Έργο:	**«ΘΑΛΗΣ: Ενίσχυση της Διεπιστημονικής ή και Διιδρυματικής έρευνας και καινοτομίας με δυνατότητα προσέλκυσης ερευνητών υψηλού επιπέδου από το εξωτερικό μέσω της διενέργειας βασικής και εφαρμοσμένης έρευνας αριστείας»**
Τίτλος Υποέργου:	**«ΕΙΚΟΣ»: Θεωρητική και αλγοριθμική θεμελίωση για Προσωποκεντρικά Συνεργατικά Πληροφοριακά Συστήματα**

Παραδοτέο Π.3.1

Διαχείριση δεδομένων συμπεριφοράς χρήστη

Σεπτέμβριος 2015

Δράση 5	**Contextualization και εξατομίκευση περιεχομένου και παρουσίαση της πληροφορίας**				
Ομάδα	Ερ. Ομάδα 3	**Έναρξη**	01/02/2012	**Λήξη**	30/11/2015
Συντονιστής ΕΟ3	Ι. Ιωαννίδης (ΕΚΠΑ)				
Υποδράση: ΥΔ 31	Διαχείριση συμπεριφοράς χρήστη				
Συμμετέχοντες	*Μέλη ΚΕΟ*	Ι. Ιωαννίδης (ΕΚΠΑ), Ε. Πιτουρά (Παν. Ιωαννίνων), Μ. Χατζόπουλος (ΕΚΠΑ)			
	Μέλη ΟΕΣ	Α. Κατηφόρη (ΕΚΠΑ), Μ. Κυριακίδη (ΕΚΠΑ), Κ. Σεμερτζίδης (Παν. Ιωαννίνων)			

Σύντομη Περιγραφή	Η Υποδράση ΥΔ3.1 ερευνά την ανάπτυξη μοντέλων και αλγοριθμικών μεθόδων που επιτρέπουν την αποδοτική αποθήκευση και τη γενικευμένη αναπαράσταση της συμπεριφοράς και των επιλογών του χρήστη στον υπερχώρο δεδομένων, καθώς και για την επεξεργασία των καταγεγραμμένων δεδομένων, που οδηγεί στην εξαγωγή ποικίλων στατιστικών και προτύπων συμπεριφοράς χρήστη. Στο πλαίσιο της υποδράσης παρέχουμε μια αρχική θεμελίωση για πρότυπα πρότασης συμπερασμάτων προς το χρήστη, τα οποία στο μέλλον μπορεί να βασίζονται σε πρότυπα συμπεριφοράς ή προτιμήσεων. Στη συνέχεια προτείνουμε ένα ευρετήριο στο χρήστη το οποίο του επιτρέπει τη γρήγορη πλοήγηση της ιστορίας ενός γραφήματος. Έτσι του δίνεται η δυνατότητα να διατυπώσει ερωτήματα που αφορούν οποιοδήποτε χρονικό διάστημα με βάσει τις προτιμήσεις του και να ανακτήσει την ιστορία του γραφήματος και να ερευνήσει το περιβάλλον που διαδραματιζόταν γύρω από αυτόν την εκάστοτε χρονική στιγμή. Τέλος, προτείνουμε ένα τρόπο διεπαφής του χρήστη με ένα πληροφοριακό σύστημα που του επιτρέπει τη πλοήγηση στη βάση δεδομένων. Με αυτόν τον τρόπο, ξεκινώντας από ένα βασικό ερώτημα, ο χρήστης με τη βοήθεια του συστήματος συστάσεων, μπορεί να διατυπώνει ερωτήματα με βάση τις προσωπικές του προτιμήσεις και τις ατομικές πληροφοριακές του ανάγκες.
Παραδοτέο	Π.3.1 Διαχείριση συμπεριφοράς χρήστη
Στόχος στο Τ.Δ.	Τεχνική αναφορά που θα περιλαμβάνει τουλάχιστον 2 δημοσιεύσεις.
Επίτευξη στόχου	100%

