ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων

«Πιθανοτικές Βάσεις Δεδομένων»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αλέξανδρος Α. Τρέντσιος

altrents@inf.uth.gr

Επιβλέπων καθηγητής: Παναγιώτης Μποζάνης

Βόλος , Φεβρουάριος 2010

Περιεχόμενα

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον καθηγητή μου Παναγιώτη Μποζάνη για την πολύτιμη καθοδήγηση του κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας .

Επίσης θα ήθελα να ευχαριστήσω την οικογένεια μου για την πολύτιμη στήριξη της και βοήθεια της σε όλη τη διάρκεια των σπουδών μου.

Σύνοψη

Δεδομένα που παίρνουν τιμές με στατιστική πιθανότητα να είναι σωστές ονομάζονται αβέβαια δεδομένα. Προσφάτως, μια σειρά από εφαρμογές συλλογής πληροφορίας έχουν επαναφέρει στην επιφάνεια το θέμα των αβέβαιων δεδομένων. Τέτοια, δεδομένα όταν αποθηκεύονται σε βάσεις δεδομένων δημιουργούν μια σειρά από προκλήσεις και προβλήματα για την αναπαράσταση τους, τη διαχείριση και χρήση τους και την ασφαλή εξαγωγή συμπερασμάτων. Βάσεις δεδομένων που περιέχουν αβέβαια δεδομένα αντιμετωπίζονται ως πιθανοτικές (probabilistic). Οι τελευταίες ερευνητικές προσπάθειες στον τομέα έχουν οδηγήσει στην ανάπτυξη μιας σειράς μοντέλων και μεθοδολογιών για την διαχείριση της αβέβαιης πληροφορίας. Στην παρούσα εργασία φιλοδοξούμε να κατηγοριοποιήσουμε τις μεθόδους μοντελοποίησης των πιθανοτικών δεδομένων και να αναλύσουμε και να κατηγοριοποιήσουμε τις σημαντικότερες προτεινόμενες λύσεις και αλγορίθμους. Επίσης, να παρουσιάζουμε τις κυριότερες μεθόδους εξόρυξης δεδομένων από πιθανοτικές βάσεις και τον τρόπου διαχείρισής τους. Τέλος, επιχειρούμε να παρουσιάσουμε τρέχοντα projects, συστήματα και εφαρμογές καθώς και προβλήματα του τομέα που δεν έχουν αντιμετωπιστεί ακόμα.

1. Εισαγωγή

Μέχρι τώρα οι βάσεις δεδομένων είχαν να κάνουν με διακριτές (deterministic) τιμές. Μια σειρά από εφαρμογές δημιουργούσαν τα δεδομένα τα οποία αποθηκεύονταν σε αυτές θεωρούμενα de facto σωστά. Μια σειρά μεθόδων έδιναν τη δυνατότητα επεξεργασίας και διαχείρισης των δεδομένων ώστε να εξαχθούν χρήσιμα συμπεράσματα, ή να γίνει επεξεργασία της πληροφορίας. Τέτοια παραδείγματα αποτελούν εφαρμογές αρχειοθέτησης, logistics, τραπεζών, λογιστηρίου κλπ.

Παρόλα αυτά όμως, σύγχρονες εξελίξεις και εφαρμογές έχουν δημιουργήσει νέες ανάγκες τις οποίες οι παραδοσιακές βάσεις δεδομένων δεν μπορούν να αντιμετωπίσουν. Οι προκλήσεις αφορούν στα δεδομένα και την αναπαράστασή τους. Πιο συγκεκριμένα εφαρμογές όπως δίκτυα αισθητήρων, κάμερες επιτήρησης, καθαρισμός δεδομένων, οικονομικές αποφάσεις, έχουν να κάνουν με δεδομένα που δεν παίρνουν διακριτές τιμές με την παραδοσιακή έννοια του όρου και παρουσιάζουν διακυμάνσεις και αβεβαιότητα. Αυτές και άλλες εφαρμογές έχουν στρέψει την προσοχή της ερευνητικής κοινότητας στις «αβέβαιες βάσεις δεδομένων» με σκοπό την αντιμετώπιση των προκλήσεων και την εύρεση λύσεων ώστε να καταστεί δυνατή η χρήση τους.

Σε αυτό το σημείο πρέπει να αναφερθεί ότι στην παρούσα εργασία δεν διαχωρίζουμε μεταξύ αβέβαιων και ανακριβών δεδομένων. Με τον όρο ανακρίβεια, για παράδειγμα, εννοούμε ότι αν ένας αισθητήρας πάρει μια μέτρηση της έντασης του φωτός αυτή μπορεί να μην είναι απόλυτα ακριβής. Από την άλλη, με τον όρο αβεβαιότητα εννοούμε ότι η ένταση που μετρήθηκε μπορεί να είναι σωστή, μπορεί και να μην είναι, αλλά δεν

υπάρχει τρόπος να γνωρίζουμε την αλήθεια αυτής της μέτρησης (Sunter, 1969). Από εδώ και στο εξής δε θα γίνεται διάκριση ανάμεσα στις δύο έννοιες και θα αντιμετωπίζονται και οι δύο ως αβέβαιες μετρήσεις.

Οι προκλήσεις και τα προβλήματα των πιθανοτικών βάσεων δεδομένων όπως έχουν περιγραφεί στη βιβλιογραφία έχουν να κάνουν με την αβέβαιη σημασιολογία (uncertain semantics), την μοντελοποίηση των δεδομένων, την επεξεργασία επερωτήσεων (query evaluation) και τη διαφοροποίηση των εφαρμογών εξόρυξης (mining) ώστε να λειτουργούν και με αβέβαια δεδομένα. Διάφορα προβλήματα παρουσιάζονται με περισσότερες λεπτομέρειες στα (Aggarwal, Managing and Mining Uncertain Data, 2009) και (Sarma A. D., Benjelloun, Halevy, & Widom, 2006).

Συνολικά, οι σημαντικότεροι τομείς (Nilesh Dalvi, 2007) στους οποίους έχει επικεντρωθεί η έρευνα (Aggarwal & Yu, A survey of uncertain data algorithms, 2007), αφορούν τη:

- *Μοντελοποίηση αβέβαιων δεδομένων*
  Ένα σημαντικό θέμα είναι το πως να μοντελοποιηθούν τα δεδομένα ώστε παρόλη την πολυπλοκότητα να είναι δυνατή η επεξεργασία και η χρήση της βάσης.

- Διαχείριση αβέβαιων δεδομένων
  Οι εφαρμογές βάσεων δεδομένων πρέπει να τροποποιηθούν ώστε να μπορούν να υποστηρίζουν αβέβαια δεδομένα, όπως επεξεργασία επερωτήσεων (queries) και σελιδοποίηση (indexing).

- *Εξόρυξη αβέβαιων δεδομένων*
  Η εξόρυξη αβέβαιων δεδομένων δεν μπορεί πια να θεωρείται ασφαλής. Είναι αναγκαία η ανάπτυξη τεχνικών εξόρυξης δεδομένων που θα μπορούν να λαμβάνουν υπόψη τους την αβεβαιότητα και να εξάγουν ασφαλή αποτελέσματα.

Αν συνολικά θέλαμε να κάνουμε μια αντιστοίχιση της ανάγκης που παρουσιάζεται στις βάσεις δεδομένων, αυτή είναι παρόμοια με τις ανάγκες που αντιμετώπισε η επιστήμη την Τεχνητής Νοημοσύνης, ότι μεταμόρφωσε τη λογική σε λογική με πιθανότητα (Heckerman, 2002). Έτσι και οι βάσεις δεδομένων πρέπει να περάσουν από τα δεδομένα στα πιθανοτικά δεδομένα.

Το υπόλοιπο αυτής της εργασίας είναι οργανωμένο ως εξής: Στο επόμενο κεφάλαιο παρουσιάζονται μια σειρά από πρακτικές εφαρμογές και χρήσεις πιθανοτικών βάσεων δεδομένων. Στα κεφάλαια 3, 4, 5 παρουσιάζουμε τις τεχνικές, τους αλγορίθμους και τις μεθοδολογίες για τις σημαντικότερες προαναφερθείσες προκλήσεις, τη μοντελοποίηση, διαχείριση και εξόρυξη αβέβαιων δεδομένων. Στη συνέχεια, στο κεφάλαιο **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.** περιγράφει τα τρέχοντα ερευνητικά εγχειρήματα (projects), την πρόοδο, τους στόχους και τις προκλήσεις. Το κεφάλαιο 6 παρουσιάζει τα ανοιχτά ζητήματα του τομέα των πιθανοτικών βάσεων δεδομένων που δεν έχουν αντιμετωπιστεί επιτυχώς μέχρι τις μέρες μας. Τέλος η εργασία κλείνει με τα συμπεράσματα του κεφαλαίου 7.

2. Εφαρμογές

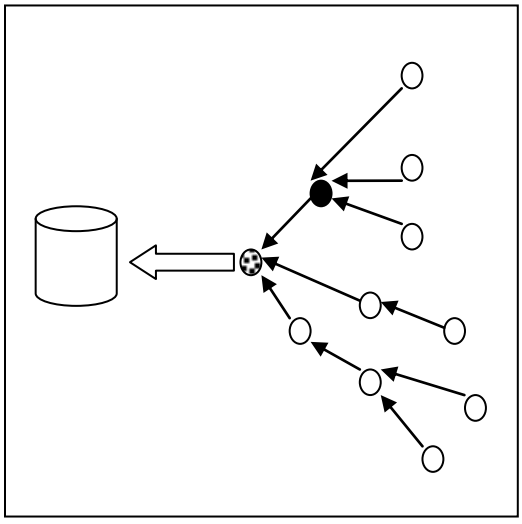
Είναι σημαντικό να παρουσιαστούν μια σειρά από εφαρμογές που αφορούν στη συλλογή και επεξεργασία αβέβαιων δεδομένων ώστε ο αναγνώστης να έχει μια εποπτική άποψη για τα κίνητρα του ενδιαφέροντος της επιστημονικής κοινότητας για το θέμα. Στα παραδείγματα που ακολουθούν θα δούμε πως παραδοσιακές εφαρμογές καθώς και νέες συνδέονται στενά με τα αβέβαια δεδομένα και γιατί είναι επιτακτική η υπερσκέλιση των εμποδίων.

Προϋπάρχουσες εφαρμογές

Η δημιουργία αναπαραστάσεων (mappings) δεδομένων και σχηματικών αναπαραστάσεων (schema) δεν υπήρξε πότε ακριβής ή απλή (Halevy, 2007). Η δημιουργία τους σε ευρύτερες κλίμακες κρίνει πολύ σημαντική την αντιμετώπιση της αβεβαιότητας. Στο (Kohler, 2006) περιγράφεται η δυσκολία ενοποίησης των βάσεων δεδομένων που περιέχουν «φυσικά» δεδομένα με σημαντικότερο πρόβλημα αυτό της αβεβαιότητας. Οι βάσεις είναι αβέβαιες και τα δεδομένα είναι πολύ δύσκολο να ενοποιηθούν με μια εποπτική μέθοδο. Ως τώρα οι επιστήμονες τα επεξεργάζονταν «με το χέρι» και έτρεχαν αλλεπάλληλες επερωτήσεις τροφοδοτώντας τα αποτελέσματα των προηγούμενων στις επόμενες, αφότου τα κατέτασσαν σε σειρά. Για να το κάνει κανείς αυτό αυτόματα θα πρέπει να έχει μια πιθανοτική αναπαράσταση της βάσης ώστε να μπορεί να ταξινομήσει τα δεδομένα. Για να το απλοποιήσουμε με ένα παράδειγμα ας θεωρήσουμε μια βάση πρωτεϊνικών δεδομένων. Αυτή η βάση περιέχει εγγραφές για της συναρτήσεις/αναπαραστάσεις των πρωτεϊνών (annotations, πχ DNA annotation είναι η αναπαράσταση της αλυσίδας του DNA). Μερικές από αυτές μπορούν να εξαχθούν στο εργαστήριο πειραματικά, οπότε μπορούμε να πούμε ότι είναι ντετερμινιστικές. Από την άλλη, υπάρχουν

και αναπαραστάσεις πρωτεϊνών που εξάγονται με υπολογιστικές μεθόδους (στατιστικές Markov, σειριακή ομοιότητα (sequence similarity)). Αυτές οι αναπαραστάσεις δεν είναι εξίσου αξιόπιστες και ενέχουν αβεβαιότητα. Στο σημείο αυτή η μοντελοποίηση πιθανοτικών δεδομένων θα έπρεπε να είναι σε θέση να δώσει λύσεις.

Παραδοσιακά, οι αισθητήρες είναι οι κατεξοχήν συσκευές που συλλέγουν δεδομένα τα οποία μπορεί να περιέχουν σφάλματα. Προβλήματα στον εξοπλισμό μέτρησης, εξάντληση της μπαταρίας, εμπόδια εκπομπής σήματος, διαλείψεις και σκεδάσεις, φυσικές αποκλίσεις μπορούν να παρεισφρήσουν και να εισάγουν αβεβαιότητα στις μετρήσεις. Οι χρήσεις αυτών των δεδομένων πρέπει να γίνεται προσεκτικά. Μέχρι τώρα, οι προτεινόμενες λύσεις για δίκτυα αισθητήρων είχαν να κάνουν με τη σύνοψη των δεδομένων πριν φτάσουν στον κόμβο συγκέντρωσης (J. Considine, 2004), (Manjhi, Nath, & Gibbons, 2005), (Σχήμα 1: **Υπάρχουσες λύσεις συγκέντρωσης δεδομένων σε δίκτυα αισθητήρων.**). Ωστόσο, αυτές λύσεις δεν αντιμετωπίζουν αποτελεσματικά τις αβέβαιες μετρήσεις και αν για παράδειγμα οι μετρήσεις έχουν μεγάλες αποκλίσεις μπορεί η βάση να εξάγει το συμπέρασμα ότι το δάσος όπου βρίσκονται οι αισθητήρες καίγεται χωρίς αυτό να είναι ακριβές. Η πιθανοτική



Σχήμα 1: Υπάρχουσες λύσεις συγκέντρωσης δεδομένων σε δίκτυα αισθητήρων.

Οι κόμβοι ● συγκεντρώνουν και ενοποιούν δεδομένα.

Ο κεντρικός κόμβος 👁 τα επεξεργάζεται πριν τοποθετηθούν στη βάση δεδομένων.

μοντελοποίηση δεδομένων αισθητήρων παρουσιάζεται στα (Deshpande, Guestrin, Madden, Hellerstein, & Hong, 2005). (Zhang, Lin, Pei, Zhang, & Fraser, 2008)

Άλλες φορές, η αβεβαιότητα στα δεδομένα μπορεί να είναι επιθυμητή. Για παράδειγμα, δημογραφικά δεδομένα όπως και κάποιες περιπτώσεις γεωγραφικών δεδομένων έχουν μερική ενοποίηση για λόγους ασφαλείας ώστε να μην είναι δυνατός ο απόλυτος εντοπισμός χρηστών. Το ίδιο συμβαίνει και σε δεδομένα διατήρησης ασφάλειας πληροφοριών στις οποίες προστίθενται «λανθασμένα» δεδομένα ώστε να παραμείνουν ασφαλείς οι πληροφορίες. Πρόσφατα στο (Aggarwal, On Unifying Privacy and Uncertain Data, 2008) προτείνονται μια σειρά από μοντέλα ασφαλείας που με τη μεταποίηση δεδομένων δημιουργούντα αβέβαια δεδομένα «φιλικά» προς τις τεχνικές επεξεργασίας αβέβαιων βάσεων δεδομένων.

Νέες εφαρμογές

Όσον αφόρα σε νέες εφαρμογές, αυτές πολλές φορές τείνουν να δημιουργούν αβέβαια δεδομένα. Τέτοιες είναι το Google Base (Google Base), Flickr (Flickr), το παιχνίδι ESP (Dabbish & Dabbish, 2004), οι οποίες γεννούν τεράστιες ποσότητες δεδομένων που συχνά παρουσιάζουν αποκλίσεις. Αλλά και συστήματα εξαγωγής πληροφοριών από κείμενο (Πίνακας 1) τείνουν να γεννούν δεδομένα με μεγάλες αποκλίσεις (Doan, και συν., 2006), (T.S. Jayram, 2006). Για παράδειγμα όταν σε μια αθλητική ιστοσελίδα αναζητήσουμε «πόσα κείμενα έχουν θέμα το άθλημα Χ», οι απαντήσεις θα ποικίλουν και θα πρέπει να ταξινομηθούν με βάση τη σχέση τους με την ερώτηση. Τι γίνεται όμως αν έχουν συνδυαστικές ερωτήσεις; Σε αυτό το σημείο τα πιθανοτικά μοντέλα δεδομένων θα πρέπει να μπορούν να δώσουν λύση.

Από την άλλη, δεδομένα XML πολλές φορές δημιουργούν μια σειρά από ειδικές προκλήσεις. Ο λόγος είναι ότι οι XML πληροφορίες είναι δομημένες,

με αποτέλεσμα οι πιθανότητες που θα πρέπει να αποδοθούν για την αναπαράστασή τους να πρέπει να ακολουθήσουν δομή δένδρου με κόμβους και ακμές που έχουν συγκεκριμένη πιθανότητα. Επιπλέον, πιθανότητες στοιχείων μπορούν να συμβούν σε διάφορα επίπεδα του δέντρου και φωλιασμένες πιθανότητες σε υποδένδρα πρέπει να ληφθούν υπόψη. (M. Keulen, 2005). Άλλο ένα θέμα με τα XML δεδομένα είναι ότι οι πιθανότητες μιας αλυσίδας κόμβων γονέων-παιδιών συνδέονται επίσης πιθανοτικά. Το κεφάλαιο 3 πραγματεύεται τις προτάσεις που έχουν γίνει για το σχετικό θέμα.

