνική μας αξιολογείται με μια σειρά πειραμάτων που αποδεικνύουν την αποδοτικότητά της.

PROTEAS: ένας Αλγόριθμος Εξόρυξης Δεδομένων Βασισμένος σε Πεπερασμένα Αυτόματα για την Εξαγωγή Κανόνων Ταξινόμησης Πρωτεϊνών

Φ. Ψωμόπουλος, Π. Μήτκας (Αριστοτέλειο Παν.)

Περίληψη: Ένα από τα σημαντικότερα προβλήματα της σύγχρονης πρωτεομικής είναι η πρόβλεψη της λειτουργικής συμπεριφοράς των πρωτεϊνών. Μοτίβα τα οποία παρουσιάζονται στην πρωτεϊνική αλυσίδα καθιστούν δυνατή μια τέτοια πρόβλεψη. Καθώς σε μια αλυσίδα μπορεί να εμφανίζονται περισσότερα του ενός μοτίβα, δεν είναι εύκολο να βρεθεί η συσχέτιση μεταξύ των ιδιοτήτων μιας πρωτεΐνης και των μοτίβων της. Έτσι η συμπεριφορά που εκδηλώνει μια πρωτεΐνη είναι συνάρτηση πολλών μοτίβων, όπου κάποια υπερισχύουν έναντι άλλων. Μια νέα προσέγγιση εξόρυξης δεδομένων στο πρόβλημα, η οποία παρουσιάζεται εδώ, κάνει χρήση πεπερασμένων αυτομάτων. Αρχικά τα δεδομένα μοντελοποιούνται στη μορφή δένδρων αποδοχής προθεμάτων, τα οποία στη συνέχεια συγχωνεύονται με κατάλληλο τρόπο σε αιτιοκρατικά πεπερασμένα αυτόματα. Τέλος προτείνεται ένα νέος αλγόριθμος για την εξαγωγή κανόνων ταξινόμησης πρωτεϊνών από τα αυτόματα. Το μοντέλο εκπαιδεύεται και ελέγχεται με διάφορα υποσύνολα πρωτεϊνών και πρωτεϊνικών οικογενειών, καθώς και με το σύνολο των γνωστών πρωτεϊνών. Τα αποτελέσματα καταδεικνύουν την αποτελεσματικότητα της μεθοδολογίας έναντι άλλων γνωστών αλγορίθμων ταξινόμησης.

CURE για Κύβους: Κατασκευή Κύβου με Χρήση Σχεσιακών Συστημάτων

Κ. Μορφονιός, Γ. Ιωαννίδης (Παν. Αθηνών)

Περίληψη: Η κατασκευή του κύβου δεδομένων έχει βρεθεί στο επίκεντρο αρκετών ερευνητικών προσπαθειών εξαιτίας της σημασίας της στη βελτίωση της απόδοσης της άμεσης αναλυτικής επεξεργασίας (OLAP). Ένα σημαντικό μέρος των εργασιών αυτών αφορά τεχνικές ROLAP, που βασίζονται στη σχεσιακή τεχνολογία. Οι υπάρχοντες αλγόριθμοι της κατηγορίας αυτής κυρίως εστιάζουν σε 'επίπεδα' δεδομένα, τα οποία δεν περιλαμβάνουν ιεραρχίες μεταξύ των διαστάσεών τους. Ωστόσο, η φύση των ιεραρχιών εισάγει διάφορες επιπλοκές στην κατασκευή του κύβου, καθιστώντας τούς υπάρχοντες αλγορίθμους ουσιαστικά ανεφάρμοστους σε ένα σημαντικό αριθμό πραγματικών εφαρμογών. Πιο συγκεκριμένα, οι ιεραρχίες εισάγουν κυρίως τρεις νέες προκλήσεις: (α) Το πλήθος των κόμβων του πλέγματος του κύβου δεδομένων αυξάνεται δραματικά. Αυτό απαιτεί εύρεση νέων τρόπων διάσχισης του πλέγματος για αποδοτική εκτέλεση.