Περιεχόμενα

1 Εισαγωγή

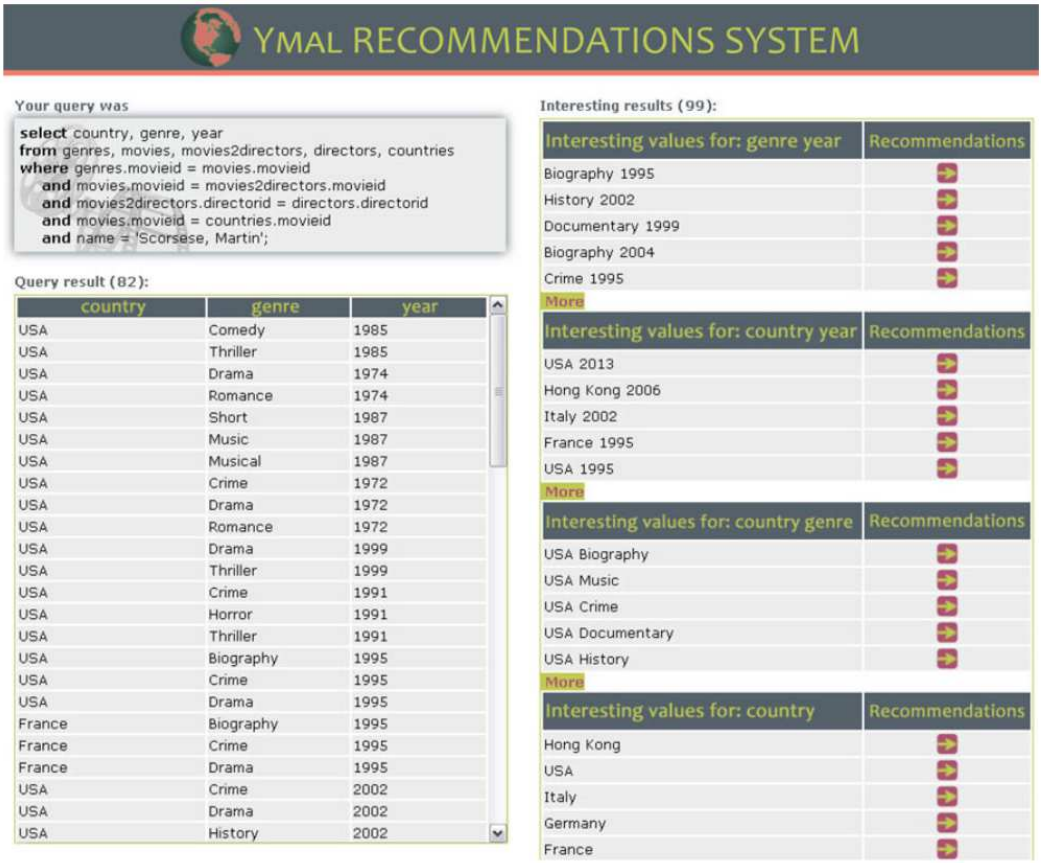
Ο βασικός στόχος του έργου ΕΙΚΟΣ είναι να προσφέρει τη μεθοδολογία, τη θεωρητική θεμελίωση, τις αλγοριθμικές τεχνικές και την αρχιτεκτονική του λογισμικού που απαιτείται ώστε τα πληροφοριακά συστήματα να μπορούν να προσφέρουν στους χρήστες αφενός την δυνατότητα εξατομίκευσης της παρεχόμενης πληροφορίας και αφετέρου τη δυνατότητα χρήσης ενσωματωμένων ετερογενών δεδομένων, ενδεχομένως διαφορετικής προέλευσης, με διαφανή τρόπο.

Στα πλαίσια του έργου, η Δράση 3 «Contextualization και εξατομίκευση περιεχομένου και παρουσίασης της πληροφορίας» αποσκοπεί στον ορισμό γενικευμένων και εκφραστικών μοντέλων, στην εξαγωγή αλγοριθμικών αποτελεσμάτων για την παροχή εξατομικευμένων και προσαρμόσιμων υπηρεσιών σε πληροφοριακά συστήματα, καθώς και στο γενικότερο σχεδιασμό ενός ισχυρού συστήματος εξατομίκευσης, ικανού να ανταποκριθεί και να ικανοποιήσει τις διαρκώς αυξανόμενες και μεταβαλλόμενες απαιτήσεις πληθώρας εφαρμογών. Η Δράση οργανώνεται σε τρεις θεμελιώδεις δράσεις, εκ των οποίων η πρώτη αφορά τη διαχείριση και τη συντακτική και σημασιολογική ολοκλήρωση των δεδομένων για τη συμπεριφορά και το περιβάλλον του χρήστη, ενώ η δεύτερη αφορά τη μοντελοποίηση του χρήστη και τη διαχείριση των προφίλ χρηστών, τα οποία περιέχουν πληροφορία για της προτιμήσεις και τα χαρακτηριστικά τους. Τέλος, η τρίτη δράση αφορά τον ευρύτερο σχεδιασμό ενός γενικευμένου συστήματος εξατομίκευσης, έχοντας υπόψη τα διαθέσιμα μοντέλα και τις ορισμένες τεχνικές από τις δύο προηγούμενες δράσεις.
Το παρόν Παραδοτέο Π.3.1 περιλαμβάνει τα αποτελέσματα της υποδράσης ΥΔ3.1. Στις ενότητες 2-4 παρουσιάζουμε τις ερευνητικές εργασίες μας και τα αποτελέσματα τους που αφορούν την υποδράση 3.1. Ανακεφαλαιώνουμε τα αποτελέσματά μας στην ενότητα 5.