Τέλος, ο «φυσικός κόσμος» κατεξοχήν παράγει δεδομένα που είναι πιθανοτικά αβέβαια. Τέτοια είναι δεδομένα από κάμερες, RFIDs. Τα (Zhao, 2006) και (M & al, 2007) προτείνουν να γίνουν αυτά τα δεδομένα διαθέσιμα μέσα από το World Wide Sensor Web. Για την εύστοχη χρήση μιας τόσο ευρείας και αβέβαιης κλίμακας δεδομένων είναι πολύ σημαντικό να ξεπεραστούν όλες οι «δυσκολίες της αβεβαιότητας».

ID	House #	Area	City	Prob
1	52	Goregaon West	Mumbai	10%
1	52-A	Goregaon West	Mumbai	50%
1	52	Goregaon West	West Mumbai	20%
1	52-A	Goregaon West	West Mumbai	30%
2	…	…	…	…

Πίνακας 1: Από (Sarawagi, 2006)

Αυτές είναι κάποιες χρήσεις που αν προστεθούν σε αρκετές άλλες γίνεται κατανοητό πόσο ευρύ είναι το πεδίο των πιθανοτικών δεδομένων και γιατί έχει λάβει τόσο μεγάλης προσοχής από την ερευνητική κοινότητα τελευταία.

3. Μοντελοποίηση αβέβαιων δεδομένων

Στο κεφάλαιο 1, αναφέρθηκαν οι σημαντικότερες προκλήσεις. Η πιο βασική είναι αυτή της αναπαράστασης των δεδομένων ώστε στη συνέχεια να είναι δυνατή η επεξεργασία τους. Η βιβλιογραφία στον τομέα είναι πλούσια και πρώιμη δουλειά είχε ήδη ξεκινήσει από τα τέλη της δεκαετίας του 1970. Το βασικό πρόβλημα είναι η απάντηση στην ερώτηση μια γραμμή (tuple) υπάρχει σε μια σχέση ή η σιγουριά για την ύπαρξή της σε μια σχέση (που θα προκύψει από απάντηση επερώτησης). Φυσικά, πρόσφατα η δουλειά στο θέμα έχει παράγει μεγάλο αριθμό από μοντέλα που το καθένα έχει τις δυνατότητες και της αδυναμίες του.

Υπάρχουν πολλοί τρόποι κατηγοριοποίησης των μοντέλων που έχουν προταθεί ανάλογα με τα χαρακτηριστικά τους. Ενδεικτικά θα μπορούσαν να χωριστούν με βάση τη

- Φύση της αβεβαιότητας (πιθανοτική, μη-πιθανοτική)
- Είδος ΒΔ (σχετική (relational), XML)
- Πολυπλοκότητα της αβεβαιότητας
    - Λεπτότητα (granularity) της αβεβαιότητας
    - Χειρισμό συσχετισμών
    - Χειρισμό χαμένης πληροφορίας
    - Είδη αβεβαιότητας που υποστηρίζεται

Επίσης, μπορούν να χωριστούν ανάλογα με τις τεχνικές σε

- Ποιοτικά μοντέλα
    - ΚΕΝΩΝ (NULL) τιμών

Οι ΚΕΝΕΣ τιμές είναι ένα τρόπος σύλληψης της αβεβαιότητας με λογική τριών τιμών (Αλήθεια, Ψέματα, Ίσως) (T,F,M)

- *Οριστικά, αόριστα, πιθανά* (Lim & Shekhar, 1996)

  Κινείται στο ίδιο μήκος κύματος με το προηγούμενο ποιοτικό μοντέλο αντιμετωπίζοντας την ύπαρξη ή όχι μιας γραμμής σε απάντηση επερώτησης με τρεις δυνατές εκφάνσεις.

Τα ποιοτικά μοντέλα μπορεί να πει κανείς είναι τα πιο πρώιμα του τομέα καθώς η έλλειψη του υπόβαθρου ανάγκαζε του θιασώτες τους στη λογική προσέγγιση του προβλήματος χωρίς να λαμβάνουν υπόψη τις υπάρχουσες πιθανοτικές μαθηματικές σχέσεις των δεδομένων.

- Ποσοτικά μοντέλα

  - Πιθανοτικά

    - ❑ Αβεβαιότητα γραμμής

      Η ύπαρξη μιας γραμμής στη σχέση είναι αβέβαιη και μοντελοποιείται σαν μια πιθανότητα που σχετίζεται με ολόκληρη τη γραμμή.

    - ❑ Αβεβαιότητα ιδιότητας

      Η τιμή κάποιας ιδιότητας μια γραμμής (tuple) δεν είναι απολύτως γνωστή, και μοντελοποιείται σαν μια σειρά από πιθανές τιμές που σχετίζονται μεταξύ τους με πιθανότητες.

    - ❑ Αβεβαιότητα αντικειμένου (object)

Προσεγγίζει την πιθανότητα εμφάνισης ενός αντικειμένου σε ένα σύνολο ως μια πιθανότητα που ανατίθεται στο αντικείμενο αυτό καθ' εαυτό.

- ❑ Αβεβαιότητα συνόλου (group)

  Η «διασπορά» ενός συνόλου αντιμετωπίζεται ως το ποσοστό των αντικειμένων στο σύνολο επί του συνόλου αυτόν.

- Αποδεικτικά (evidence-oriented) (Lee, 1992)

  Παλαιότερα μοντέλα που βασίζονται στην Θεωρία της Απόδειξης των Dempster-Shafer εφαρμόζοντάς την στα πιθανοτικά δεδομένα.

- Περίτεχνα (fuzzy)

  Τα μοντέλα αυτά περιέχουν περίτεχνες οντότητες, περίτεχνες σχεσιακές συνδέσεις, περίτεχνες ενώσεις, κλπ που χρησιμοποιούνται για να αντιμετωπίσουν τις ανακρίβειες στα δεδομένα (Galindo, Urrutia, & Piattini).

Αφού αναφέραμε τις σημαντικότερες κατηγορίες μοντέλων θα εστιάσουμε στα σημαντικότερα από αυτά, παρουσιάζοντάς τα. Σημειωτέον ότι τα περισσότερα ανήκουν στην πιθανοτική κατηγορία καθώς είναι πιο ρεαλιστικά και υλοποιήσιμα.

- Γενικό

  Στη γενική περίπτωση πιθανοτικών βάσεων δεδομένων, οι τιμές των δεδομένων συσχετίζονται. Αν η πιθανότητα εμφάνισης μια τιμής R_i είναι $P(R_i)$ τότε η πιθανότητα εμφάνισης της είναι $\sum_{1\leq i\leq n} P(R_i) \leq 1$. Ένα σύνολο από αποκλειστικά δεδομένα ονομάζεται γενετήριος κανόνας (generation rule) C. Η πιθανότητα εμφάνισης του C είναι $\sum_{R\in C} P(R)$. Ο γενετήριος κανόνας αποτελείται από μια γραμμή και διαφορετικοί κανόνες είναι ανεξάρτητοι μεταξύ τους. Δεδομένου

ενός συνόλου κανόνων $G_D = \{R_1, R_2, \dots, R_n\}$, ένας πιθανός κόσμος[1] ορίζεται σαν ένα στοιχείο του $\prod_{R \in G'} R$ όπου G' είναι ένα υποσύνολο του G_D και περιέχει κάθε γενετήριο κανόνα R τέτοιο ώστε $P(R) = 1$. Αν |R| είναι ο αριθμός των κανόνων στο R, τότε ο αριθμός όλων των πιθανών κόσμων σε σχέση με το G_D είναι $|W| = \prod_{R \in G_D, P(R)=1} |R| \prod_{R \in G_D, P(R)<1} (|R| + 1)$. Η πιθανότητα εμφάνισης ενός πιθανού κόσμου W είναι

$$P(W) = \prod_{R \in G_D, R \cap W \neq \emptyset} P(R \cap W) \prod_{R \in G_D, R \cap W = \emptyset} (1 - P(R))$$

όπου $P(R \cap W)$ είναι η πιθανότητα εμφάνισης ενός στοιχείου και στο R και στο W.

Σαν παράδειγμα, ο Πίνακας 2 παρουσιάζει τις μετρήσεις για τη ταχύτητα συγκεκριμένων οχημάτων από διάφορες συσκευές μέτρησης ταχύτητας ενός οχήματος σε συγκεκριμένο τόπο και ώρα και με συγκεκριμένη πιθανότητα η μέτρηση να είναι σωστή.

Πίνακας 2: **Μέτρηση ταχύτητας οχήματος (παρόμοιος με (Zhang, Lin, Pei, Zhang, & Fraser, 2008)**

Αριθμός	Μέτρηση	Ώρα	Τόπος	Όχημα	Ταχύτητα	Πιθανότητα
A1	M1	2:00μμ	T1	YAP2345	120	0.7
A2	M2	2:00μμ	T2	YAP2345	150	0.2
A3	M3	3:45μμ	T17	HBA1234	170	0.9

Πίνακας 2: Μέτρηση ταχύτητας οχήματος (παρόμοιος με (Zhang, Lin, Pei, Zhang, & Fraser, 2008))

Οι γραμμές A1 και A2 είναι αμοιβαία αποκλειστικές καθώς δεν μπορούν να ισχύουν ταυτόχρονα. Ο κανόνας A3 είναι ανεξάρτητος μόνο ως προς το σύνολο $\{A_1, A_2\}$ και υπάρχουν 6 πιθανοί κόσμοι που φαίνονται στον Πίνακας 3.

Κόσμος	Πιθανότητα

[1] Αν ήθελε κανείς να δώσει μια πιο εποπτική ιδέα για το τι είναι ένας πιθανός κόσμος, θα μπορούσε να πει ότι είναι όλες οι δυνατές γραμμές που μπορούν να απαρτίζουν μια βάση δεδομένων.

$\{\emptyset\}$	0.01
$\{A_1\}$	0.07
$\{A_2\}$	0.02
$\{A_3\}$	0.09
$\{A_1, A_3\}$	0.63
$\{A_2, A_3\}$	0.18

Πίνακας 3: Πιθανοί κόσμοι (παρόμοιος με (Zhang, Lin, Pei, Zhang, & Fraser, 2008))

- Ανεξάρτητο μοντέλο

  Μια τεχνική αναπαράστασης είναι να αντιστοιχίσουμε την ύπαρξη μια γραμμής στη ΒΔ με μια πιθανότητα $P(R)$ ($P(R) > 0$). Ένας πιθανός κόσμος W ορίζεται και πάλι ως το υποσύνολο του D για το οποίο ένα στοιχείο $R \in D$ περιλαμβάνεται στο W με πιθανότητα P(R) = 1. Είναι φανερό ότι η πιθανότητα ύπαρξης ενός κόσμου W είναι $P(W) = \prod_{R \in W} P(R) \prod_{R \in W}(1 - P(R))$. Στην ειδική περίπτωση, αν W είναι το σύνολο των πιθανών κόσμων και N είναι ο αριθμός των στοιχείων με πιθανότητα εμφάνισης μικρότερη του 1, τότε $|W| = 2^N$. Η συνολική πιθανότητα εμφάνισης όλων των πιθανών κόσμων είναι $\sum_{W \in W} P(W) = 1$.

- ProbView (Lakshmanan, Leone, Ross, & Subrahmanian, 1997)

  Στο ίδιο μήκος κύματος κινείται και το Probview μοντέλο αλλά χρησιμοποιώντας πιθανοτικό ?-πίνακα. Αντιμετωπίζει τη ΒΔ μοντελοποιώντας την πιθανότητα μιας τιμής μιας ιδιότητας/στοιχείου κάποιας γραμμής. Η τιμή του στοιχείου είναι πρακτικά ένα διαζευκτικό «ή» σε μια σειρά από πιθανές τιμές με αντίστοιχες τιμές εμπιστοσύνης. Στη συνέχεια η εμπιστοσύνη στοιχείου γίνεται εμπιστοσύνη γραμμής (tuple). Η αρχικοποίηση της ΒΔ γίνεται παίρνοντας τις πιθανές τιμές του κάθε στοιχείου ανεξάρτητα και οι «πιθανοί κόσμοι» εξάγονται με «άνω» και «κάτω» φράγματα στις τιμές εμπιστοσύνης. Ο υπολογισμός των πιθανοτήτων γίνεται με συναρτήσεις ορισμένες από το χρήστη, ενώ

επίσης οι παραδοσιακή σχεσιακή άλγεβρα διαφοροποιείται για να αντιμετωπιστούν τα πιθανοτικά φράγματα.

- (Rolleke & FuhrT., 1997)
  Το μοντέλο του (Rolleke & FuhrT., 1997) είναι προγενέστερο και δεν βασίζεται σε κανονική μορφή (normal form (NF2)). Οι εγγραφές μιας σχέσης έχουν πιθανοτικά βάρη. Πιθανοτικές τιμές στοιχείων χωρίζονται σε υπο-σχέσεις. Επίσης, το μοντέλο προτείνει μια πιθανοτική σχεσιακή άλγεβρα που είναι επέκταση της κλασσικής σχεσιακής άλγεβρας. Κάθε γραμμή (tuple) t της ΒΔ έχει τρία χαρακτηριστικά, την τιμή των στοιχείων, την έκφραση γεγονότος (event expression) τ.η και τη πιθανότητα του γεγονότος (event probability) τ.β. Ο Πίνακας 4, σύμφωνα με το μοντέλο, δείχνει πως ένα βιβλίο έχει ατομικά (atomic) χαρακτηριστικά *BNO*, *YEAR* και χαρακτηριστικά PRICE, INDEX, AUTHOR που μοντελοποιούνται σαν υπο-σχέσεις. Αυτονόητο είναι ότι οι τιμές PRICE, INDEX, AUTHOR δεν είναι ακριβείς και περιέχουν σφάλματα. Υπάρχουν διάφορα είδη πιθανοτικών σχέσεων όπως ντετερμινιστικές (deterministic), ανεξάρτητες (independent), χωριστές (disjoint) και εξαρτημένες (dependent). Αν θελήσουμε να επανέλθουμε και πάλι στον Πίνακας 4, τότε η υπο-σχέση για την PRICE είναι χωριστή, πράγμα που σημαίνει ότι ένα και μόνο ένα γεγονός μεταξύ *BEP*1 και *BEP*2 μπορεί να είναι αλήθεια τη φορά. Η υπο-σχέση *INDEX είναι ανεξάρτητη καθώς και το BEI1 και το BEI2 μπορούν να αληθεύουν. Τέλος, το AUTHOR είναι ντετερμινιστικό γιατί δύο τιμές του για το NAME έχουν την ίδια πιθανότητα με την γραμμή στην οποία ανήκουν (1). Το Ανεξάρτητο μοντέλο που παρουσιάσαμε παραπάνω μπορεί να μοντελοποιήσει και τέτοιες NF2 σχέσεις.*

BOOK														
η	β	BNO	YEAR	PRICE			INDEX			AUTHOR		
				η	β	VAL	η	β	TERM	η	β	NAME
$B\hat{E}$	1.0	1	92	$B\hat{E}P1$ $B\hat{E}P2$	0.6 0.4	30 25	$B\hat{E}I1$ $B\hat{E}I2$	0.9 0.8	IR DB	$B\hat{E}A\hat{E}$ $B\hat{E}A\hat{E}$	1.0 1.0	Smith Jones
$B\hat{E}$	1.0	2	93	$B\hat{E}P3$	1.0	29	$B\hat{E}I3$ $B\hat{E}I4$	0.9 0.8	AI DB	$B\hat{E}A\hat{E}$	1.0	Miller
$B\hat{E}$	1.0	3	92	$B\hat{E}P4$ $B\hat{E}P5$	0.7 0.3	28 25	$B\hat{E}I5$	0.9	DB	$B\hat{E}A\hat{E}$	1.0	Jones
$B\hat{E}$	1.0	4	90	$B\hat{E}P6$ $B\hat{E}P7$	0.5 0.5	32 28	$B\hat{E}I6$	0.9	DB	$B\hat{E}A\hat{E}$	1.0	Jones

Πίνακας 4: Από (Zhang, Lin, Pei, Zhang, & Fraser, 2008)

- (Sarma, Benjelloun, Halevy, & Widom, 2005)

  Το (Sarma, Benjelloun, Halevy, & Widom, 2005) βασίζεται στις τεχνικές καταγωγής (lineage) για να μοντελοποιήσουν την αβεβαιότητα (Sarma A. D., Benjelloun, Halevy, & Widom, 2005). Καταγωγή, όπως σημαίνει και η λέξη, είναι η ιδιότητα κάποιου στοιχείου που διαθέτει πληροφορία για το στοιχείο από το οποίο προέρχεται. Ομοίως, το (O. Benjelloun A. D., 2006) παρουσιάζει ένα μοντέλο για ΒΔ (ULDBs (Uncertainty-Lineage Databases)) που περιέχουν δεδομένα που περιέχουν πληροφορία για την καταγωγή τους. Το βασικά χαρακτηριστικά του μοντέλου αυτού που επεκτείνει τα κλασσικά σχεσιακά μοντέλα είναι

  1) δίνει εναλλακτικές για τον εντοπισμό της αβεβαιότητας στα περιεχόμενα μια γραμμής
  2) εισάγει πιθανούς όρους «?» που αναπαριστούν την αβεβαιότητα μιας γραμμής
  3) εισάγει τιμές εμπιστοσύνης για την αβεβαιότητα
  4) εξάγει πληροφορίες καταγωγής από εναλλακτικές γραμμές.

  Αξίζει να σημειωθεί ότι το μοντέλο καταγωγής μοιάζει με το ανεξάρτητο όσον αφορά την αβεβαιότητα στοιχείου.

Παράδειγμα:

Ο Πίνακας 5 δείχνει ένα παράδειγμα μιας Uncertainty Lineage Database (ULDB) για οχήματα ανάλογα με τον αριθμό πινακίδας και τη μέγιστη ταχύτητα τους. Η γραμμή για το όχημα O1 μπορεί να πάρει δύο πιθανές τιμές με την αναγραφόμενη τιμή εμπιστοσύνης. Το O2 υπάρχει στον πίνακα με πιθανότητα 90%.