(β) Ο αριθμός των διαφορετικών τιμών στα υψηλότερα επίπεδα ιεραρχίας κάθε διάστασης είναι συνήθως πολύ μικρός, επομένως, η διαμέριση των δεδομένων σε τμήματα που χωρούν στη μνήμη και περιλαμβάνουν όλες τις εμφανίσεις μιας συγκεκριμένης τιμής είναι συχνά αδύνατη. Αυτό απαιτεί νέες τεχνικές διαμέρισης. (γ) Το πλήθος των πλειάδων που πρέπει να αποθηκευθούν στον κύβο τελικά αυξάνεται δραματικά. Αυτό απαιτεί νέα σχήματα αποθήκευσης που αφαιρούν όλους τους τύπους πλεονασμού για αποδοτική χρήση του αποθηκευτικού χώρου. Στο παρόν άρθρο, προτείνουμε τον CURE, έναν καινοτόμο αλγόριθμο κατασκευής κύβων ROLAP που αντιμετωπίζει τα προαναφερθέντα θέματα και κατασκευάζει πλήρεις κύβους πάνω από πολύ μεγάλα σύνολα δεδομένων με αυθαίρετη ιεραρχική οργάνωση. Ο CURE συνεισφέρει έναν πρωτότυπο τρόπο διάσχισης του πλέγματος, μία βελτιστοποιημένη μέθοδο διαμέρισης, και μία συλλογή σχεσιακών σχημάτων αποθήκευσης για όλους τους τύπους πλεοναστικών δεδομένων. Δείχνουμε την αποτελεσματικότητα του CURE με τη βοήθεια πειραμάτων τόσο πάνω σε πραγματικά όσο και σε συνθετικά σύνολα δεδομένων. Μεταξύ των πειραματικών αποτελεσμάτων, ξεχωρίζουμε εκείνα που κατέστησαν τον CURE την πρώτη μέθοδο ROLAP που ολοκλήρωσε την κατασκευή του κύβου πάνω από σύνολα δεδομένων με τη μεγαλύτερη παράμετρο πυκνότητας στη δοκιμασία επιδόσεων APB-1 (12 GB). Ο CURE ήταν στην πραγματικότητα αρκετά αποδοτικός σε αυτό, δίνοντας μεγάλες υποσχέσεις σχετικά με τις δυνατότητές του συνολικά.

Νέες Τεχνικές για Αποδοτική Εκτέλεση XPath Επερωτήσεων
Χ. Γεωργιάδης, Β. Βασσάλος (Οικονομικό Παν. Αθηνών)

Περίληψη: Η εργασία αυτή περιγράφει μια μέθοδο για επεξεργασία XPath επερωτήσεων σε σχεσιακά συστήματα, η οποία μειώνει τον αριθμό των απαιτούμενων συνενώσεων (joins), εκμεταλλεύεται τις δυνατότητες των σύγχρονων επεξεργαστών SQL επερωτήσεων, αξιοποιεί την ύπαρξη XML σχήματος και έχει χαμηλό κόστος υλοποίησης. Η μέθοδος βασίζεται στην διαίρεση XPath εκφράσεων σε πρωτεύοντα τμήματα, που ονομάζουμε PPF, και στον μετέπειτα συνδυασμό τους χρησιμοποιώντας μια αποδοτική μέθοδο συνένωσης, ενώ καλύπτει όλους του δομικούς άξονες της XPath. Ακολουθεί αναλυτική περιγραφή της μεθόδου καθώς και πειραματική μελέτη που δείχνει τα προτερήματά της έναντι άλλων μεθόδων και συστημάτων.

Précis: Ένας Καινούριος Τρόπος Απαντήσεων σε Επερωτήσεις Πάνω σε Σχεσιακά Δεδομένα
Γ. Κουτρίκα, Α. Σιμιτσής, Γ. Ιωαννίδης (Παν. Αθηνών)