Όχημα	(Όχημα, Ταχύτητα)
O1	(YAP1235, 150): 0.7 // (HAP1235, 150): 0.2
O2	(HAB2568, 190): 0.9 ?

Πίνακας 5: Όχημα και χαρακτηριστικά του

Ο Πίνακας 6 ασχολείται με τον αριθμό πινακίδας και το όνομα του οδηγού. Αν ενώσουμε (join) τους δύο πίνακες και εστιάσουμε στην πινακίδα θα πάρουμε μόνο μια γραμμή, αυτή του Γιάννη. Ας ονομάσουμε αυτή τη γραμμή O5. Η τεχνική καταγωγής του μοντέλου που περιγράφουμε μπορεί να εντοπίσει την καταγωγή του O5 από τις εναλλακτικές γραμμές O1 και O3 μέσω μια συνάρτησης λ, λ(O5, 1) = ((O1, 2), (O3, 1)). Η παραπάνω συνάρτηση σημαίνει ότι η πρώτη εναλλακτική της O5 μπορεί να προκύψει από τη δεύτερη εναλλακτική της O1 και την πρώτη της O3.

Όχημα	Αρ. πινακ. Ταχύτητα
O3	HAP1235, Γιάννης
O4	IKE2478, Ελένη

Πίνακας 6: Πινακίδα και οδηγός

- (Suciu & Dalvi, 2005)

  Στο (Suciu & Dalvi, 2005), οι Dalvi και Suciu εξερευνούν παρατεταμένες αποτιμήσεις επερωτήσεων θεωρώντας τις περιπτώσεις του στατιστικού συσχετισμού και της πιθανοτικής γέννησης των δεδομένων από την πηγή τους. Για να το πετύχουν αυτό χρησιμοποιούν μια επέκταση του μοντέλου Local As View (LAV). Επίσης υπολογίζουν την πιθανότητα τιμών γραμμών με τη χρήση τελεστών. Ωστόσο, η μέθοδος που προτείνουν μπορεί σε μερικές

περιπτώσεις να οδηγήσει σε ανακριβή και πολύπλοκα αποτελέσματα. Γι' αυτό το λόγο εισάγουν την έννοια των σχεδίων ασφαλών επερωτήσεων (safe query plans) με τη χρήση ενός ειδικού αλγορίθμου που μπορεί να τα αναπτύξει. Το μειονέκτημα του αλγόριθμου είναι ότι δεν εγγυάται ότι θα είναι σε θέσει να υπολογίσει το πλάνο για όλες τις περιπτώσεις. Επίσης, το (Suciu & Dalvi, 2005) προτείνει ευριστικά πλάνα και προσεγγίσεις για την περίπτωση που η πολυπλοκότητα της επερώτησης είναι #P-complete[2]. Τέλος, εξηγεί πως πολύπλοκες επερωτήσεις μπορεί να αποτιμηθούν σπάζοντας και ξαναγράφοντάς τες σε μικρότερες.

- (Sen & Deshpande, 2007)

  Οι Sen και Deshpande αυτό που χρησιμοποίησαν είναι ένα πιθανοτικό γραφικό μοντέλο (Pearl, 1988) για να μπορέσουν να υπολογίσουν επερωτήσεις σε αβέβαια δεδομένα με γενικές μορφές συσχέτισης. Ασχολούνται με τη ανεξαρτησία και τον αμοιβαίο αποκλεισμό (η ύπαρξη μια γραμμής σημαίνει τη μη ύπαρξη της άλλης) και την υψηλή θετική συσχέτιση ανάμεσα στις δύο γραμμές. Κάθε γραμμή συνδέεται με μια λογική (boolean) τυχαία μεταβλητή X_i που μπορεί να είναι λάθος ή σωστή. Στο πιθανοτικό γραφικό μοντέλο που προτείνεται οι κόμβοι αντιπροσωπεύουν τυχαίες μεταβλητές και οι ακμές αντιπροσωπεύουν συσχετίσεις μεταξύ των μεταβλητών. Έτσι μοντελοποιούνται διαφόρων ειδών συσχετίσεις όπως αμοιβαίος αποκλεισμός, απόλυτη ανεξαρτησία και θετικός συσχετισμός. Με αυτόν τον τρόπο η αποτίμηση των επερωτήσεων με συσχετισμούς μετατρέπεται σε ένα αντίστοιχο πρόβλημα του πιθανοτικού γραφικού μοντέλου και μπορεί να λυθεί με γνωστές μεθόδους και αλγορίθμους.

[2] Ένα πρόβλημα είναι #P-complete όταν και μόνο όταν ανήκει στα #P, και κάθε πρόβλημα στο #P μπορεί να αναχθεί σε αυτό σε πολυωνυμικό χρόνο. Από την άλλη, #P είναι μια κλάση προβλημάτων συναρτήσεων της μορφής $f(x)$ όπου f είναι μονοπάτια διάσχισης μιας NP-μηχανής Turing.

- XML δεδομένα

  Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, τα XML δεδομένα παρουσιάζουν μια σειρά από προκλήσεις λόγω της εξαρτημένης δομής τους. Το (Nierman & Jagadish, 2002) προτείνει τη μοντελοποίηση κάποιων κλάσεων εξάρτησης που διαθέτουν τα δεδομένα (πχ αμοιβαίος αποκλεισμός (mutual exclusion) για κάποια XML tags) που είναι ευκολότερα να μοντελοποιηθούν. Επίσης παρουσιάζει μια σειρά τεχνικών για επερωτήσεις σε XML δεδομένα και επιχειρεί πιθανοτική κατασκευή δεδομένων με χρήση πιθανοτικών XML δένδρων που επιχειρούν να μιμηθούν τη δενδροειδή δομή της γλώσσας. Τα δένδρα αυτά έχουν κόμβους που δείχνουν τις μεταβάσεις και κόμβους που δείχνουν τις πιθανότητες μετάβασης. Ο υπολογισμός μιας επερώτησης γίνεται διατρέχοντας το δέντρο αναδρομικά και κατασκευάζοντας όλους τους πιθανούς κόσμους. Στη συνέχεια η επερώτηση αποτιμάται σε κάθε κόσμο.

Εκτός από αυτές που αναφέρθηκαν εδώ, υπάρχει και μια σειρά άλλων προσπαθειών, αλγορίθμων και μοντέλων που προτάθηκαν στη βιβλιογραφία και δεν παρουσιάζονται γιατί ξεφεύγουν από τα όρια της κατηγοριοποίησης και παρουσίασης των σημαντικότερων από αυτές.

4. Επεξεργασία αβέβαιων δεδομένων

Ίσως το πιο σημαντικό κομμάτι της έρευνας είναι η επεξεργασία των πιθανοτικών δεδομένων. Όλα τα μοντέλα που προαναφέρθηκαν εξυπηρετούν ένα σκοπό και αυτός δεν είναι άλλος από το να καταστήσουν δυνατή την επεξεργασία των δεδομένων αυτών όπως γίνεται και σε παραδοσιακές ΒΔ. Ο όρος επεξεργασία αναφέρεται σε επεξεργασία επερωτήσεων, σελιδοποίηση (indexing), ενώσεις. Οι σημαντικότερες από αυτές τις τεχνικές παρουσιάζονται εδώ συνοπτικά όπως εμφανίζονται στη βιβλιογραφία.

4.1 Επεξεργασία επερωτήσεων

Πρώτα από όλα στις βάσεις δεδομένων είναι η αποτίμηση επερωτήσεων. Στις παραδοσιακές βάσεις οι επερωτήσεις δεν είναι παρά απλές ερωτήσεις γλώσσας SQL, αλλά όπως έχει ήδη εννοηθεί αυτό δεν είναι το ίδιο απλό όταν έχουμε να κάνουμε με αβέβαια δεδομένα.

4.1.1 Σημασιολογία

Συχνά η επεξεργασία των δεδομένων πρέπει να γίνει πάνω σε μια σειρά από πιθανότητες ενώ επίσης οι επερωτήσεις μπορεί να είναι φωλιασμένες πράγμα που τις κάνει πιο πολύπλοκες. Οι προσεγγίσεις που υπάρχουν έχουν να κάνουν με τη σημασιολογία που χρησιμοποιείται:

- Εκούσια
  Χρησιμοποιεί ένα μοντέλο που υπολογίζει όλους τους δυνατούς κόσμους μέσα από μια δομή δένδρου. Η δομή αυτή απαριθμεί όλες τις πιθανότητες μέσα από τις οποίες μπορεί να υπολογιστεί η επερώτηση. Η πολυπλοκότητα υπολογισμού είναι εκθετική αλλά εγγυάται σωστά αποτελέσματα.

- Επεκτατική
  Χρησιμοποιεί ένα προσεγγιστικό μοντέλο χωρίς να υπολογίζει ολόκληρο το δένδρο πιθανοτήτων. Χρησιμοποιεί συναρτήσεις που συνδέονται με την ύπαρξη ή όχι αβεβαιότητας και υπολογίζει προσεγγιστικά τα αποτελέσματα. Όπως είναι φυσικό δεν είναι πάντα ακριβής.

Η ευριστική φύση της μεθόδου επεκτατικής σημασιολογίας την κάνει αποτελεσματική για απλές και μη φωλιασμένες ερωτήσεις. Φυσικά η εκούσια λειτουργεί καλύτερα για πολύπλοκες. Για παράδειγμα, ας υποθέσουμε ότι έχουμε της πιθανές γραμμές $t_1, t_2, \dots, t_n$ με ίδια πιθανότητα εμφάνισης στη ΒΔ. Ας θεωρήσουμε ότι θέλουμε να βρούμε την πιθανότητα ύπαρξης $P(\mathrm{e}(t_1) \cap \mathrm{e}(t_2))$ των t_1, t_2 στη βάση, όπου $\mathrm{e}(t_1)$, $\mathrm{e}(t_2)$ είναι οι μεταβλητές εμφάνισης της κάθε γραμμής. Ανάλογα το πώς το κάθε σημασιολογικό μοντέλο απαριθμεί τις μεταβλητές και τις πιθανότητες τους μπορεί να χρησιμοποιηθεί και το καταλληλότερο. Αν οι γραμμές ήταν ανεξάρτητες τότε $P\big(\mathrm{e}(t_1) \cap \mathrm{e}(t_2)\big) = P(\mathrm{e}(t_1))P(\mathrm{e}(t_2))$ και το επεκτατικό μοντέλο θα ήταν αρκετό. Αν αντίθετα οι γραμμές δεν ήταν ανεξάρτητες και σχετίζονταν πολύπλοκα μόνο η εκούσια σημασιολογία μπορούσε να φέρει χρήσιμα αποτελέσματα.

Οι πιθανοτικές ΒΔ έχουν ανάγκη από μια αποτελεσματική μέθοδο που να μειώνει την πολυπλοκότητα αποτίμησης επερωτήσεων. Μια πρώτη προσπάθεια είχε γίνει στο (Fuhr & Rolleke, 1997) που παρουσιάστηκε στο κεφάλαιο 3. Πιο πρόσφατα, το (Dalvi & Suciu, Efficient Query Evaluation on Probabilistic Databases, 2004) υπολογίζει ένα επεκτατικό μοντέλο. Παρόλο

που δεν υπάρχει επέκταση για όλες τις επερωτήσεις, αποδεικνύεται ότι 80% έχουν. Για επερωτήσεις που δεν υπάρχει, το (Dalvi & Suciu, Efficient Query Evaluation on Probabilistic Databases, 2004) περιγράφει δύο επιπλέον τεχνικές. Η μία είναι μια ευρυστική μέθοδος που μειώνει τα λάθη υπολογισμού και η άλλη χρησιμοποιεί προσομοίωση Monte Carlo βασισμένη σε δείγματα δεδομένων. Η δεύτερη έχει μεγάλο υπολογιστικό κόστος αλλά είναι πιο ακριβής από την πρώτη. Στη βιβλιογραφία παρουσιάζονται και άλλες μέθοδοι που χρησιμοποιούν τις τεχνικές του προαναφερθέντος μοντέλου οι οποίες ξεφεύγουν από το στόχο αυτής της εργασίας.

4.1.2 Top-k επερωτήσεις

Ας υποθέσουμε ότι έχουμε ένα σύνολο από ασθενοφόρα που διαθέτουν συσκευή GPS και βρίσκονται κοντά σε κάποια περιοχή που έχει συμβεί ατύχημα. Θέλουμε να υπολογίσουμε τα 3 ασθενοφόρα που βρίσκονται πιο κοντά στο ατύχημα για να δούμε πιο είναι ευκολότερο να βρεθεί στον τόπο γρηγορότερα. Είναι φανερό ότι οι μετρήσεις δεν είναι ακριβείς λόγο εξασθένησης σήματος και σκεδάσεις μέσα στην πόλη καθώς και σφάλματα του οργάνου. Έτσι ο υπολογισμός του 3 πιο κοντινών σημείων μπορεί να αποδειχτεί πρόκληση.

Με μία top-k επερώτηση επιδιώκουμε να βρούμε τις καλύτερες k απαντήσεις ως προς μια συνάρτηση βαθμολόγησης (πχ απόσταση) σε ντετερμινιστικές εφαρμογές. Σε μη ντετερμινιστικές εφαρμογές ωστόσο δεν είναι τόσο απλό γιατί η εύρεση των k καλύτερων απαντήσεων δεν εξαρτάται μόνο από τη συνάρτηση βαθμολόγησης αλλά και από την πιθανότητα ύπαρξης. Άλλο ένα πρόβλημα είναι και η πιθανή συσχέτιση μεταξύ γραμμών που μπορεί να βρίσκονται στον πιθανό κόσμο και στην απάντηση.

Το (Soliman, Ilyas, & Chang, 2007) ασχολείται με το θέμα και παρουσιάζει μια σειρά από μεθόδους υπολογισμού top-k επερωτήσεων. Αν έχουμε μια

συσκευή μέτρησης της ταχύτητας, η οποία εισάγει σφάλματα μέτρησης και σφάλματα του παρατηρητή, τότε δύο top-k ερωτήσεις είναι

- Ποια είναι τα top-k αυτοκίνητα που έτρεχαν περισσότερο την τελευταία ώρα;
- Ποια είναι η σειρά μοντέλων των top-k γρηγορότερων αυτοκινήτων;

Οι επερωτήσεις αυτές δεν είναι εύκολες δεδομένης της αβεβαιότητας. Το (Soliman, Ilyas, & Chang, 2007) μοντελοποιεί τις επερωτήσεις σαν ένα χώρο καταστάσεων αναζήτησης. Για τη δημιουργία του χώρου, εισάγει μια Μηχανή Κανόνων (Rule Engine) η οποία υπολογίζει τις πιθανότητες μετάβασης καταστάσεων και είναι βασισμένη στα δίκτυα Bayes. Για την αναζήτηση δημιουργεί έναν αλγόριθμο αναζήτησης που εγγυάται την εύρεση των top-k στοιχείων. Επίσης, προτείνει ένα αλγόριθμο πλοήγησης σε υπάρχοντες DBMS χώρους.

Επιπλέον, το (Soliman, Ilyas, & Chang, 2007) δημιουργεί δύο νέες έννοιες top-k επερωτήσεων, τις U-kRank και U-Topk. Το U-kRank επιστρέφει τα k ταξινομημένα στοιχεία όπου το ι-οστό στοιχείο έχει την μεγαλύτερη πιθανότητα να βρεθεί στην ι-οστή θέση ανάμεσα σε όλους τους πιθανούς κόσμους. Για παράδειγμα, η παραπάνω επερώτηση «Ποια είναι η σειρά μοντέλων των top-k γρηγορότερων αυτοκινήτων» θα μπορούσε να είναι ίδια με την επερώτηση σε «Ποια είναι τα πιο πιθανά top-ιth γρηγορότερα αυτοκίνητα μέσα από όλους του δυνατούς συνδυασμούς αυτοκινήτων που μπορεί να πέρασαν». Το U-Topk επιστρέφει το σύνολο των k στοιχείων που έχουν τη μεγαλύτερη πιθανότητα να βρίσκονται στα top-k αποτελέσματα όλων των πιθανών κόσμων. Η επερώτηση τώρα γίνεται «Ποια είναι τα πιο πιθανά top-k γρηγορότερα αυτοκίνητα από όλα αυτά που πέρασαν». Ο μαθηματικός ορισμός των παραπάνω όρων βρίσκεται στη δημοσίευση αλλά ξεφεύγει από το στόχο αυτής της εργασίας.

Επιπλέον, το έργο στο (Yi, Li, Srivastava, & Kollios, 2007) βελτιώνει την απόδοση των αλγορίθμων στο (Soliman, Ilyas, & Chang, 2007). Περαιτέρω σχετική δουλειά παρουσιάζεται και στα Hua (Hua, Pei, & Zhang, 2008) και (Re, Dalv, & Suciu, 2007).

4.1.3 Ενώσεις (joins)

Στις πιθανοτικές ΒΔ κάθε ζευγάρι γραμμών έχει μια πιθανότητα να βρίσκεται σε μια ένωση. Αυτό σημαίνει ότι ορισμένες φορές μπορεί να έχουμε λανθασμένες γραμμές ως αποτελέσματα ενώσεων (false positives). Το (Cheng, Xia, Prabhakar, Shah, Vitter, & Xia, Efficient Join Processing over Uncertain Data, 2005) παρουσιάζει την έννοια των πιθανοτικών επερωτήσεων ένωσης που χρησιμοποιεί όρια για το αν μια γραμμή βρίσκεται στην ένωση ή όχι. Φυσικά, ανάλογα με το πώς επιλεγούν τα όρια μπορεί να εισάγει περισσότερα σφάλματα με γραμμές που λανθασμένα μένουν έξω από την ένωση (false negatives).