Περίληψη: Στο άρθρο αυτό, παρουσιάζουμε ένα νέο είδος επερωτήσεων. Πρόκειται για επερωτήσεις ελεύθερης μορφής των οποίων οι απαντήσεις αποτελούν μια σύνθεση αποτελεσμάτων, που δεν εμπεριέχουν μόνο πληροφορία σχετική με την επερώτηση, αλλά και πληροφορία που σχετίζεται με αυτή με διάφορους τρόπους. Η προσέγγισή μας επιφέρει δύο βασικές καινοτομίες σε σχέση με την υπάρχουσα δουλειά σε αναζήτηση βασισμένη σε λέξεις-κλειδιά: (α) οι επερωτήσεις δεν παράγουν ανεξάρτητες σχέσεις, αντίθετα συνθέτουν λογικά υποσύνολα της βάσης δεδομένων, και (β) τα αποτελέσματα των επερωτήσεων εξατομικεύονται ανάλογα με το χρήστη ή με περιορισμούς κυριότητας. Παράλληλα, παρουσιάζουμε ένα σύνολο πειραμάτων που αναδεικνύει τη σπουδαιότητα της μεθόδου μας.

UPR: Κατάταξη Σελίδων Βασισμένη στα Δεδομένα Χρήσης για Εξατομίκευση Δικτυακών Τόπων
Μ. Ειρηνάκη, Μ. Βαζυργιάννης (Οικονομικό Παν. Αθνών)

Περίληψη: Οι αλγόριθμοι παραγωγής προτάσεων έχουν ως σκοπό τη σύσταση 'επόμενων' σελίδων στους χρήστες, ανάλογα με τη πλοήγησή τους. Η μεγάλη πλειοψηφία των προτεινόμενων αλγορίθμων βασίζεται μόνο στα δεδομένα χρήσης του δικτυακού τόπου προκειμένου να παραχθούν αυτές οι προτάσεις. Σε αυτή την εργασία υποστηρίζουμε ότι η ενσωμάτωση της δομής του δικτυακού τόπου, καθώς και η χρήση αλγορίθμων ανάλυσης υπερσυνδέσμων βελτιώνει την ποιότητα των προτάσεων. Παρουσιάζουμε τον UPR, έναν αλγόριθμο εξατομίκευσης ο οποίος συνδυάζει τα δεδομένα χρήσης με τεχνικές ανάλυσης υπερσυνδέσμων προκειμένου να κατατάξει και να προτείνει σελίδες στον τελικό χρήστη. Χρησιμοποιώντας τη δομή του δικτυακού τόπου, τις συνεδρίες προηγούμενων χρηστών και την παρούσα επίσκεψη, δημιουργούμε εξατομικευμένους υπο-γράφους πλοήγησης (prNGs) στους οποίους εφαρμόζεται ο UPR. Τα πειραματικά αποτελέσματα δείχνουν ότι οι προτάσεις που παράγονται με αυτή τη μέθοδο έχουν μεγαλύτερη ακρίβεια από αυτές που παράγονται από αμιγώς βασισμένες σε δεδομένα χρήσης προσεγγίσεις.

Ερώτησεις Προτίμησης με βάση το Συμφραζόμενο Περιβάλλον: μια Επισκόπηση

Κ. Στεφανίδης, Ε. Πιτουρά, Π. Βασιλειάδης (Παν. Ιωαννίνων)

Περίληψη: Για τη διαχείριση της μεγάλης ποσότητας πληροφορίας που είναι διαθέσιμη σήμερα, τα συστήματα εξατομίκευσης επιτρέπουν στους χρήστες να καθορίσουν τις πληροφορίες που τους ενδιαφέρουν μέσω των προτιμήσεων. Γενικά, οι χρήστες μπορούν να έχουν διαφορετικές προτιμήσεις ανάλογα με το συμφραζόμενο περιβάλλον, για παράδειγμα, τον καιρό, το χρόνο, ή τη θέση του χρήστη. Σε αυτήν την εργασία, ορίζουμε ένα μοντέλο για την έκφραση των βασισμένων στα συμφραζόμενα προτιμήσεων. Μοντελοποιούμε το συμφραζόμενο περιβάλλον ως ένα διατεταγμένο σύνολο πολυδιάστατων γνωρισμάτων. Έτσι, οι προτιμήσεις των χρηστών μπορούν να εκφραστούν μέσω αυτών των γνωρισμάτων. Στη συνέχεια, διατυπώνουμε το πρόβλημα εύρεσης των προτιμήσεων που είναι πιο σχετικές με μία ερώτηση και παρουσιάζουμε έναν αλγόριθμο που τις εντοπίζει. Εισάγουμε επίσης, δύο δομές που εκμεταλλεύονται τις βασισμένες στα συμφραζόμενα πληροφορίες για την αποθήκευση (α) των προτιμήσεων και (β) των αποτελεσμάτων των ερωτήσεων που βασίζονται στο συμφραζόμενο περιβάλλον.