Η αβεβαιότητα για ένα δεδομένο a παραμετροποιείται με ένα όριο $a.U$ και συνάρτηση πυκνότητας πιθανότητας $a.f(x)$. Οι τελεστές που ορίζονται είναι ισότητα, ανισότητα, μεγαλύτερο και μικρότερο. Δύο δεδομένα a και b θεωρούνται ίσα όταν $|a-b| \leq c$, όπου c είναι το επονομαζόμενο όριο επίλυσης (resolution). Η πιθανότητα το a να είναι ίσο με το b δίνεται από τη σχέση

$$P(a =_c b) = \int_{-\infty}^{+\infty} a.f(x)\,(b.F(x+c) - b.F(x-c)\,)\mathrm{dx}$$

όπου $b.F(x)$ είναι η συνάρτηση η αθροιστική κατανομή πιθανότητας (cumulative distribution function CDF) του b. Για άλλους τελεστές όπως του μεγαλύτερου ή του μικρότερου ο υπολογισμός της πιθανότητα είναι πιο απλός.

Επιπλέον μια σειρά από τεχνικές αποκοπής εφαρμόζονται για τη βελτιστοποίηση των επερωτήσεων ένωσης. Τέτοιες είναι η αποκοπή σε επίπεδο στοιχείου, αποκοπή σε επίπεδο σελίδας και επίπεδο δείκτη. Όλες

αφορούν στο αντικείμενο που αποκλείεται από τον υπολογισμό πιθανότητας σύμφωνα με συγκεκριμένα μαθηματικά στατιστικά κριτήρια.

Το (Cheng, Xia, Prabhakar, Shah, Vitter, & Xia, Efficient Join Processing over Uncertain Data, 2005) παρουσιάζει με μεγαλύτερη λεπτομέρεια όλους του τρόπους υπολογισμού πιθανοτήτων και τα κριτήρια αποκλεισμού από τον υπολογισμό για τη βελτιστοποίηση της αποτίμησης της επερώτησης, ο οποίοι ξεφεύγουν από αυτή εδώ την εργασία. Ο αναγνώστης μπορεί να ανατρέξει στη δημοσίευση για περισσότερες πληροφορίες. Τέλος, πιο πρόσφατα, το (Ljosa & Singh, 2008) ασχολείται με ενώσεις σε χωρικές πιθανοτικές ΒΔ.

4.1.4 Ενώσεις ομοιότητας

Ένα πολύ κοινό πρόβλημα ένωσης ομοιότητας είναι αυτό της απόστασης. Αν υποθέσουμε ότι έχουμε δύο σημεία, η ένωση γίνεται μόνο αν αυτά τα σημεία απέχουν μεταξύ τους λιγότερο από κάποια τιμή ε. Το πρόβλημα ερευνάται στο (Kriegel, Kunath, Pfeifle, & Renz, 2006). Η πιθανότητα η απόσταση μεταξύ δύο σημείων U, V να είναι μεταξύ $a \leq \varepsilon \leq b$ είναι

$$P(a \leq distance(U,V) \leq b) = \int_a^b f_d(U,V)xdx$$

όπου f_d είναι η συνάρτηση πυκνότητας απόστασης μεταξύ U και V. Αν θέλαμε να γενικεύσουμε το παραπάνω πρόβλημα, θα θέλαμε να ενώσουμε δύο σχέσεις μόνο αν η απόσταση τους είναι μικρότερη από ε, πράγμα που δεν είναι και τόσο εύκολο γιατί η αναμενόμενη απόσταση επηρεάζεται από τις ουρές των συναρτήσεων πυκνότητας πιθανότητας των στοιχείων. Αυτό όπως είναι αναμενόμενο μπορεί να δημιουργήσει λανθασμένες ενώσεις καθώς η πιθανότητα να βρίσκεται η ένωση εντός του ορίου μπορεί να μην αντικατοπτρίζεται στις αναμενόμενες τιμές. Το (Kriegel, Kunath, Pfeifle, & Renz, 2006) προτείνει τη δημιουργία μιας πιθανότητας για κάθε ζευγάρι τιμών s χρησιμοποιώντας Monte-Carlo

τεχνικές δειγματοληψίας. Η πιθανότητα αυτή σημαίνει το πόσο πιθανό είναι τα δύο στοιχεία να βρίσκονται μέσα στο σύνολο της ένωσης (να ανήκουν στην ένωση). Η πιθανότητα αυτή δίνεται από τον τύπο

$$P(distance(U,V) \leq \varepsilon) = \frac{|\{(u_i, v_j) | distance(u_i, v_j) \leq \varepsilon, i \geq 1, j \leq s\}|}{s^2}$$

όπου $u_i \in U, v_j \in V$. Σημειωτέον ότι, ο Kriegel και οι λοιποί χρησιμοποιούν μια ντετερμινιστική περίπτωση όπου οι αποστάσεις είναι γνωστές και γι' αυτό οι ενώσεις ομοιότητας είναι ειδικές περιπτώσεις των ενώσεων.

4.1.5 Συσχετισμένες επερωτήσεις

Πολλές φορές τα μοντέλα θεωρούν ότι οι γραμμές της ΒΔ είναι ανεξάρτητες μεταξύ τους (Dalvi & Suciu, Efficient Query Evaluation on Probabilistic Databases, 2004). Ωστόσο σπάνια αυτό είναι αληθές. Για παράδειγμα, σε μια ΒΔ που αποθηκεύει στοιχεία όπως η διεύθυνση και το κόστος σπιτιών, ποτέ οι τιμές που υπάρχουν σε κάθε γραμμή δεν είναι ανεξάρτητες μεταξύ τους. Σπίτια που βρίσκονται σε κοντινές διευθύνσεις έχουν συσχετισμένες τιμές κόστους. Ακόμα και αν οι γραμμές σε μια ΒΔ ήταν ανεξάρτητες πολλά ενδιάμεσα αποτελέσματα κατά την αποτίμηση μιας επερώτησης είναι συσχετισμένα μεταξύ τους.

Το (Sen & Deshpande, 2007) αναπτύσσει μια τεχνική στατιστικής μοντελοποίησης για την αποτίμηση των συσχετισμένων δεδομένων. Χρησιμοποιεί κοινές κατανομές πιθανότητας για να εκφράσει τις συσχετίσεις και να κατασκευάσει μια δομή αποτίμησης. Ταυτόχρονα βασίζεται σε γραφικά μοντέλα τα οποία μπορούν να αναπαραστήσουν πολύπλοκες συσχετίσεις τυχαίων μεταβλητών. Υπάρχουν μια σειρά από τεχνικές/αλγόριθμοι που έχουν αναπτυχθεί και μπορούν να χρησιμοποιηθούν για πιθανοτικά γραφικά μοντέλα και αυτό κάνει την πρόταση του (Sen & Deshpande, 2007) ιδιαίτερα ενδιαφέρουσα αν

προϋποτεθεί η επιλογή του κατάλληλου αλγορίθμου ανάλογα με τις ανάγκες ταχύτητας και ακρίβειας.

4.2 Σελιδοποίηση

Το πρόβλημα της σελιδοποίησης (indexing) υπάρχει σε διάφορες εφαρμογές όπως αυτές των κινούμενων αντικειμένων ή των δεδομένων αισθητήρων. Αυτό ισχύει γιατί τα δεδομένα συγκεντρώνονται περιοδικά (όχι ανά πάσα στιγμή) και μόνο ο δείκτης στοιχείο ενημερώνεται. Για παράδειγμα, δεν αποθηκεύονται όλες οι μετρήσεις θερμοκρασίας για ένα σημείο του χώρου αλλά μόνο ο δείκτης της τιμής του σημείου αυτού. Υπάρχουν διαφορετικά είδη επερωτήσεων με τα οποία ασχολείται η σελιδοποίηση

- Όρια τιμών: Σκοπός είναι να εντοπιστούν όλες οι τιμές μεταξύ δύο ορίων. Τα στοιχεία είναι αβέβαια διότι οι θέσεις τους δεν είναι ακριβείς οπότε η ύπαρξη τους ή μη μεταξύ δύο ορίων δεν είναι απλή. Γι’ αυτό χρησιμοποιούνται πιθανότητες ύπαρξης τους στα όρια και όταν αυτή υπερβαίνει κάποια τιμή θεωρούνται ότι τα σημεία είναι εντός των ορίων.
- Κοντινότερος γείτονας: Προσπαθούν να βρουν το στοιχείο που έχει τη μικρότερη σχέση γειτονίας με το δεδομένο στόχο. Και πάλι επειδή τα στοιχεία είναι πιθανοτικά το πρόβλημα μεταλλάσσεται στην εύρεση της πιθανότητας ένα στοιχείο να είναι ο κοντινότερος γείτονας ενός άλλου.
- Αθροιστικά: Ασχολούνται με αθροιστικά στατιστικά δεδομένων όπως το άθροισμα ή το μέγιστο.

Τα βασικότερα θέματα σελιδοποίησης παρουσιάζονται στα υποκεφάλαια που ακολουθούν.

4.2.1 Επερωτήσεις με όρια

Υπάρχουν κάποιες επερωτήσεις που υπολογίζουν τα στοιχεία που ικανοποιούν κάποια κριτήρια με συγκεκριμένη πιθανότητα. Αυτές ονομάζονται επερωτήσεις με όρια και παρουσιάζονται στο (Cheng, Xia, Prabhakar, Shah, & Vitter, 2004). Αν θέλουμε να ορίσουμε αυστηρά ένα επερώτημα με όρια, τότε θα πρέπει πούμε δεδομένου ενός ορίου [c,d], όπου $c, d \in \mathbb{R}$ και $c \leq d$, ένα πιθανοτικό οριακό επερώτημα επιστρέφει γραμμές T_i, τέτοιες ώστε η πιθανότητα p_i του $T_i.a$ να βρίσκεται μεταξύ [c,d] είναι μεγαλύτερη ή ίση με p ($0 \leq p \leq 1$), όπου $T_i.a$ είναι το ένα στοιχείο της γραμμής T_i.

Το (Cheng, Xia, Prabhakar, Shah, & Vitter, 2004) παρουσιάζει μια σειρά από δομές και τεχνικές για τον υπολογισμό αυτών των επερωτήσεων. Η πιο απλή είναι να υπολογίσει κανείς όλες τις γραμμές και να τις συγκρίνει με το όριο [c,d]. Όταν βρει όλες τις γραμμές που είναι φραγμένες από το όριο, ο υπολογισμός της πιθανότητας είναι απλή υπόθεση. Το μειονέκτημα αυτής της μεθόδου είναι ότι υπολογίζει όλες τις γραμμές χωρίς όλες να είναι χρήσιμές πράγμα που μπορεί να είναι πρακτικά υπολογιστικά πολύ ακριβό.

Μια άλλη τεχνική είναι αυτή της Σελιδοποίησης Ορίου Πιθανότητας (Probability Threshold Indexing). Βασίζεται σε μια δομή ενός «μεταλλαγμένου» R-δένδρου. Η πιθανότητες ενσωματώνονται στους κόμβους του δένδρου για να δώσουν τη δυνατότητα αποκοπής. Επίσης δημιουργεί ένα όριο που το ονομάζει x-όριο και είναι αυστηρότερο από το γνωστό μας Ελάχιστο Ορθογώνιο Φράγμα (Minimum Bounding Rectangle (MBR)) για κάθε κόμβο. Η ύπαρξη του ορίου αυτού μας επιτρέπει να βρούμε τα MBR χωρίς να χρειάζεται να προχωρήσουμε μέσα στο βάθος τον κόμβων. Από την άλλη αν το φράγμα [c,d] δεν τέμνεται με το χ-όριο με πιθανότητα τουλάχιστον p τότε ο κόμβος μπορεί να αποκοπεί από το δένδρο, πράγμα που βελτιστοποιεί τη διάσχιση του. Στη συνέχεια η διάσχιση του δένδρου συνεχίζεται για τους κόμβους που δεν έχουν

αποκοπεί. Ο παραπάνω αλγόριθμος είναι πιο αποτελεσματικός όταν η πιθανότητα p είναι σταθερή για όλες τις επερωτήσεις.

Για περισσότερες τεχνικές λεπτομέρειες ο αναγνώστης μπορεί να ανατρέξει στο (Cheng, Xia, Prabhakar, Shah, & Vitter, 2004). Αξίζει να σημειωθεί ότι άλλη μια μεθοδολογία που αυτή τη φορά χρησιμοποιεί U-δένδρα (U-tree) παρουσιάστηκε στο (Tao, Cheng, Xiao, Ngai, Kao, & Prabhakar, 2005).

4.2.2 Επερωτήσεις κοντινότερου γείτονα

Υπάρχουν πολλές εφαρμογές που μετρούν τη θέση και την κίνηση αντικειμένων. Τέτοιες εφαρμογές δεν μπορούν να αποθηκεύσουν την απόλυτη θέση κάθε αντικειμένου ανά πάσα στιγμή φυσικά. Οπότε κρατούν τη θέση σε κάθε στιγμή μέτρησης και όσο κινείται το αντικείμενο δεν είναι γνωστό που βρίσκεται και η αβεβαιότητα της θέσης του αυξάνεται. Το σφάλμα σε επερωτήσεις τέτοιων δεδομένων μειώνεται μειώνοντας την αβεβαιότητα (πιο συχνές μετρήσεις).

Υπάρχουν διάφορα μοντέλα για την κίνηση αντικειμένων. Τα κυριότερα ασχολούνται με την απόσταση που διανύει το αντικείμενο καθώς κινείται μεταξύ δύο μετρήσεων. Υπάρχει μια αναμενόμενη απόσταση που το αντικείμενο αναμένεται να βρίσκεται και σε κάθε μέτρηση αυτή η θέση ανανεώνεται. Άλλες προσεγγίσεις εισάγουν και την τροχιά της κίνησης του αντικειμένου (πχ ευθεία γραμμή).

Πιο συγκεκριμένα, το (Cheng, Kalashnikov, & Prabhakar, 2004) ορίζει μια περιοχή όπου το αντικείμενο μπορεί να βρίσκεται με πιθανότητα ίση με 1 καθώς και μια συνάρτηση πυκνότητας πιθανότητα για το αντικείμενο να βρίσκεται σε μια συγκεκριμένη θέση συντεταγμένων σε κάποιο χρόνο. Επίσης ορίζει και την επερώτηση του κοντινότερου γείτονα που υπολογίζει τα υποψήφια αντικείμενα να βρίσκονται κοντινότερα σε μια θέση με τη μεγαλύτερη πιθανότητα. Η τεχνική που παρουσιάζεται αποτιμά την

πιθανότητα κάθε αντικειμένου να είναι ο κοντινότερος γείτονας με μεγαλύτερη πρόκληση το γεγονός ότι αυτή η πιθανότητα ενός σημείου δεν είναι ανεξάρτητη από τα υπόλοιπα σημεία. Ο τρόπος υπολογισμού χρησιμοποιεί τις μεθόδους της προβολής (projection), αποκοπής (pruning), φραγής (bounding) και αποτίμησης (evaluation). Κατά την προβολή υπολογίζεται η περιοχή που μπορεί να βρίσκεται το αντικείμενο. Αυτό γίνεται σύμφωνα με την τελευταία του θέση, το μοντέλο της κίνησης (τροχιά), το χρόνο που πέρασε και την ταχύτητα του αντικειμένου. Κατά την αποκοπή ορισμένα αντικείμενα αποκλείονται χωρίς να γίνει αποτίμηση της πιθανότητας. Για παράδειγμα αν μια περιοχή που υπολογίστηκε κατά την προβολή έχει ελάχιστη απόσταση μεγαλύτερη από τη μέγιστη απόσταση μια άλλης περιοχή, τότε το αντικείμενο που όρισε την πρώτη περιοχή αποκλείεται να είναι ο κοντινότερος γείτονας. Με άλλα λόγια κάθε αντικείμενο που έχει ελάχιστη απόσταση από το στόχο μεγαλύτερη από την ελάχιστη των μεγίστων αποστάσεων των περιοχών από το στόχο αποκλείεται. Κατά τη φραγή, εξετάζονται μόνο μη αποκλεισμένες περιοχές. Πιο συγκεκριμένα δε χρειάζεται να εξεταστούν ολόκληρες οι περιοχές αλλά μόνο τα μέρη τους που απέχουν λιγότερο από την ελάχιστη των μεγίστων αποστάσεων. Τέλος, τα τμήματα των περιοχών που έχουν απομένει από τις προηγούμενες φάσεις ερευνώνται στη φραγή. Κατά τη φραγή γίνεται ολοκλήρωση του γινομένου της πυκνότητας πιθανότητα για ένα αντικείμενο να βρίσκεται σε συγκεκριμένη απόσταση από το στόχο επί το την πυκνότητα όλων των υπολοίπων αντικειμένων να βρίσκονται σε μεγαλύτερη απόσταση από αυτόν.

4.2.3 Αθροιστικές επερωτήσεις

Πρόσφατα, η ομάδα του Infolab του Stanford παρουσίασε μια μέθοδο για αποτίμηση αθροιστικών επερωτήσεων (M. Mutsuzaki, 2007). Αθροιστικές ονομάζονται οι επερωτήσεις που εφαρμόζονται σε σύνολο δεδομένων και υπολογίζουν αθροιστικές τιμές. Τέτοιες είναι οι ΑΘΡΟΙΣΜΑ (SUM), ΜΕΓΙΣΤΟ (MAX), ΑΡΙΘΜΟΣ (COUNT). Το (M. Mutsuzaki, 2007) ασχολείται

με τις ΑΡΙΘΜΟΣ, ΑΘΡΟΙΣΜΑ, ΜΕΓΙΣΤΟ (MIN), ΕΛΑΧΙΣΤΟ, ΜΕΣΗ ΤΙΜΗ (AVG). Υπάρχουν αρκετές μέθοδοι υπολογισμού των ΑΡΙΘΜΟΣ, ΜΕΓΙΣΤΟ, ΕΛΑΧΙΣΤΟ που έχουν προέλθει κυρίως από δημοσιεύσεις σε δίκτυα αισθητήρων καθώς τέτοιες επερωτήσεις είναι κατεξοχήν ερωτήσεις που χρησιμοποιούνται σε δενδροειδή δίκτυα αισθητήρων που συλλέγουν τιμές. Ωστόσο οι ΑΘΡΟΙΣΜΑ, ΜΕΣΗ ΤΙΜΗ αλλάζουν ανάλογα με τους πιθανούς κόσμους και είναι #P-complete (Chui & Kao, A Decremental Approach for Mining Frequent Itemsets from Uncertain Data, 2008).

Για την αποφυγή της ενδελεχούς αποτίμησης των αθροιστικών επερωτήσεων υπάρχουν τρεις εναλλακτικές: η ελάχιστη πιθανή τιμή, η αναμενόμενη τιμή και η μέγιστη πιθανή τιμή. Για παράδειγμα LSUM (ελάχιστη πιθανή τιμή αθροίσματος) είναι το ΑΘΡΟΙΣΜΑ των ελαχίστων τιμών των αβέβαιων στοιχείων. Αντίστοιχα το EAVG (αναμενόμενη τιμή μέσης τιμής) είναι $ESUM/ECOUNT$. Το Trio Project (κεφάλαιο 6) χρησιμοποιεί αυτές τις τεχνικές.

Μια εντελώς διαφορετική προσέγγιση ακολουθείται από το γνωστό OLAP μοντέλο δεδομένων. Το OLAP δεν είναι καινούριο και μέσα στα όρια του έχουν παρουσιαστεί μια σειρά από σημαντικές δημοσιεύσεις με χρήσιμες μεθόδους για αθροιστικά δεδομένα που χρησιμοποιούνται σε ένα μέρος του κύβου των δεδομένων. Χαρακτηριστικό του μοντέλου είναι η ιεραρχική δομή και φυσικά η ασάφεια/αβεβαιότητα για τα δεδομένα. Στο (Burdick, Deshpande, Jayram, Ramakrishnan, & Vaithyanathan, 2005) παρουσιάζονται κριτήρια για τα διαχείριση της ασάφειας. Αυτά είναι

- Συνέπεια
  Παρόμοιες επερωτήσεις εφαρμοζόμενες στην ίδια δομή του κύβου δεδομένων πρέπει να επιστρέφουν αντίστοιχα αποτελέσματα. Για παράδειγμα το ΑΘΡΟΙΣΜΑ μιας περιοχής πρέπει να είναι ίδιο με το ΑΘΡΟΙΣΜΑ των αποτελεσμάτων των ΑΘΡΟΙΣΜΑΤΩΝ υποπεριοχών της αρχικής περιοχής.
- Πιστότητα

Όσο πιο ακριβή είναι τα δεδομένα τόσο πιο ακριβή θα είναι και τα αποτελέσματα των επερωτήσεων. Παράδειγμα αποτελεί το ΑΘΡΟΙΣΜΑ μη αρνητικών τιμών,. Είναι αυτονόητο ότι όσο πιο πολύ αποκλίνουν οι τιμές από την αλήθεια τόσο πιο λανθασμένο θα είναι το αποτέλεσμα $(\sum_{i,v,\varepsilon>0}(v+\varepsilon) > \sum_{i,v,\varepsilon>0} v)$.

- Διατήρηση συσχέτισης
  Η συσχέτιση μεταξύ δύο δεδομένων δεν επηρεάζεται από την εισαγωγή αβεβαιότητας στην τιμή.

Εκτός από τα παραπάνω, το (Burdick, Deshpande, Jayram, Ramakrishnan, & Vaithyanathan, 2005) κάνει και κάποιες επεκτάσεις που αφορούν στα χαρακτηριστικά διάστασης (τοποθεσίας) και προσθέτει ένα νέο χαρακτηριστικό στις τιμές που έχει να κάνει με την αβεβαιότητα (σαν συνάρτηση πυκνότητας πιθανότητας). Για την αντιμετώπιση της αβεβαιότητας προτείνει

- Ανάθεση επερώτησης, σύμφωνα με την οποία δεδομένα που βρίσκονται στα ανώτερα στρώματα της ιεραρχικής δομής πρέπει να ανατεθούν σε κόμβους φύλλα κατά την ανάθεση. Η ανάθεση γίνεται με βάρη σε κόμβους διαφορετικών επιπέδων.
- Σύνθεση αβέβαιων μετρήσεων
  Η επερώτηση αποτιμάται σε διαφορετικές συναρτήσεις πυκνότητας πιθανότητας. Το πρόβλημα αντιμετωπίζεται παρόμοια με το opinion polling πρόβλημα στη στατιστική.

Περισσότερες λεπτομέρειες μπορεί κανείς να βρει στο (Burdick, Deshpande, Jayram, Ramakrishnan, & Vaithyanathan, 2005)

4.2.4 Κατηγορικά δεδομένα

Κατηγορικά είναι τα δεδομένα που χαρακτηρίζουν ένα στοιχείο. Με άλλα λόγια είναι ιδιότητες για κάποια στοιχεία της ΒΔ. Για παράδειγμα, ένα χαρακτηριστικό ενός ανθρώπου για το αν είναι ψηλός (μπορεί να ορίζεται από την ιδιότητα του να ξεπερνά το μέσο όρο) μπορούμε να πούμε ότι είναι ένα διακριτό κατηγορικό δεδομένο. Με τα αβέβαια κατηγορικά δεδομένα ασχολείται το (Singh, Mayfield, Prabhakar, Shah, & Hambrusch, Indexing Uncertain Categorical Data, 2006).

Ορισμός: Δεδομένου ενός κατηγορικού χώρου $D = \{d_1, d_2, \dots, d_n\}$, η κατανομή πιθανότητας στο D ονομάζεται αβέβαιο διακριτό χαρακτηριστικό (UDA) και μπορεί να αναπαρασταθεί ως $u.P = \{p_1, p_2, \dots, p_n\}$ τέτοιο ώστε $\Pr(u = d_i) = u.p_i$

Αν υπολογίσουμε την πιθανότητα ισότητας μέσα σε όλες τις πιθανές τιμές, τότε μπορούμε να πάρουμε την πιθανότητα δύο αβέβαιες τιμές να είναι ίσες. Έτσι μπορούμε να πούμε ότι γνωρίζοντας δύο UDAs u και v, η πιθανότητα να είναι ίσες είναι $\Pr(u = v) = \sum_{i=1}^{n} u.p_i \times v.p_i$.

Το (Singh, Mayfield, Prabhakar, Shah, & Hambrusch, Indexing Uncertain Categorical Data, 2006) επίσης παρουσιάζει μια μέθοδο εύρεσης τη πιθανότητας ομοιότητας με χρήση συναρτήσεων απόστασης και συναρτήσεων Kullback-Leibler. Χρησιμοποιεί δομές δεικτών για την επίλυση κατηγορικών επερωτήσεων, όπως ο πιθανοτικός ανεστραμμένος δείκτης. Σε αυτή τη δομή τοποθετούνται σε μια B-δένδρο λίστα όλες οι τιμές των γραμμών που μπορεί να περιέχουν ένα κατηγόρημα καθώς και η πιθανότητα τους. Μια σειρά από τεχνικές πρόσθεσης στοιχείων, αφαίρεσης και αποκοπής εφαρμόζονται στη λίστα για γραμμές που δεν προκρίνονται. Για την απάντηση της επερώτησης συγκεντρώνονται γραμμές που ικανοποιούν τα κατηγορικά κριτήρια και που στη συνέχεια βρίσκονται πάνω από το πιθανοτικό όριο.

Η παραπάνω τεχνική χρησιμοποιείται σε

- Probabilistic equality query (PEQ): Επιστρέφει όλες τις γραμμές για τις οποίες $\Pr(q = t.a) \geq 0$, όπου q είναι ένα UDA

- Probabilistic equality threshold query (PETQ): Επιστρέφει όλες τις γραμμές για τις οποίες $\Pr(q = t.a) \geq \tau$, όπου q είναι ένα UDA και τ είναι ένα αυθαίρετο όριο.

- Distributional similarity threshold query (DSTQ): Επιστρέφει όλες τις γραμμές για τις οποίες $F(q, t.a) \leq \tau$, όπου q είναι ένα UDA, τ είναι ένα αυθαίρετο όριο και F η συνάρτηση απόκλισης.

- Probabilistic equality threshold join (PETJ): Επιστρέφει όλες τις τιμές του $R \bowtie_{R_a = R_b \tau} S$ για τις οποίες ισχύει $\Pr(r.a = s.b) \geq \tau$, όπου a,b είναι UDAs, R και S είναι δύο σχέσεις και τ είναι ένα αυθαίρετο όριο.

4.2.5 R-δένδρα

Στη συνέχεια θα παρουσιάσουμε τα πιθανοτικά R-δένδρα (R-trees). Τα R-δένδρα είναι μια εναλλακτική για τη σελιδοποίηση UDAs. Σε γενικές γραμμές αυτό που επιδιώκεται είναι να σελιδοποιηθεί το διάνυσμα των πιθανοτήτων των πιθανών τιμών των στοιχείων. Για N αριθμό πιθανών τιμών τότε τα σημεία ανήκουν στο $\mathbb{R}^N$. Η τεχνική μοιάζει με τα γνωστά μας στις ΒΔ R-δένδρα, αλλά υπάρχουν διαφορές στη σημασιολογία. Οι επερωτήσεις είναι επερωτήσεις στον N-διάστατο κύβο ενώ το Ελάχιστο Ορθογώνιο Φράγμα (Minimum Bounding Rectangle (MBR)) του R-δένδρου ορίζεται με τις πιθανοτικές τιμές. Οι γνωστές τεχνικές αποκοπής (pruning) του δένδρου μπορούν να εφαρμοστούν ακόμα.

Οι δύο τελευταίες τεχνικές σελιδοποίησης συγκρίνονται στο (Singh, Mayfield, Prabhakar, Shah, & Hambrusch, 2006) χωρίς να μπορούμε να

πούμε ότι κάποια από τις δύο επικρατεί. Αποδεικνύεται ότι η καθεμία έχει τα πλεονεκτήματα της ανάλογα με τα δεδομένα και τις επερωτήσεις.

4.3 Skyline επερωτήσεις

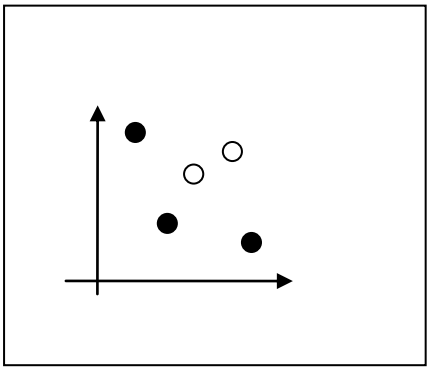
Ας θεωρήσουμε μια ομάδα μπάσκετ που παίζει στο πρωτάθλημα και το τεχνικό προσωπικό της ομάδας θέλει αν πάσα στιγμή να γνωρίζει τον παίχτη που βρίσκεται πρώτος σε όλες μαζί τις κατηγορίες των ριμπάουντ, ασίστ, πόντων και κλεψιμάτων. Το πρόβλημα του skyline εφαρμόζεται σε «πολύκριτηριακές» εφαρμογές όπως οι παραπάνω.

Ορισμός: Για δύο n-διάστατα σημεία $u = (u_1, u_2, \dots, u_n)$, $v = (v_1, v_2, \dots, v_n)$, το u λέγεται ότι επικρατεί επί του v ($u \prec v$), αν $\forall i \in \{1,2, \dots n\}$, υπάρχει v_i: $u_n \leq v_n$ και για κάποιο $i_0 \in \{1,2, \dots n\}$ για $u_{i_0} \leq v_{i_0}$.

Ο παραπάνω ορισμός ισχύει αν προτιμότερες είναι οι μικρότερες τιμές αλλά είναι εύκολο να επεκταθεί και για το αντίστροφο.

Ορισμός: Δεδομένου ενός συνόλου από στοιχεία S, ένα σημείο u είναι σημείο γραμμής του ουρανού (skyline) αν δεν υπάρχει άλλο σημείο $v \in S$ που να επικρατεί επί του u. Η skyline γραμμή αποτελείται από όλα τα σημεία skyline.

Για παράδειγμα, το Σχήμα 2 παρουσιάζει εποπτικά τα σημεία που ανήκουν στη skyline. Στα σημεία αυτά δεν επικρατεί κανένα άλλο σημείο στο χώρο. Σύμφωνα με τον ορισμό, ένα σημείο u επικρατεί σε ένα άλλο v αν δεν υπάρχει διάσταση στην οποία το u έχει μεγαλύτερη τιμή από το v και υπάρχει μια τουλάχιστον διάσταση που το u έχει μικρότερη τιμή από το v.



Σχήμα 2: Σημεία Skyline στο δισδιάστατο χώρο

Αν επιστρέψουμε πάλι στον παράδειγμα με τους παίχτες του μπάσκετ. Είναι φανερό ότι οι παίχτες της γραμμής ουρανού είναι πολύ καλοί παίχτες. Ωστόσο, κάθε παίχτης της γραμμής δεν μπορεί να είναι πάντα στη γραμμή σε κάθε παιχνίδι. Λόγω διαφόρων παραγόντων η απόδοση αλλάζει πιθανοτικά. Έτσι, η ανάλυση για τους καλύτερους μπορεί να γίνει με χρήση των μέσων τιμών στην στατιστική κατηγορία για κάθε παίχτη ή μπορεί να ακολουθηθεί μια μέθοδος με δεδομένα αβεβαιότητας.

Η εύρεση της skyline δεν είναι απλή υπόθεση σε πιθανοτικά δεδομένα. Το δυσκολότερο κομμάτι είναι να υπολογιστεί η σχέση επικράτησης μεταξύ αβέβαιων δεδομένων. Το (Pei, Jiang, Lin, & Yuan, 2007) αναλύει το θέμα διεξοδικά. Εισάγει την έννοια της πιθανοτικής γραμμής ουρανού, που είναι η πιθανότητα σ' ένα στοιχείο να μην επικρατεί κανένα άλλο στοιχείο. Δεδομένης της πιθανότητας p, το (Pei, Jiang, Lin, & Yuan, 2007) ορίζει την p-skyline σαν ένα σύνολο από αβέβαια στοιχεία, τέτοιο ώστε όλα τα στοιχεία να έχουν πιθανότητα τουλάχιστον p να βρίσκονται στη skyline. Για δεδομένα συνεχούς φάσματος η πιθανότητα για ένα σημείο U να βρίσκεται στη skyline δίνεται από τον τύπο

$$\Pr(U) = \int_{u \in U} f(u) \prod_{\forall V \neq U} (1 - \int_{v<u} f'(v)dv)du$$

όπου f είναι η συνάρτηση πυκνότητας πιθανότητας του U και f' είναι αυτή του V. $1 - \int_{v<u} f'(v)dv$ είναι η πιθανότητα το $u \in U$ να μην το επικρατεί

κανένα άλλο σημείο. Αντίστοιχα, για διακριτού χώρου σημεία, η πιθανότητα του U να βρίσκεται στη skyline δίνεται από το

$$\Pr(U) = \sum_{u \in U} (P(u) \prod_{\forall V \neq U} 1 - \sum_{v \in V, v < u} P(v)))$$

όπου $\prod_{\forall V \neq U} 1 - \sum_{v \in V, u < v} P(v))$ είναι πάλι πιθανότητα το $u \in U$ να μην το επικρατεί κανένα άλλο σημείο.

Το πρόβλημα στην αποτίμηση των παραπάνω πιθανοτήτων είναι η εύρεση της πυκνότητας πιθανότητας των σημείων. Για να γίνει αυτό χρειάζονται μια σειρά από δεδομένα για τον προσεγγιστικό υπολογισμό της. Με άλλα λόγια θα χρειάζονται πολλά παιχνίδια για να πάρουμε τα στατιστικά των παιχτών και να εξάγουμε την πυκνότητα πιθανότητας από αυτά, πράγμα που σε αρκετές περιπτώσεις είναι πρακτικά δύσκολο.

Το (Pei, Jiang, Lin, & Yuan, 2007) προτείνει δύο ακόμα αλγορίθμους για την εύρεση της πιθανότητας. Ο πρώτος είναι ένας από-κάτω-προς-τα-πάνω (bottom-up) αλγόριθμος που ξεκινάει από τιμές των δεδομένων και χρησιμοποιεί αυτές την τιμές για να υπολογίσει τις αρχικές πιθανότητες και στη συνέχεια να αποκόψει άλλες. Ο δεύτερος είναι από-πάνω-προς-τα-κάτω (top-down) αλγόριθμος που χωρίζει τα δεδομένα σε τμήματα και αποκόπτει κομμάτια πριν υπολογίσει την πιθανότητα. Και οι δύο είναι αλγόριθμοι που χρησιμοποιούν τεχνικές ορίων και αποκοπής με αλλεπάλληλες σαρώσεις των δεδομένων.

5. Εφαρμογές εξόρυξης αβέβαιων δεδομένων

Τον τελευταίο καιρό μια σειρά από εφαρμογές εξόρυξης δεδομένων έχουν αναπτυχθεί και αφορούν κυρίως στην συσταδοποίηση (clustering) και ταξινόμηση. Αξιοσημείωτο είναι ότι η ύπαρξη αβεβαιότητας μπορεί να επηρεάσει σημαντικά τέτοιες εφαρμογές. Για παράδειγμα, ένα στοιχείο μέτρησης που έχει μεγαλύτερη αβεβαιότητα πρέπει να αντιμετωπίζεται διαφορετικά από ένα άλλο που είναι ακριβές και ένα στοιχείο με μεγαλύτερη βεβαιότητα είναι πιο χρήσιμο στις τεχνικές ταξινόμησης.