Ένας Συνδυασμός Δένδρων Trie και Αντεστραμμένων Αρχείων για τη Δεικτοδότηση Χαρακτηριστικών Βασισμένων σε Σύνολα

Μ. Τερροβίτης, Σ. Πασσάς (ΕΜΠ), Π. Βασιλειάδης (Παν. Ιωαννίνων), Τ. Σελλής (ΕΜΠ)

Περίληψη: Δεδομένα σε μορφή τιμών-συνόλων εμφανίζονται σε περιοχές εφαρμογής όπως η ανάλυση καλαθιού αγοράς και χρηματιστηριακών τάσεων. Η πρόσφατη ερευνητική βιβλιογραφία επικεντρώνεται σε set containment joins και ερωτήσεις εξόρυξης γνώσης, χωρίς να ασχολείται με απλά ερωτήματα για τις τιμές-σύνολα. Σε αυτή την εργασία αντιμετωπίζουμε ερωτήσεις υπερσυνόλου, ισότητας και υποσυνόλου και προτείνουμε ένα καινοτόμο ευρετήριο για την επεξεργασίας τους σε τιμές-σύνολα. Το προτεινόμενο ευρετήριο υπερθέτει ένα trie σε ένα ευρετήριο ανεστραμμένου αρχείου που δεικτοδοτεί τιμές-σύνολα. Δείχνουμε ότι μπορούμε να απαντήσουμε αποδοτικά τις προαναφερθείσες ερωτήσεις με το να δεικτοδοτούμε μόνο ένα υποσύνολο των αντικειμένων που εμφανίζονται συχνότερα στην προκείμενη σχέση. Τέλος, δείχνουμε, μέσα από εκτεταμένα πειράματα, ότι η προσέγγιση μας ξεπερνάει σε απόδοση τους βέλτιστους αντίστοιχους μηχανισμούς και κλιμακώνεται καλύτερα καθώς το μέγεθος της βάσης αυξάνει.

Συνεχής Ελεγχος Κορυφαίων-k Ερωτημάτων σε Συνεχείς Ροές Δεδομένων

Κ. Μουρατίδης, Σ. Μπακίρας, Δ. Παπαδιάς (Hong Kong U of Science & Technology)

Περίληψη: Δοθέντος ενός συνόλου δεδομένων P και μιας συνάρτησης προτίμησης f, ένα κορυφαίο-k ερώτημα ανακτά τα k αντικείμενα από το P με την υψηλότερη βαθμολογία σύμφωνα με την f. Οι υπάρχουσες μέθοδοι για συμβατικές βάσεις δεδομένων είναι μη εφαρμόσιμες σε δυναμικά περιβάλλοντα που περιλαμβάνουν πολυάριθμα και μακροπρόθεσμα ερωτήματα. Η παρούσα εργασία μελετά το συνεχή έλεγχο κορυφαίων-k ερωτημάτων σε ένα παράθυρο σταθερού μεγέθους W με τα πιο πρόσφατα δεδομένα. Το μέγεθος του παραθύρου μπορεί να εκφραστεί είτε ως ο αριθμός των ενεργών αντικειμένων είτε ως χρονικές μονάδες. Προτείνουμε μια γενική μεθοδολογία που περιορίζει την επεξεργασία στις υποπεριοχές του χώρου που επηρεάζουν το αποτέλεσμα κάποιου ερωτήματος. Για να αντεπεξέλθουμε στους υψηλούς ρυθμούς ροής δεδομένων, τα δεδομένα στο W παραμένουν στην κύρια μνήμη. Τα έγκυρα αντικείμενα δεικτοδοτούνται από μια δομή πλέγματος η όποια διατηρεί επίσης λογιστική πληροφορία. Παρουσιάζουμε δύο τεχνικές επεξεργασίας: η πρώτη υπολογίζει τη νέα απάντηση ενός ερωτήματος όταν μερικά από τα τρέχοντα κορυφαία-k αντικείμενα λήγουν, ενώ η δεύτερη προϋπολογίζει μερικώς τις μελλοντικές αλλαγές στο αποτέλεσμα, επιτυγχάνοντας καλύτερο χρόνο επεξεργασίας εις βάρος ελαφρώς υψηλότερων απαιτήσεων μνήμης. Αναλύουμε την απόδοση και των δύο αλγορίθμων και αξιολογούμε την αποδοτικότητά τους μέσω εκτενών πειραμάτων. Τέλος, επεκτείνουμε το προτεινόμενο πλαίσιο σε άλλους τύπους ερωτημάτων και σε ένα διαφορετικό πρότυπο ροής δεδομένων.