5.1 Ταξινόμηση δεδομένων

Ένα ακόμα πρόβλημα εξόρυξης είναι αυτό της ταξινόμησης ενός δεδομένου σε μια κλάση τιμών από ένα σύνολο κλάσεων τιμών. Το (Bi & Zhang, 2004) προτείνει μια τεχνική υποστήριξης μηχανής διανυσμάτων (support vector machine (SVM)) για αβέβαια δεδομένα. Η τεχνική αυτή βασίζεται σε ένα διακριτό μοντέλο που με τη σειρά του ξεκινάει από τη μέθοδο ελαχίστων τετραγώνων (least squares method). Είναι θα λέγαμε ένα γεωμετρικής λογικής μοντέλο. Η μέθοδος ελαχίστων τετραγώνων θεωρεί ένα περιβάλλον προσθετικού θορύβου όπου οι τιμές έχουν ένα πρόσθετο σφάλμα λόγω θορύβου. Η διαφορά είναι ότι στο (Bi & Zhang, 2004) δεν έχουμε Gaussian θόρυβο αλλά απλά ένα μοντέλου φραγμένου θορύβου (με ελάχιστη και μέγιστη τιμή). Η υποστήριξη μηχανής διανυσμάτων λειτουργεί με δημιουργία ορίων ανάμεσα σε σύνολα τιμών. Στη συνέχεια το κενό που δημιουργείται από την SVM μπορεί να διαφοροποιηθεί με τη χρήση της αβεβαιότητας των τιμών που βρίσκονται

μέσα στο όριο. Για παράδειγμα, αν υπάρχουν μια σειρά από σημεία που βρίσκονται στη μια πλευρά του ορίου (με συγκεκριμένη πιθανότητα/αβεβαιότητα) το κενό μεταξύ ορίων μπορεί να αλλαχτεί από την μηχανή ταξινόμησης. Ο λόγος είναι επειδή η αβεβαιότητα του δεδομένου μπορεί να το κάνει να βρίσκεται από τη μία ή την άλλη πλευρά του ορίου. Το (Bi & Zhang, 2004) προτείνει ένα γεωμετρικό αλγόριθμο που βελτιστοποιεί την πιθανοτική διαφοροποίηση μεταξύ δύο κλάσεων στις δύο πλευρές του ορίου. Αυτό δεν διαφέρει πολύ από τις κλασικές SVM μόνο που εσωκλείει και την πιθανότητα μια τιμή να βρίσκεται σε κάποια τιμή του ορίου καθώς υπολογίζει την ανεξαρτησία των δύο κλάσεων.

5.2 Συσταδοποίηση (Clustering)

Η αβεβαιότητα επηρεάζει τη φύση των συστάδων (clusters) δεδομένων γιατί επηρεάζει την απόσταση μεταξύ διαφορετικών δεδομένων. Σύμφωνα με προϋπάρχουσες τεχνικές (Kriegel & Pfeifle, 2005) κανείς μπορεί να υπολογίσει πιθανοτικές αποστάσεις μεταξύ τιμών που ορίζονται πιθανοτικά. Αυτή η απόσταση ορίζεται σαν μια συνάρτηση κατανομής απόστασης που εμπεριέχει την πιθανότητα η απόσταση μεταξύ δύο στοιχείων να βρίσκεται μεταξύ δύο τιμών. Αν τα αβέβαια στοιχεία είναι X, Y τότε αν p(X,Y) η συνάρτηση πυκνότητας απόστασης μεταξύ τους, η πιθανότητα μια απόσταση d(X,Y) να βρίσκεται μεταξύ τιμών X, Y να είναι a, b είναι:

$$P(a \leq d(X,Y) \leq b) = \int_a^b p(X,Y)(z)dz$$

Το (Kriegel & Pfeifle, 2005) μπορεί και ορίζει την πιθανότητα δύο δεδομένα να είναι «δίπλα» το ένα στο άλλο έτσι ώστε κάθε σημείο που τα ενώνει να έχει πυκνότητα μεγαλύτερη από κάποια τιμή. Με άλλα λόγια, ορίζει τα

σημεία που το διάστημα που τα ενώνει αποτελείται από σημεία που με τη σειρά τους έχουν πυκνότητα μεγαλύτερη από κάποιο όριο. Ολόκληρη αυτή η τεχνική είναι εμπνευσμένη από τον DBSCAN αλγόριθμο του (Ester, Kriegel, Sander, & Xu, 1996). Ο αλγόριθμος ξεκινάει από μια μικρή περιοχή και συνεχίζει να προσθέτει σημεία που πληρούν τις προϋποθέσεις. Φυσικά είναι φανερό ότι όσο μεγαλύτερη είναι η τιμή φράγμα για την συνάρτηση πυκνότητας τόσο μεγαλύτερος θα είναι και ο αριθμός των συστάδων που σχηματίζονται. Η επέκταση του DBSCAN είναι ο FDBSCAN, είναι παρόμοια μόνο που η πυκνότητα τώρα είναι επίσης αβέβαια λόγω της φύσης των δεδομένων. Αυτό όμως κάνει με τη σειρά του την απόσταση αβέβαια καθώς στα σημεία που είναι μεταξύ δύο σημείων μπορεί να παρεμβάλλονται άλλα. Για το λόγο αυτό ο FDBSCAN προσθέτει και έναν επιπλέον περιορισμό φράγματος της πιθανότητας της απόστασης δύο σημείων.

Εκτός από τα παραπάνω, στο (Kriegel & Pfeifle, 2005) προτείνεται ο αλγόριθμος OPTICS ο οποίος κινείται στο ίδιο μήκος κύματος με τον DBSCAN. Η διαφορά τους είναι ότι ο OPTICS αφορά στην ιεραρχική πυκνότητα απόστασης. Με άλλα λόγια, η συνάρτηση πυκνότητας απόστασης τώρα φράσσεται από μια σειρά από ιεραρχικά φράγματα και όχι μόνο από μια τιμή όπως στο DBSCAN. Η απόσταση ανάμεσα σε δύο σημεία μπορεί να μην έχει την ίδια πυκνότητα και έτσι να χρειάζονται διαφορετικά όρια πυκνότητας πιθανότητας για να υπολογιστούν οι συστάδες. Στόχος είναι να οριστούν οι όροι ταξινόμησης σημείων ώστε όταν εφαρμοστεί ο αλγόριθμος DBSCAN να μπορούν να εξαχθούν οι συστάδες σε οποιοδήποτε επίπεδο για διαφορετικές τιμές πυκνότητας. Σημαντικό είναι να εξασφαλιστεί οι συστάδες διαφορετικών επιπέδων να είναι συνεπείς μεταξύ τους. Αξιοσημείωτο επίσης είναι ότι συστάδες που σχηματίζονται με μεγαλύτερα όρια πυκνότητα εμπεριέχονται μέσα στις συστάδες με μεγαλύτερα όρια. Για να επιτευχθεί αυτό το (Kriegel & Pfeifle, 2005) αποδεικνύει ότι η συσταδοποίηση πρέπει να γίνει πρώτα σε περιοχές υψηλής πυκνότητας και ύστερα σε περιοχές μικρότερης πυκνότητας. Το αποτέλεσμα του αλγορίθμου OPTICS δεν είναι η

συσταδοποίηση αυτή καθεαυτή αλλά η σειρά επεξεργασίας των σημείων. Δεδομένου ότι ο OPTICS προέρχεται από τον DBSCAN έτσι και ο OPTICS επεκτάθηκε σε F OPTICS.

Επιπλέον, μια ακόμα επέκταση ενός γνωστού αλγορίθμου (K-means) παρουσιάστηκε στο (Ngai, Kao, Chui, Cheng, Chau, & Yip, 2006). Ο νέος αλγόριθμος ονομάζεται UK-means. Κάθε συστάδα έχει ένα αντιπροσωπευτικό στοιχείο. Ένα στοιχείο ανατίθεται στη συστάδα όταν έχει τη μικρότερη απόσταση από το αντιπροσωπευτικό σημείο τις συστάδας αυτής σε σχέση με την απόσταση του από τα άλλα σημεία άλλων συστάδων. Το μοντέλο παρουσιάζει υψηλή υπολογιστική πολυπλοκότητα για να μπορέσει να υπολογίσει τις αποστάσεις και γι’ αυτό προτείνει και μια σειρά από τεχνικές βελτιστοποίησης και υπολογισμού με προϋποθέσεις τις απόστασης χωρίς να γίνεται αυτό δυνατό για όλες τις τιμές. Γενικά ο UK-means χρησιμοποιείται για αβέβαιες τιμές δεδομένων τοποθεσίας. Η βιβλιογραφία περιέχει και κάποιες άλλες σχετικές τεχνικές τις οποίες δεν παρουσιάζουμε εδώ.

5.3 Επαναλαμβανόμενα δεδομένα

Ένα ακόμα θέμα είναι η εξόρυξη συχνών και επαλαμβανόμενων τιμών (Chui, Kao, & Hung, 2007). Στο μοντέλο αυτό γίνεται η υπόθεση ότι το κάθε στοιχείο έχει μια πιθανότητα ύπαρξης στο αποτέλεσμα μια επερώτησης. Αυτό που μοντελοποιείται σε αυτή την περίπτωση είναι η πιθανότητα P(ι) ένα στοιχείο ι να βρίσκεται σε μια επερώτηση και αυτό το στοιχείο θεωρείται συχνά επαναλαμβανόμενο αν η αναμενόμενη χρήση του ξεπερνά κάποιο όριο.

Ο αλγόριθμος που κλήθηκε να δώσει λύση σε αυτό το ερώτημα είναι ο U-apriori ο οποίος μοιάζει με τον Apriori επεκτείνοντας τον σε διαφορετικά

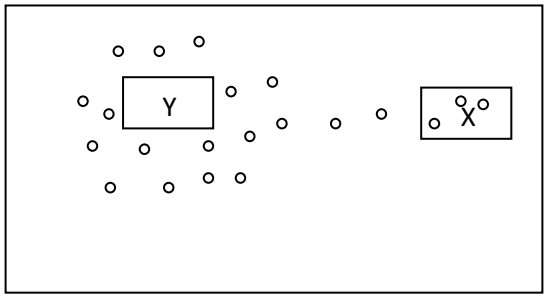
σύνολα τιμών. Συνοπτικά, η συχνότητα χρήσης ενός συνόλου τιμών S σε μια επερώτηση είναι ίση με το γινόμενο των επιμέρους πιθανοτήτων.

$$freq(S) = \prod_i P(i)$$

Περαιτέρω, η τεχνική την αποκοπής δεδομένων (data trimming) τα στοιχεία που έχουν πολύ μικρή πιθανότητα εμφάνισης σε μια επερώτηση δε συμμετέχουν στο σύνολο (trim) και τότε μπορεί να εφαρμοστεί ο U-apriori στα εναπομείναντα δεδομένα. Όπως αποδεικνύεται στο (Chui, Kao, & Hung, 2007) η παραπάνω τεχνική υπολογίζει τα συχνά επαναλαμβανόμενα δεδομένα με αποτελεσματικό τρόπο. Στη βιβλιογραφία επίσης μπορεί κανείς να βρει και άλλες μεθόδους αποκοπής των άχρηστων δεδομένων που κάνουν την εφαρμογή του αλγορίθμου πιο αποδοτική.

5.4 Εντοπισμός λαθών

Εντοπισμός λαθών είναι η δυνατότητα εξαγωγής λανθασμένων ή αβέβαιων τιμών μέσα από πολυάριθμα δεδομένα. Η ύπαρξη αβεβαιότητας σε διαφορετικές διαστάσεις κάνει αυτή τη δουλειά ακόμα πιο δύσκολη. Ας πάρουμε για παράδειγμα το Σχήμα 3 που παρουσιάζει τη αβεβαιότητα για δύο σημεία X,Y. Το σημείο Y φαίνεται ότι είναι πιο κοντά στη διασπορά των σημείων αλλά παρόλα αυτά δεν πέφτει κανένα σημείο σε αυτό. Αντίθετα το X είναι σημείο που περιέχει πιθανά σημεία παρόλο που δεν είναι πάνω στη διασπορά των πολλών σημείων. Το X όμως είναι πιο πιθανό να βρίσκεται μέσα στη κατανομή των δεδομένων.



Σχήμα 3: Αβεβαιότητα εντοπισμού λαθών

Ένα λάθος δεδομένο μπορεί να οριστεί σαν την πιθανότητα ένα σημείο να βρίσκεται σε μια πυκνή σε σημεία περιοχή της συνολικής περιοχής των σημείων. Το (Aggarwa & Yu, 2008) την ορίζει σαν η-πιθανότητα. Η-πιθανότητα ενός σημείου X_i, είναι η πιθανότητα το σημείο να βρίσκεται σε μια περιοχή με πυκνότητα τουλάχιστον η. Επειδή ο υπολογισμός της πυκνότητας αυτής προϋποθέτει υπολογισμό πυκνότητας σημείου και τομή του με την συνολική πυκνότητα περιορισμένη από η, είναι ευκολότερη η χρήση δείγματος για τον υπολογισμό της αντί για τον αναλυτικό υπολογισμό της. Με δειγματοληψία υπολογίζεται το υποσύνολο των σημείων για το οποίο το όριο πυκνότητας ζητείται. Τελικά, το (Aggarwa & Yu, 2008) ορίζει ένα (δ,η)-σφάλμα σαν το σημείο X_i για το οποίο αν η-πιθανότητα του, τότε για κάποιο υποχώρο ισχύει ότι $\eta < \delta$, καθώς και μια σειρά από αλγορίθμους βελτιστοποίησης που κάνουν τον αλγόριθμου πιο αποδοτικό.

5.5 Μέθοδοι εξόρυξης

Οι μέθοδοι που παρουσιάστηκαν παραπάνω είναι χρήσιμες και βρίσκουν εφαρμογή σε διάφορες εφαρμογές όπως η συσταδοποίηση και η ταξινόμηση. Ωστόσο, οι παραπάνω τεχνικές προϋποθέτουν ότι είναι γνωστή η συνάρτηση πυκνότητας πιθανότητας για την απόσταση μεταξύ

σημείων. Αυτό δεν είναι πάντοτε πρακτικά εφικτό χωρίς τη χρήση εκτεταμένης δειγματοληψίας μαζί με χρήση μαθηματικών μοντέλων. Το (Aggarwal C. , 2007) παρουσιάζει μια εναλλακτική μεθοδολογία που δεν έχει την προαναφερθείσα υπόθεση. Το (Aggarwal C. , 2007) κάνει μια πιο χαλαρή υπόθεση, ότι είναι γνωστά μόνο τα σφάλματα (κανονική απόκλιση) των τιμών και όχι όλη η συνάρτηση πυκνότητας πιθανότητας. Η υπόθεση αυτή είναι ρεαλιστική καθώς γνωρίζοντας τα σημεία μπορεί κανείς να υπολογίσει την απόκλιση και τη μέση τιμή για μια αβέβαιη μεταβλητή χωρίς χρήση ιδιαίτερων μαθηματικών μοντέλων. Φυσικά ακόμα και αν η συνάρτηση είναι γνωστή, κανείς μπορεί να εφαρμόσει τη μέθοδο εξάγοντας την απόκλιση από τη πυκνότητα πιθανότητας.

Το (Aggarwal C. , 2007), φιλοδοξεί και σχεδιάζει μια ενδιάμεση αναπαράσταση των δεδομένων η οποία μπορεί στη συνέχεια να χρησιμοποιηθεί για την εξόρυξη των δεδομένων (data mining). Η αναπαράσταση αυτή ονομάζεται προσαρμοσμένη προσέγγιση πυκνότητας (adjusted density estimate). Ονομάζεται προσαρμοσμένη γιατί εμπεριέχει την αβεβαιότητα όταν υπολογίζεται. Αν υποθέσουμε ότι η μέση τιμή των δεδομένων είναι X_i και η κανονική απόκλιση σ_i(), τότε η συνάρτηση πυκνότητα για N στοιχεία δίνεται από τον τύπο

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{h + \sigma_i(X_i)}{\sqrt{2\pi}} e^{\frac{-(x-X_i)^2}{2(h^2+\sigma_i(X_i)^2)}}$$

όπου h είναι μια παράμετρος (Hurst). Για περισσότερες λεπτομέρειες ο αναγνώστης μπορεί να ανατρέξει στη δημοσίευση. Όσον αφορά στις πιθανές πρακτικές χρήσεις εξόρυξης δεδομένων του παραπάνω μοντέλου, αυτές μπορεί να είναι:

- Υπάρχουν διάφορες μέθοδοι συσταδοποίησης που χρησιμοποιούν παρόμοιους αλγορίθμους πυκνότητας πιθανότητας. Συνήθως χρησιμοποιούν ένα όριο πυκνότητας για να απομονώσουν τις περιοχές (clusters).

- Υπάρχουν τεχνικές που χρησιμοποιούν ανώτατα όρια για να απομονώσουν «αραιές» περιοχές δεδομένων (Aggarwa & Yu, 2008). Στη συνέχεια αυτές οι περιοχές μπορεί να εντοπίσουν λανθασμένες ασυνήθεις τιμές (outlier) στα δεδομένα.
- Όπως παρουσιάζεται στο (Aggarwal C. , On Unifying Privacy and Uncertain Data, 2008) η παραπάνω τεχνική χρησιμοποιείται για ταξινόμηση.
- Γενικά οποιαδήποτε εφαρμογή χρησιμοποιεί τη συμπεριφορά-πυκνότητα των διαφορετικών περιοχών δεδομένων μπορούν να βρουν χρησιμότητα στη μέθοδο που παρουσιάστηκε στο (Aggarwal C. , 2007).