Επιτυγχάνοντας την Ενημέρωση Όψεων χωρίς Παρενέργειες

Γ. Κωτίδης, D. Srivastava, Γ. Βελεγράκης (AT&T Labs)

Περίληψη: Η χρήση εικονικών όψεων είναι πολύ διαδεδομένη σε συστήματα βάσεων δεδομένων. Για τις περισσότερες εφαρμογές οι όψεις δε διαφέρουν από κοινούς πίνακες, επομένως θα πρέπει να μπορούν να ενημερώνονται όταν υπάρχει ανάγκη. Επειδή οι όψεις είναι εικονικές, οι εντολές ενημέρωσης πρέπει να με-

ταφραστούν σε αλλαγές πάνω στους πίνακες της βάσης. Όπως είναι γνωστό, συνήθως τέτοιες μεταφράσεις προκαλούν ανεπιθύμητες παρενέργειες. Στην εργασία αυτή προτείνουμε μία λύση στο πρόβλημα της ενημέρωσης όψεων μέσο του διαχωρισμού της υπόστασης κάθε όψης σε φυσικό και λογικό σχήμα. Με αυτό το διαχωρισμό επιτρέπουμε την ενημέρωση των όψεων χωρίς ανεπιθύμητες παρενέργειες στην όψη και στους πίνακες της βάσης. Παρουσιάζουμε μία υλοποίηση της αρχιτεκτονικής σε ένα πραγματικό σύστημα βάσεων δεδομένων και δείχνουμε ότι μπορεί να επιτρέψει οποιαδήποτε ενημέρωση πάνω στις όψεις με πολύ μικρό κόστος.

Χρονοδρομολόγηση Ενημερώσεων και Ερωτήσεων με Συμβόλαια Ποιότητας για Βάσεις Δεδομένων του Παγκόσμιου Ιστού

H. Qu, A. Λαμπρινίδης (U of Pittsburgh)

Περίληψη: Σε μοντέρνα συστήματα βάσεων δεδομένων στον παγκόσμιο ιστό πληροφοριών, οι χρήστες συνήθως υποβάλλουν ερωτήσεις, ενώ όλες οι ενημερώσεις γίνονται στο υπόβαθρο, ταυτόχρονα με τις ερωτήσεις. Στο άρθρο αυτό, εισάγουμε την έννοια των Συμβολαίων Ποιότητας (Σ.Π.), τα οποία δίνουν τη δυνατότητα στους χρήστες να εκφράσουν τις προτιμήσεις τους αναθέτοντας τιμές κέρδους για την Ποιότητα Εξυπηρέτησης (Π.Ε.) και Ποιότητα Δεδομένων (Π.Δ.) που περιμένουν στις απαντήσεις που θα πάρουν. Συγκεκριμένα, προτείνουμε έναν προσαρμοστικό αλγόριθμο, QUTS, που μεγιστοποιεί το συνολικό κέρδος από τα Συμβόλαια Ποιότητας που έχουν έρθει στο σύστημα και, με τον τρόπο αυτό, μεγιστοποιεί το συνολικό βαθμό ικανοποίησης των χρηστών. Ο αλγόριθμος QUTS δίνει λύση στο πρόβλημα της ανάθεσης προτεραιότητας για τη χρονοδρομολόγηση των ενημερώσεων (που είναι κρίσιμες για την Π.Δ.) και των ερωτήσεων από τους χρήστες (που είναι κρίσιμες για την Π.Ε.), χρησιμοποιώντας ένα διεπίπεδο σχήμα χρονοδρομολόγησης, το οποίο αναθέτει δυναμικά υπολογιστικούς πόρους μεταξύ ερωτήσεων και ενημερώσεων. Σε αυτό το άρθρο, παρουσιάζουμε τα αποτελέσματα μιας εκτενούς πειραματικής μελέτης, βασισμένης σε πραγματικά δεδομένα πρόσβασης και ενημέρωσης, και δείχνουμε ότι ο προτεινόμενος αλγόριθμος QUTS έχει πολύ καλύτερη απόδοση από τους αλγορίθμους που χρησιμοποιήσαμε ως βάση αναφοράς. Η υπεροχή του QUTS έναντι των άλλων αλγορίθμων είναι σταθερή σε ολόκληρο το φάσμα των Συμβολαίων Ποιότητας. Τέλος, ο αλγόριθμος QUTS προσαρμόζεται γρήγορα σε μεταβαλλόμενα περιβάλλοντα και δεν έχει ευαισθησία στις παραμέτρους του.

Σχετικά με τα Γνωρίσματα των Γράφων των Σχημάτων του Σημασιολογικού Ιστού

Γ. Θεοχάρης, Β. Χριστοφίδης, Γ. Τζίτζικας (Παν. Κρήτης)

Περίληψη: Στην παρούσα δημοσίευση μετρούμε και αναλύουμε τα γνωρίσματα των γράφων των σχημάτων του Σημασιολογικού Ιστού. Δεδομένου ότι η ανάπτυξή τους είναι αποτέλεσμα κοινωνικής συνεργασίας, εστιάζουμε στη διερεύνηση γνωρισμάτων γράφων που αναδύονται στην ανάλυση των κοινωνικών δικτύων, όπως η κατανομή των βαθμών των κόμβων που ακολουθεί τον 'νόμο της δύναμης' και το φαινόμενο του 'μικρού κόσμου'. Τα κύρια συμπεράσματα της ανάλυσής μας είναι: α) το 88,5% των σχημάτων με σημαντικό αριθμό κλάσεων προσεγγίζουν το 'νόμο της δύναμης' για την κατανομή των υποκλάσεων, β) το 94,3% των σχημάτων με σημαντικό αριθμό κλάσεων και ιδιοτήτων προσεγγίζουν το 'νόμο της δύναμης' για την κατανομή των βαθμών των κόμβων, και γ) το 69,8% των σχημάτων με σημαντικό αριθμό ιδιοτήτων εμφανίζουν το φαινόμενο του 'μικρού κόσμου'.

Αποτελεσματικές Προσεγγιστικές Τεχνικές για Υπολογισμό Ερωτήσεων σε Κατανεμημένες Βάσεις Δεδομένων Υψηλής Κλίμακας

B. Arai, G. Das, Δ. Γουνόπουλος, Β. Καλογεράκη (U of California, Riverside)

Περίληψη: Οι υψηλής κλίμακας κατανεμημένες βάσεις δεδομένων (peer-to-peer) έχουν γίνει πολύ διαδεδομένες στο διαδίκτυο για τη διανομή και την αποθήκευση δεδομένων, εφαρμογών, και άλλων ψηφιακών μέσων. Στην εργασία αυτή εξετάζουμε το πρόβλημα της γρήγορης και αποτελεσματικής προσεγγιστικής απάντησης αθροιστικών ερωτήσεων σε τέτοια συστήματα. Ακριβείς απαντήσεις απαιτούν μεγάλο χρονικό διάστημα επεξεργασίας των δεδομένων, και η υλοποίησή τους μπορεί να είναι δύσκολη, δεδομένης της κατανεμημένης και δυναμικής φύσης των peer-to-peer βάσεων δεδομένων. Στην εργασία αυτή δίνουμε νέες τεχνικές για την προσεγγιστική απάντηση αθροιστικών ερωτήσεων σε τέτοιες βάσεις δεδομένων με την χρήση τεχνικών τυχαίας δειγματοληψίας. Ο υπολογισμός ενός υψηλής ποιότητας δείγματος από την βάση δεδομένων είναι δύσκολος για διάφορους λόγους: τα δεδομένα είναι διαχωρισμένα