6. Εγχειρήματα

Όταν λέμε ότι ένας κλάδος έχει λάβει υψηλό βαθμό προσοχής, αυτό σίγουρα πρέπει να αντικατοπτρίζεται και στις ερευνητικές δουλειές που διενεργούνται σε διάφορα πανεπιστήμια και ερευνητικά κέντρα προκειμένου να έλθουν τα επιθυμητά αποτελέσματα. Σε αυτό το κεφάλαιο παρουσιάζουμε τα πιο γνωστά ερευνητικά projects μαζί με το αντικείμενο, τις σημαντικότερες προτάσεις και επιτυχίες τους. Σε γενικές γραμμές τα περισσότερα από αυτά ασχολούνται και σχεδιάζουν πιθανοτικές βάσεις με βάση συγκεκριμένες εφαρμογές. Πιο συγκεκριμένα, το

- Orion

από το αμερικανικό πανεπιστήμιο Purdue λειτουργεί στο θεωρητικό φάσμα των συνεχών τιμών και προσπαθεί να παρουσιάσει τεχνικές επεξεργασίας επερωτήσεων και σελιδοποίησης (indexing). Είναι βασισμένο σε PostgreSQL (ευρύτατα γνωστό DBMS σύστημα βάσεων δεδομένων ανοιχτού κώδικα) (Douglas, 2005). Οι πρακτικές εφαρμογές του project αυτού δε φαίνεται να είναι άμεσες άλλα έθεσε τις βάσεις για τη μετέπειτα εξέλιξη του τομέα (R. Cheng, 2003).

- ConQuer

Το project από το πανεπιστήμιο του Toronto ,εισάγει μια σειρά από αλγορίθμους μετασχηματισμού επερωτήσεων για την εξαγωγή «καθαρών» και συνεπών απαντήσεων από «βρώμικα» (αβέβαια) δεδομένα (P. Andritsos, 2006), (Fuxman, Fazli, & Miller, 2005). Το Conquer καταφέρνει να προτείνει μεθόδους που εξάγουν αποτελέσματα επερωτήσεων σε

πραγματικό χρόνο με συνέπεια. Επίσης, επιχειρεί να εξάγει πιθανοτικές τιμές για τα αβέβαια δεδομένα χρησιμοποιώντας τα δεδομένα σαν δείγμα.

- Trio

από το Stanford, το οποίο μοντελοποιεί, επεξεργάζεται και υπολογίζει την ακρίβεια δεδομένων και αναφέρθηκε στο κεφάλαιο 4.2.3. Το Trio έχει τη δική του γλώσσα (TrioQL) και λειτουργεί στα συμφραζόμενα μιας Uncertainty Lineage Database (ULDB).

Όσο αναφορά το σχεδιασμό του συστήματος, μπορεί να αποτιμήσει και κανονικά SQL queries, ενώ μέσω ενός ειδικού API μπορεί και μετατρέπει τα TrioQL queries σε παραδοσιακά queries. Η σχεσιακή (relational) DBMS βάση δεδομένων στην οποία βασίζεται διαθέτει μεθόδους κωδικοποίησης των πινάκων με βάση πιθανότητες και αποθηκευμένες συναρτήσεις (stored procedures) για τον υπολογισμό της ακρίβειας της αποτιμώμενης επερώτησης. Με βάση το ULDB μοντέλο η ΒΔ είναι σε θέση ανάλογα με την πηγή των δεδομένων να γνωρίζει την επιρροή τους στη ΒΔ. Με άλλα λόγια, μία «έξυπνη ΒΔ» σαν το Trio μπορεί να γνωρίζει πόσο το αποτέλεσμα μια επερώτησης επηρεάζεται από μια συγκεκριμένη πηγή πράγμα που μπορεί να χρησιμοποιηθεί για την αποδοχή του αποτελέσματος ή την απόρριψη του ως λανθασμένου (O. Benjelloun, 2006), (Das, Benjelloun, Halevy, & J., 2006), (O. Benjelloun D. S., 2006), (M. Mutsuzaki, 2007).

- MystiQ

από το πανεπιστήμιο της Wasighton, εισάγει μια νέα γλώσσα μοντελοποίησης (mDML), μια νέα γλώσσα ορισμού (mDDL) και μια νέα

μηχανή αναζήτησης επερωτήσεων. Πρακτικά το σύστημα διαφοροποιεί τις γλώσσες ώστε να εμπεριέχουν την έννοια της πιθανότητας. Η mDDL μπορεί να ορίσει επερωτήσεις με πιθανότητα, περιορισμούς, συναρτήσεις μέτρησης ομοιότητας κ.α . Όλα ενοποιημένα λειτουργούν με τη μηχανή αναζήτησης όπως διαφανώς μετατρέπονται σε συμβατικά SQL queries (Boulos, Dalvi, Mandhani, Mathur, Re, & Suciu, 2005), (Re, Dalv, & Suciu, 2007).

- URank

από το Waterloo University του Καναδά, που ασχολείται με U-Topk επερωτήσεις (κεφάλαιο 4.1.2). Βασίζεται στις παραδοσιακές ΒΔ (DBMS) προσθέτοντας κι αυτό δύο νέες έννοιες: το στρώμα αποθήκευσης (storage) και το στρώμα επεξεργασίας (processing). Μία νέα μηχανή «κανόνων» (rule) χρησιμοποιεί δεδομένα από το στρώμα αποθήκευσης υπολογίζοντας πιθανότητες και στη συνέχεια το στρώμα αποθήκευσης υπολογίζει το αποτέλεσμα της επερώτησης χρησιμοποιώντας ειδικές τεχνικές (M. A. Soliman I. F., 2007), (M. A. Soliman, 2007).

- MayBMS

στο πανεπιστήμιο των ΗΠΑ Cornell, επιτυγχάνει να επεκτείνει την PostgreSQL και να δημιουργήσει κι αυτό μια νέα γλώσσα. Επιπλέον, μεγιστοποιεί το χώρο για τα αβέβαια δεδομένα, πράγμα που δεδομένης της αβεβαιότητας είναι σημαντικό για την αναπαράστασή τους (L. Antova C. K., 2007), (L. Antova, 2008). Οι πρακτικές εφαρμογές και συμβολή του MayBMS είναι ιδιαίτερα σημαντικές.

- SPROUT

Στο πανεπιστήμιο του Oxford στόχος του project είναι να αναπτυχθούν τεχνικές για την εξελικτική επεξεργασία ερωτήσεων στις πιθανοτικές βάσεις δεδομένων και να χρησιμοποιηθούν για να χτιστεί μια "γερή" μηχανή ερωτήσεων(robust query engine) με την ονομασία SPROUT (S calable PRO cessing on U ncertain T ables).Αυτήν την περίοδο υπάρχουν τρεις κύριες ερευνητικές κατευθύνσεις(Jiewen Huang,2009).

α) Γίνετε έρευνα στα ανοικτά προβλήματα γύρω από την αποδοτική αξιολόγηση μιας ερώτησης(efficient query evaluation).Συγκεκριμένα, στοχεύουν στην ανακάλυψη κλάσεων βατών ερωτήσεων σε πιθανοτικές ΒΔ(δηλαδή υπολογίσιμων σε πολυωνυμικό χρόνο σε σχέση με την πολυπλοκότητα των δεδομένων)

β) Για την περίπτωση των μη-βατών ερωτήσεων, ερευνούν την κατά προσέγγιση αξιολόγηση ερώτησης (Rasmus Wissmann,2009). Σε αντίθεση με την ακριβή αξιολόγηση, που υπολογίζει απαντήσεις ερωτήσεων μαζί με τα ακριβή στοιχεία τους, η κατά προσέγγιση αξιολόγηση υπολογίζει τις απαντήσεις ερωτήσεων με τα κατά προσέγγιση στοιχεία τους.

γ)Η ανοιχτού κώδικα μηχανή επερωτήσεων για πιθανοτικά συστήματα διαχείρισης δεδομένων, χρησιμοποιούν τις ιδέες που αποκομίζονται από τις πρώτες δύο κατευθύνσεις που προαναφέραμε. Αυτή η μηχανή είναι βασισμένη στους αποδοτικούς αλγορίθμους αξιολόγησης δευτεροβάθμιας-αποθήκευσης, ακριβείς και κατά προσέγγιση για τις αυθαίρετες ερωτήσεις (Smitha Mysore-Shankar: Learning Probabilistic Databases, Oxford 2009).

- MCDM

Συνεργασία της IBM και του πανεπιστημίου της Florida το σύστημα σχεσιακής βάσης δεδομένων MCDB έχει αναπτυχθεί για τη διαχείριση των αβέβαιων στοιχείων, βασισμένων σε μια προσέγγιση του Monte Carlo. Η

MCDB αντιπροσωπεύει την αβεβαιότητα μέσω των VG(value genaration) συναρτήσεων που χρησιμοποιούνται για να παράγουν ψευδό-τυχαίες πραγματικές τιμές για αβέβαιες ιδιότητες. Οι VG συναρτήσεις μπορούν να παραμετροποιηθούν στα αποτελέσματα των SQL ερωτήσεων πέρα από τους πίνακες παραμέτρων(parameter tables) που αποθηκεύονται στη βάση δεδομένων, διευκολύνοντας την what-if ανάλυση (L. Antova, C. Koch, and D. Olteanu.,2007) Αποθηκεύοντας παραμέτρους, και όχι πιθανότητες, και εκτιμώντας παρά υπολογίζοντας την πιθανοτική κατανομή των πιθανών απαντήσεων στις ερωτήσεις το MCDB αποφεύγει πολλούς από τους περιορισμούς των προγενέστερων συστημάτων(Periklis Andritsos , Ariel Fuxman,2006).

- BayesStore

Το BayesStore(Berkley University) είναι μια νέα πιθανοτική αρχιτεκτονική διαχείρισης δεδομένων που στηρίζεται στην αρχή να χειρίζεται στατιστικά μοντέλα και πιθανοτικά εργαλεία συμπεράσματος ως πρώτης κλάσεως “πολίτες” του συστήματος βάσεων δεδομένων. Το BayesStore αντιπροσωπεύει μοντέλα δεδομένων και πραγματικά δεδομένα ως συγγενικούς πίνακες, εφαρμόζονται συμπερασματικοί αλγόριθμοι με καλή απόδοση στην SQL, προσθέτει πιθανοτικούς συγγενικούς χειριστές στη μηχανή ερωτήσεων, βελτιστοποιεί τις ερωτήσεις και με τους συγγενικούς χειριστές και με τους χειριστές συμπεράσματος.

Οι σχεδιαστικοί στόχοι του BayesStore είναι:

α) Να είναι σε θέση να υποστηρίξει αποδοτικά την επεξεργασία ερώτησεών ανάμεσα σε διαφορετικά μοντέλα έναντι των off-the-shelf εκμάθησης βιβλιοθηκών της μηχανής.

β) Να είναι σε θέση να υποστηρίξει το επεκτάσιμο API για τη σύνδεση σε νέα μοντέλα και σε νέους αλγορίθμους συμπεράσματος.

γ)Να είναι σε θέση να ανταποκριθεί σε πολύ μεγάλα σύνολα στοιχείων.

Περισσότερα για το BayesStore μπορούμε να δούμε στην ηλεκτρονική σελίδα που φιλοξενείται από το πανεπιστήμιο του Berkley http://www.cs.berkeley.edu/~daisyw/BayesStore.html.

7. Ανοιχτά/μελλοντικά θέματα

Μια σειρά από θέματα στον κλάδο των πιθανοτικών βάσεων δεδομένων δεν μπορούμε να πούμε ότι έχουν αντιμετωπιστεί αποτελεσματικά. Τα κυριότερα από αυτά έχουν να κάνουν με βελτιστοποιήσεις και επεκτάσεις επερωτήσεων. Παρακάτω παρουσιάζουμε μερικά από αυτά τα οποία δεν έχουν λυθεί, τουλάχιστον στη βιβλιογραφία που συναντήσαμε.

- Διχοτόμηση

Η ιδιότητα της διχοτόμησης είναι ένα βασικό αποτέλεσμα στον υπολογισμό επερωτήσεων πιθανοτικών ΒΔ και δίνει μια ολοκληρωμένη εικόνα για την πολυπλοκότητα συνδετικών (που περιέχουν ενώσεις) επερωτήσεων. Η διχοτόμηση δεν έχει οριστικά υπολογιστεί για τις πιθανοτικές βάσεις δεδομένων. Πιο συγκεκριμένα, αυτό αφορά στην επέκταση της διχοτόμησης σε κατηγορήματα (≤, ≠, <), ενώσεις (unions) και αυτό-ενώσεις (self-joins). Οι ενώσεις και αυτό-ενώσεις φαίνονται να συνδέονται καθώς $P(q1 \cup q2) = P(q1) + P(q2) - P(q1q2)$, όπου $q1$, $q2$ είναι δύο επερωτήσεις. Με άλλα λόγια, επερωτήσεις με αυτό-ενώσεις έχουν αποδειχτεί ότι διαθέτουν την ιδιότητα της διχοτόμησης σε ανεξάρτητες (independent) ΒΔ (N. Dalvi, 2007). Όσον αφορά στη διχοτόμηση όμως για χωρισμένες (disjoint) ανεξάρτητες ΒΔ, αυτή είναι ακόμα ανοιχτή.

- Βελτιστοποίηση με υπο-επερωτήσεις

Ακόμα και όταν είναι δύσκολος ο υπολογισμός μιας επερώτησης, αυτή ίσως να περιέχει υπο-επερωτήσεις όπου να είναι βατές. Σε αυτή την περίπτωση, ο επεξεργαστής επερωτήσεων μιας ΒΔ θα μπορεί να εργαστεί για τον απολογισμό των υπο-επερωτήσεων και στη συνέχεια να τρέξει fully

polynomial-time randomized approximation scheme (FPTRAS). Ο FPTRAS είναι ένα αλγόριθμος υπολογισμού κατά προσέγγιση για προβλήματα βελτιστοποίησης. Ο αλγόριθμος αυτός πρέπει να λύνεται σε πολυωνυμικό χρόνο ως προς το μέγεθος του προβλήματος. Το (Re, Dalv, & Suciu, 2007) παρουσιάζει ενθαρρυντικά πειραματικά αποτελέσματα, αλλά ένα γενικό υπόβαθρο για τέτοιο είδους βελτιστοποιήσεις δεν έχει δημιουργηθεί ακόμα. Για να παρουσιάσουμε ένα παράδειγμα, ας υποθέσουν την επερώτηση $q = R(x,y), S(y,z), T(y,z,u)$. Αν αυτή μπορεί να διασπασθεί σε $q = R(x,y)\mathrm{SQ(y)}$ όπου $\mathrm{SQ(y)}$ είναι ένα προσωρινός πίνακας που διαθέτει τον υπολογισμό της υπο-επερώτηση στην οποία διασπάστηκε η αρχική, τότε παρόλο που η αρχική δυσκολία της επερώτησης δεν άλλαξε, ο υπολογισμός της με χρήση FPTRAS είναι πιο εφικτή καθώς το αρχικό πρόβλημα έχει καταστεί μικρότερο.

- Πιθανοτική αναγωγή

Στη βιβλιογραφία υπάρχουν πολλές αναφορές αναγωγής πιθανοτικών προβλημάτων σε δίκτυα Markov και Bayes (πχ κεφάλαιο 4.1.2). Η αναγωγή αυτή μπορεί να χρησιμοποιηθεί για το πρόβλημα υπολογισμού επερωτήσεων σε πιθανοτικές ΒΔ. (Y. Zabiyaka, 2006) και (Dalvi & Suciu, 2007) παρουσιάζουν μια τεχνική όπου με χρήση δυαδικών (binary) και Δ-δέντρων (dtree) ένα query ανάγεται σε ένα Δ-δέντρο με πιθανότητες μετάβασης στα φύλλα του. Η προτεινόμενη λύση δεν είναι όμως πρακτική καθώς είναι δύσκολος ο υπολογισμός του δέντρου. Στο μέλλον ίσως σχετικά εγχειρήματα ανοίξουν νέους δρόμους.

- Διαρροή πληροφορίας

Όπως έχει ήδη αναφερθεί, η προσθήκη λανθασμένων πληροφοριών είναι μερικές φορές θεμιτή για την προστασία των δεδομένων. Ας υποθέσουμε ότι έχουμε μια ΒΔ I και θέλουμε να κάνουμε διαθέσιμη μια όψη (view) υ,

χωρίς όμως να είναι δυνατός ο υπολογισμός κάποιας επερώτησης q από αυτή την όψη. Το πρόβλημα της διαρροής πληροφορίας είναι πως και αν μπορεί να υπολογίσει κανείς το q αν γνωρίζει το υ;

Το (Miklau & Suciu, 2004) περιγράφει πως αν γνωρίζουμε την κατανομή πιθανότητας P της I μπορούμε να υπολογίσουμε την πιθανότητα P(q). Οπότε μετά τη δημοσίευση του υ, κάποιος εξωτερικός παράγοντας μπορεί να υπολογίσει τη δεσμευμένη πιθανότητα P(q|υ). Για να αποφευχθεί η διαρροή πληροφορίας πρέπει να ισχύει ότι P(q|υ)≈P(q) και τότε λέμε ότι το q είναι ασφαλές ως προς το υ. Σκοπός είναι να διαπιστωθεί αν το q είναι ασφαλές ως προς το υ και αν υπάρχουν πολλές όψεις υ1,υ2 αν P(q|υ1,υ2)≈P(q). Το (Miklau & Suciu, 2004) αντιμετωπίζει ως ασφαλείς όλες τις κατανομές P για τις οποίες q είναι ασφαλές ως προς υ. Ωστόσο αυτό δεν είναι απόλυτα σωστό καθώς υπάρχουν πολλά q που είναι πρακτικά ασφαλή ως προς υ, ενώ επίσης δεν αποδεικνύει ότι αν P(q|υ1)≈P(q) και P(q|υ2)≈P(q) μπορεί να μην ισχύει P(q|υ1,υ2)≈P(q). Τελικά, στη βιβλιογραφία δεν έχει αποτελεσματικά αντιμετωπιστεί πως μπορεί να αποδειχτεί για μια βάση ότι $\lim_{n\to\infty}P(q|\upsilon)=1$ και $\lim_{n\to\infty}P(q|\upsilon 1,u2)=0$, όπου n είναι το μέγεθος του χώρου και που πρακτικά θα έκανε τη διαρροή πληροφορίας ένα λυμένο πρόβλημα.