σε ανισομερή κομμάτια σε διαφορετικούς κόμβους, τα δεδομένα σε κάθε κόμβο είναι συχνά συσχετισμένα, και επιπλέον, είναι συχνά δύσκολο να πάρουμε ένα τυχαίο δείγμα από το σύνολο των κόμβων. Για να αντιμετωπίσουμε αυτά τα προβλήματα, έχουμε σχεδιάσει μια τεχνική δειγματοληψίας με δύο φάσεις, που βασίζεται σε τυχαίες διαδρομές στο γράφο των κόμβων, και σε τεχνικές δειγματοληψίας block-sampling. Δίνουμε επίσης μια εκτεταμένη πειραματική μελέτη που επιβεβαιώνει τις δυνατότητες της τεχνικής που προτείνουμε.

Η Προσέγγιση SOWES για Αναζήτηση στον Παγκόσμιο Ιστό με χρήση Σημασιολογικών Δικτύων Επικάλυψης

K. Noervaag, Χ. Δουλκερίδης, Μ. Βαζυργιάννης (Οικονομικό Παν. Αθηνών)

Περίληψη: Η αναζήτηση στον παγκόσμιο ιστό με χρήση δικτύων ομότιμων κόμβων έχει προσελκύσει αρκετό ενδιαφέρον τελευταία, λόγω των χαρακτηριστικών των δικτύων ομότιμων κόμβων, δηλαδή της επεκτασιμότητας, της ανοχής σε σφάλματα και το διαμοιρασμό του φόρτου εργασίας. Μολαταύτα, η έλλειψη καθολικής γνώσης σε ένα τεράστιο και δυναμικά εξελισσόμενο περιβάλλον, όπως ο Παγκόσμιος Ιστός, δημιουργεί μια μεγάλη πρόκληση για οργάνωση περιεχομένου και παροχή αποδοτικών μηχανισμών αναζήτησης. Τα σημασιολογικά δίκτυα επικάλυψης έχουν προταθεί σαν μια προσέγγιση που μειώνει το κόστος και αυξάνει την ποιότητα των αποτελεσμάτων σε αδόμητα δίκτυα ομότιμων κόμβων, και σε αυτό το άρθρο παρουσιάζουμε τη SOWES προσέγγιση για κατανεμημένη κατασκευή σημασιολογικών δικτύων επικάλυψης σε αδόμητα δίκτυα ομότιμων κόμβων. Παρουσιάζονται αποδοτικές στρατηγικές αναζήτησης ιστοπεριεχομένου με βάση τα δημιουργηθέντα σημασιολογικά δίκτυα ομότιμων κόμβων και αποτιμάται η απόδοση της προσέγγισή μας μέσω πειραμάτων προσομοίωσης.

Αυτόματη Κατανομή Επερωτήσεων σε Αυτόνομες Κατανεμημένες Βάσεις Δεδομένων

Φ. Πεντάρης, Γ. Ιωαννίδης (Παν. Αθηνών)

Περίληψη: Σε μεγάλες ομοσπονδίες αυτόνομων βάσεων δεδομένων, η ανάθεση των επερωτήσεων στους κόμβους είναι ένα σημαντικό θέμα. Αναλύουμε το πρόβλημα αυτό χρησιμοποιώντας θεωρία μικροοικονομικών και δείχνουμε πώς αυτή μπορεί να χρησιμοποιηθεί για τη κατασκευή ενός κατανεμημένου και αποδοτικού μηχανισμού που μεγιστοποιεί την απόδοση αυτών των συστημάτων. Ειδικότερα, περιγράφουμε μία λύση που βασίζεται στην έννοια των αγορών επερωτήσεων. Χρησιμοποιώντας τις ιδιότητες αυτών των αγορών αποδεικνύουμε πώς αυτές οδηγούν σε βέλτιστες Pareto κατανομές των επερωτήσεων στους κόμβους του δικτύου. Με ένα ευρύ σύνολο από πειράματα, τόσο με προσομοιωτή, όσο και με μία πραγματική υλοποίηση του αλγόριθμου μας πάνω από ένα γνωστό ΣΔΒΣ, δείχνουμε ότι η προτεινόμενη από εμάς λύση βελτιώνει σημαντικά την απόδοση των κατανεμημένων ΣΔΒΔ.