- Γραφικές αναπαραστάσεις

Γραφικά μοντέλα θα μπορούσαν να έχουν εφαρμογές και σε πιθανοτικές ΒΔ (Verma & Pearl, 1990). Τα γραφικά μοντέλα είναι μια σειρά συσχετίσεων τυχαίων μεταβλητών που στη συνέχεια δημιουργούν μια αναπαράσταση της κατανομής τους. Μια πιθανοτική ΒΔ δεδομένων μπορεί να αναπαρασταθεί σαν μια όψη μιας χωρισμένης ανεξάρτητης ΒΔ της οποίας η πολυπλοκότητα εξαρτάται από το γραφικό μοντέλο που περιγράφει την αρχική βάση. Έτσι μπορεί να γίνουν υπολογισμοί σε πιθανοτικές βάσεις μέσω αναγωγής του σε μοντέλο μιας όψης μιας ανεξάρτητης ΒΔ. Ωστόσο, δεν έχει γίνει σημαντική δουλειά που να περιγράφει τα παραπάνω αποτελεσματικά.

- Περιορισμοί

Όταν τα δεδομένα είναι αβέβαια, τότε περιορισμοί μπορούν να χρησιμοποιηθούν ώστε να βελτιωθεί η ποιότητα τους και άρα να αποτελέσουν ένα σημαντικό εργαλείο για τα πιθανοτικά δεδομένα. Μέχρι στιγμής δεν έχει υπάρξει ένας γενικά αποδεικτός ορισμός περιορισμών για πιθανοτικές ΒΔ.

Ας υποθέσουμε ότι έχουμε ένα περιορισμό Γ που είναι μια πρόταση που πρέπει να ισχύει για όλες τις πιθανές λέξεις της ΒΔ. Τότε μια επερώτηση q μετατρέπεται σε μια δεσμευμένη πιθανότητα $P(q|\Gamma)=\frac{P(q,\Gamma)}{P(\Gamma)}$. Η αποτίμηση του P(q,Γ) είναι το επόμενο ανοιχτό πρόβλημα. Ένα ακόμη θέμα είναι αυτό των «μαλακών» (soft) περιορισμών, όπως για παράδειγμα «πολλοί αγοράζουν τις ίδια ηλεκτρονικές συσκευές» και πως θα μπορούσαν αυτοί να εισαχθούν στη ΒΔ.

8. Συμπέρασμα

Στην παραπάνω εργασία παρουσιάσαμε το σύνολο της πρόσφατης δουλειάς που έχει γίνει στον κλάδο τον βάσεων δεδομένων που περιέχουν αβέβαια δεδομένα. Ξεκινήσαμε από μια εισαγωγή που εξηγούσε το υπόβαθρο καθώς και τους λόγους που το ερευνητικό ενδιαφέρον έχει στραφεί προς αυτή την κατεύθυνση. Παρουσιάσαμε σύντομα εφαρμογές και παραδείγματα όπου πιθανοτικά δεδομένα αποθηκεύονται και χρησιμοποιούνται. Εξηγήσαμε τα βασικά προβλήματα και προκλήσεις που παρουσιάζονται και συνοψίσαμε τα σημαντικότερα μοντέλα για τα πιθανοτικά δεδομένα και τις κατηγορίες τους. Στη συνέχεια, ο αναγνώστης γνώρισε συνοπτικά και επιγραμματικά τις σημαντικότερες τεχνικές επεξεργασίας των δεδομένων και πως γίνεται η εξόρυξή τους. Τέλος, σημαντικό είναι ότι ο αναγνώστης γνωρίζει τα σημαντικότερα projects και τα άλυτα προβλήματα του κλάδου.

Κλείνοντας αυτή την εργασία, κανείς πρέπει να γνωρίζει συνοπτικά και επιγραμματικά τον τομέα των πιθανοτικών ΒΔ, χωρίς να μπορούμε να πούμε ότι θα είναι σε θέση να έχει καταλάβει διεξοδικά τις τεχνικές λεπτομέρειες κάθε μεθόδου που παρουσιάστηκε στο παρόν κείμενο. Για την κατανόηση των επιμέρους κανείς θα έπρεπε να ανατρέξει στις δημοσιεύσεις και τη βιβλιογραφία. Ωστόσο θεωρούμε ότι το θέμα έχει καλυφθεί σφαιρικά και παρέχει το υπόβαθρο για κάποιον να είναι σε θέση να εκτιμήσει τις βασικές έννοιες και αιτίες ανάπτυξης του τομέα, να ξεκινήσει να ασχολείται με τις λεπτομέρειες, ή να αρχίσει να ερευνά τον κλάδο ή να έχει μια σφαιρική εικόνα γύρω από τον κλάδο των πιθανοτικών βάσεων δεδομένων.

Βιβλιογραφία

Aggarwa, C., & Yu, P. (2008). Outlier Detection with Uncertain Data. *Proc. SIAM Int'l Conf. Data Mining (SDM).*

Aggarwal, C. (2009). Managing and Mining Uncertain Data. *Springer* .

Aggarwal, C. (2007). On Density Based Transformations for Uncertain Data Mining. *Proc. 23rd IEEE Int'l Conf. Data Eng. (ICDE).*

Aggarwal, C. (2008). On Unifying Privacy and Uncertain Data. *Proc. 24th IEEE Int'l Conf. Data Eng (ICDE).*

Aggarwal, C., & Yu, P. S. (2007). *A survey of uncertain data algorithms.* IBM Research Report.

Bi, J., & Zhang, T. (2004). Support Vector Machines with Input Data Uncertainty. *Proc. Advances in Neural Information Processing Systems (NIPS).*

Boulos, J., Dalvi, N., Mandhani, B., Mathur, S., Re, C., & Suciu, D. (2005). MystiQ: A system for finding more answers by using probabilities. *SIGMOD.*

Burdick, D., Deshpande, P., Jayram, T., Ramakrishnan, R., & Vaithyanathan, S. (2005). OLAP over Uncertain and Imprecise Data,". *Proc. 31st Int'l Conf. Very Large Data Bases (VLDB)*, (σσ. 970-981).

Cheng, R., Kalashnikov, D., & Prabhakar, S. (2004). Querying Imprecise Data in Moving Object Environments. *IEEE Trans. Knowledge and Data Eng.*, (σσ. pp. 1112-1127).

Cheng, R., Xia, Y., Prabhakar, S., Shah, R., & Vitter, J. (2004). Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data. *Proc. 30th Int'l Conf. Very Large Data Bases (VLDB).*

Cheng, R., Xia, Y., Prabhakar, S., Shah, R., Vitter, J., & Xia, Y. (2005). *Efficient Join Processing over Uncertain Data.* CSD TR# 05-004, Dept. of Computer Science, Purdue Univ.

Chui, C.-K., & Kao, B. (2008). A Decremental Approach for Mining Frequent Itemsets from Uncertain Data. *Proc. 12th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD).*

Chui, C.-K., Kao, B., & Hung, E. (2007). Mining Frequent Itemsets from Uncertain Data. *Proc. 11th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD).*

Dabbish, v. A., & Dabbish, L. (2004). Labeling images with a computer game. *ACM Conference on Human Factors in Computing Systems.* Vienna, Austria.

Dalvi, N., & Suciu, D. (2004). Efficient Query Evaluation on Probabilistic Databases. *Proc. 30th Int'l Conf. Very Large Data Bases (VLDB).*

Dalvi, N., & Suciu, D. (2007). Management of probabilistic data: foundations and challenges. *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, (σσ. 1-12).

Das, S., Benjelloun, O., Halevy, A., & J., W. (2006). Working Models for Uncertain Data. *Proc. 22nd IEEE Int'l Conf. Data Eng.(ICDE).*

Deshpande, A., Guestrin, C., Madden, S., Hellerstein, J. M., & Hong, W. (2005). Using probabilistic models for data management in acquisitional environments. *Conference on Innovative Data Systems (CIDR)*, (σσ. 317–328).

Doan, A., Ramakrishnan, R., Chen, F., DeRose, P., Lee, Y., McCann, R., και συν. (2006). Community information management. *IEEE Data Engineering Bulletin, Special Issue on Probabilistic* , 64–72.

Douglas, K. (2005). *PostgreSQL.* Σαμσ.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (KDD),.*

Flickr. (n.d.). Ανάκτηση από flickr: www.flickr.com

Fuhr, N., & Rolleke, T. (1997). A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems. *ACM Trans. Information Systems.*

Fuxman, A., Fazli, E., & Miller, R. (2005). Conquer: Efficient Management of Inconsistent Databases. *Proc. ACM SIGMOD.*

Galindo, J., Urrutia, A., & Piattini, M. *Fuzzy databases: Modeling, design, and implementation.* Idea Group Publishing.

Google Base. (n.d.). Ανάκτηση από google: base.google.com

Halevy, A. (2007, January). *Alon Halevy's Blog*. Ανάκτηση από Blogspot: http://alonhalevy.blogspot.com/2007/01/uncertainty-and-data-integration.html

Heckerman, D. (2002). *Tutorial on graphical models.*

Hua, M., Pei, J., & Zhang, W. L. (2008). Efficiently answering probabilistic threshold top-k queries on uncertain data. *ICDE.*

J. Considine, F. L. (2004). Approximate aggregation techniques for sensor database. *ICDE.* Boston, Massachusetts.

Kohler, S. P. (2006). Addressing the problems. *Nature Reviews Genetics* , 481–488.

Kriegel, H.-P., & Pfeifle, M. (2005). Density-Based Clustering of Uncertain Data. *Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD).*

Kriegel, H.-P., & Pfeifle, M. (2005). Hierarchical Density Based Clustering of Uncertain Data. *Proc. Fifth IEEE Int'l Conf. Data Mining (ICDM).*

Kriegel, H.-P., Kunath, P., Pfeifle, M., & Renz, M. (2006). Probabilistic Similarity Join over Uncertain Data. *Proc. 11th Int'l Conf. Database Systems for Advanced Applications (DASFAA).*

L. Antova, C. K. (2007). Worlds and Beyond: Efficient Representation and Processing of Incomplete Information. *Proc. 23rd IEEE Int'l Conf. Data Eng. (ICDE).*

L. Antova, T. J. (2008). Fast and Simple Relational Processing of Uncertain Data. *Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE).*

Lakshmanan, L., Leone, N., Ross, R., & Subrahmanian, V. (1997). ProbView: A Flexible Database System. *ACM Trans. Database Systems*, (σσ. 419-469).

Lee, S. K. (1992). Imprecise and uncertain information in databases: an evidential approach. *ICDE.*

Lim, & Shekhar, S. (1996). An Evidential Reasoning Approach to Attribute Value Conflict Resolution in Database Integration. *IEEE Transactions on Knowledge and Data Engineering*, (σσ. Vol. 8, No. 5).

Ljosa, V., & Singh, A. (2008). Top-k Spatial Joins of Probabilistic Objects. *Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE).*

M, B., & al, e. (2007). Data management in the world-wide sensor web. *IEEE Pervasive Computing.*

M. A. Soliman, I. F. (2007). Top-k query processing in uncertain databases. *ICDE.*

M. A. Soliman, I. F.-C. (2007). Urank: Formulation and efficient evaluation of top-k queries in uncertain databases. *SIGMOD.*

M. Keulen, A. K. (2005). A Probabilistic XML Approach to Data Integration. *Proc. 21st IEEE Int'l Conf. Data Eng.*

M. Mutsuzaki, M. T. (2007). TrioOne: Layering uncertainty and lineage on a conventional dbms. *CIDR.*

Manjhi, A., Nath, S., & Gibbons, P. B. (2005). Tributaries and deltas: E±cient and robust aggregation in sensor network. *SIGMOD.* Baltimore, Maryland.

Miklau, G., & Suciu, D. (2004). A formal analysis of information disclosure in data exchange. *SIGMOD.*

N. Dalvi, D. S. (2007). The dichotomy of conjunctive queries on random structures. *PODS.*

Ngai, W., Kao, B., Chui, C., Cheng, R., Chau, M., & Yip, K. (2006). Efficient Clustering of Uncertain Data. *Proc. Sixth IEEE Int'l Conf. Data Mining (ICDM).*

Nierman, A., & Jagadish, H. (2002). ProTDB: Probabilistic Data in XML. *Proc. 28th Int'l Conf. Very Large Data Bases (VLDB).*

Nilesh Dalvi, D. S. (2007). Management of Probabilistic Data Foundations and Challenges. *Proceedings of 26th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, (σσ. 1-12).

O. Benjelloun, A. D. (2006). ULDBs: Databases with Uncertainty and Lineage. *Proc. 32nd Int'l Conf.Very Large Data Bases (VLDB).*

O. Benjelloun, D. S. (2006). An introduction to ULDBs and the Trio system. *IEEE Data Engineering Bulletin*, (σσ. 5-15).

P. Andritsos, A. F. (2006). Clean Answers over Dirty Databases: A Probabilistic Approach,". *Proc. 22nd IEEE Int'l.*

Pearl, J. (1988). Probabilistic reasoning in intelligent systems.

Pei, J., Jiang, B., Lin, X., & Yuan, Y. (2007). Probabilistic Skylines on Uncertain Data. *Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB).*

R. Cheng, D. K. (2003). Evaluating Probabilistic Queries over Imprecise Data., (σ. Proceedings. ACM SIGMOD).

Re, C., Dalv, N., & Suciu, D. (2007). Efficient top-k query evaluation on probabilistic data. *ICDE.*

Rolleke, N., & FuhrT. (1997). *A probabilistic NF2 relational algebra for imprecision in databases.*

Sarawagi, R. G. (2006). Creating probabilistic databases from information extraction models. *International Conference on Very Lage Databases*, (σσ. 965–976).

Sarma, A. D., Benjelloun, O., Halevy, A., & Widom, J. (2005). Working models for uncertain data. *ICDE.*

Sarma, A. D., Benjelloun, O., Halevy, A., & Widom, J. (2006). Working Models for Uncertain Data. *Proc. 22nd IEEE Int'l Conf. Data Eng.*

Sen, P., & Deshpande, A. (2007). Representing and querying correlated tuples in probabilistic databases. *ICDE.*

Singh, S., Mayfield, C., Prabhakar, S., Shah, R., & Hambrusch, S. (2006). Indexing Uncertain Categorical Data. *Proc. 22nd IEEE Int'l Conf Data Eng. (ICDE).*

Soliman, M., Ilyas, I., & Chang, K.-C. (2007). Top k-Query Processing in Uncertain Databases. *Proc. 23rd IEEE Int'l Conf. Data Eng.(ICDE).*

Suciu, D., & Dalvi, N. (2005). Answering Queries from Statistics and Probabilistic Views. *VLDB.*

Sunter, I. F. (1969). A theory for record. *Journal of the American Statistical Society,* , 1183–1210.

T.S. Jayram, R. K. (2006). Avatar information extraction system. *IEEE Data Engineering Bulletin* , 40–48.

Tao, Y., Cheng, R., Xiao, X., Ngai, W. K., Kao, B., & Prabhakar, S. (2005). Indexing multi-dimensional uncertain data with arbitrary probability density functions. *VLDB.*

Verma, T., & Pearl, J. (1990). Causal networks: Semantics and expressiveness. *Uncertainty in Artificial Intelligence*, (σσ. 69-76).

Y. Zabiyaka, A. D. (2006). Functional treewidth Bounding complexity in the presence of functional dependencies. *SAT*, (σσ. 116–129).

Yi, K., Li, F., Srivastava, D., & Kollios, G. (2007). *Efficient processing of top-k queries in uncertain databases.* Florida State University.

Zhang, W., Lin, X., Pei, J., Zhang, Y., & Fraser, S. (2008). Managing Uncertain Data: Probabilistic Approaches. *The 9th International Conference on Web-Age Information Management (WAIM)*, (σσ. 405-412).

Zhao, G. B. (2006). A Framework for Clustering Evolving Data Streams. *UW-MSR Summer Institute Semiahmoo Resort.* Blaine, Wasighton.

Parag Agrawal , Omar Benjelloun , Anish Das Sarma , Chris Hayworth , Shubha Nabar , Tomoe Sugihara , Jennifer Widom, Trio: a system for data, uncertainty, and lineage, Proceedings of the 32nd international conference on Very large data bases, September 12-15, 2006, Seoul, Korea

Periklis Andritsos , Ariel Fuxman , Renee J. Miller, Clean Answers over Dirty Databases: A Probabilistic Approach, Proceedings of the 22nd International Conference on Data Engineering, p.30, April 03-07, 2006 [doi>10.1109/ICDE.2006.35]

L. Antova, C. Koch, and D. Olteanu. 10106 worlds and beyond: Efficient representation and processing of incomplete information. In ICDE, pages 606--615, 2007.

Bahar Biller , Barry L. Nelson, Modeling and generating multivariate time-series input processes using a vector autoregressive technique, ACM Transactions on Modeling and Computer Simulation (TOMACS), v.13 n.3, p.211-237, July 2003 [doi>10.1145/937332.937333]

Doug Burdick , AnHai Doan , Raghu Ramakrishnan , Shivakumar Vaithyanathan, OLAP over imprecise data with domain constraints,

Proceedings of the 33rd international conference on Very large data bases, September 23-27, 2007, Vienna, Austria

Amol Deshpande, Carlos Guestrin, Sam Madden, Joseph M.Hellerstein, and Wei Hong.Model-driven data acquisition in sensor networks.In International Conference on Very Large Data Bases, 2004

Debabrata Dey and Sumit Sarkar.A probabilistic relational model and algebra.ACM Transactions on Database Systems., 1996

Nilesh Dalvi and Dan Suciu.Efficient query evaluation on probabilistic databases.In International Conference on Very Large Data Bases, 2004.

Norbert Fuhr and Thomas Rolleke.A probabilistic relational algebra for the integration ofinformation retrieval and database systems.ACM Transactions on Information Systems, 1997.

Sunil Choenni, Henk Ernst Blok, and Erik Leertouwer.Handling uncertainty and ignorance in databases: A rule tocombine dependent data.In Database Systems for Advanced Applications, 2006.