2 Σύστημα Διαχείρισης Βάσεων Δεδομένων που βασίζεται στην πλοήγηση

Ο τυπικός τρόπος αλληλεπίδρασης ενός χρήστη με ένα Σύστημα Διαχείρισης Βάσεων Δεδομένων γίνεται μέσω της υποβολής ερωτημάτων (queries) διατυπωμένων σε κάποια γλώσσα ερωτημάτων. Ωστόσο, συχνά οι χρήστες μπορεί να μη γνωρίζουν πλήρως το περιεχόμενο της βάσης δεδομένων ή να μην έχουν σαφή κατανόηση των πληροφοριακών τους αναγκών. Αυτό συμβαίνει ιδιαίτερα στις περιπτώσεις των υπερχώρων δεδομένων λόγω του αυξημένου όγκου των δεδομένων και της ποικιλομορφίας των πηγών από τις οποίες προέρχονται αυτά τα δεδομένα.

Ο στόχος στην εργασία [1], είναι η παρουσίαση ενός νέου τρόπου διεπαφής ενός χρήστη με ένα Σύστημα Διαχείρισης Βάσεων Δεδομένων που βασίζεται στην πλοήγηση. Συγκεκριμένα, υπολογίζουμε και προτείνουμε (recommend) στους χρήστες δεδομένα που αν και δεν ανήκουν στο αποτέλεσμα του αρχικού τους ερωτήματος σχετίζονται στενά με αυτό. Αυτά τα αποτελέσματα ονομάζονται Ymal (“You May Also Like”) αποτελέσματα. Για παράδειγμα, ας υποθέσουμε ότι ένας χρήστης ρωτά για τα χαρακτηριστικά (όπως είδος, έτος ή χώρα παραγωγής) ταινιών από ένα συγκεκριμένο σκηνοθέτη, π.χ. Μ. Σκορσέζε. Το σύστημά μας θα αναδείξει τις ενδιαφέρουσες πτυχές αυτών των αποτελεσμάτων, π.χ. ενδιαφέρουσα χρόνια, τα ζεύγη του είδους και των χρονιών, και ούτω καθεξής (Σχήμα 1).

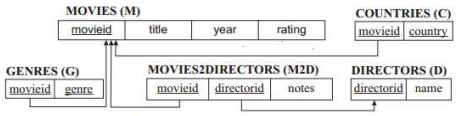


Σχήμα 1 YmalDB: Στην αριστερή πλευρά, το αρχικό ερώτημα του χρήστη Q εμφανίζεται στη κορυφή και το αποτέλεσμά του στο κάτω μέρος. Το Q ρωτά για τις χώρες, τα είδη και χρόνια των ταινιών που σκηνοθέτησε ο Μ. Σκορσέζε. Στη δεξιά πλευρά, ενδιαφέροντα μέρη του αποτελέσματος παρουσιάζονται ομαδοποιημένα με βάση τα χαρακτηριστικά τους και κατατάσσονται κατά σειρά ενδιαφέροντος.

Ο υπολογισμός των YMAL αποτελεσμάτων βασίζεται στον εντοπισμό ενδιαφερουσών τιμών και συνόλων τιμών στο αποτέλεσμα της αρχικής ερώτησης. Ο υπολογισμός του βαθμού ενδιαφέροντος ενός συνόλου τιμών γίνεται με βάση τη συχνότητα εμφάνισης του συνόλου στο αποτέλεσμα αλλά και στη βάση δεδομένων. Για τον υπολογισμό της συχνότητας ενός συνόλου τιμών στη βάση δεδομένων χρησιμοποιούμε μια πρωτότυπη μέθοδο που στηρίζεται στη διατήρηση ενός αντιπροσωπευτικού συνόλου από σπάνια itemsets.

Συγκεκριμένα, το αποτέλεσμα της ερώτησης του Σχήματος 2, περιέχει τον ίδιο αριθμό ταινιών τύπου "Biography" και "Drama". Ωστόσο αν ο τύπος "Biography" εμφανίζεται λιγότερες φορές σε όλη την βάση απ' ότι ο τύπος "Drama", τότε θεωρούμε τις ταινίες τύπου "Biography" πιο ενδιαφέρουσες. Για τον σκοπό αυτό υπολογίζουμε το ποσοστό εμφάνισης στο αποτέλεσμα ενός χαρακτηριστικού του ερωτήματος και το ποσοστό εμφάνισης του σε όλη την βάση. Επεκτείνουμε το σύστημα προτάσεων συνενώνοντας τα πιο σημαντικά χαρακτηριστικά του

αποτελέσματος ενός ερωτήματος με άλλα στοιχεία που απουσιάζουν στην αρχική ερώτηση ώστε να παρέχουμε στον χρήστη έξτρα πληροφορία που μπορεί να τον ενδιαφέρει. Τέλος επειδή το πλήθος των σπάνιων χαρακτηριστικών ενός αποτελέσματος (που μπορεί να είναι σημαντικά για τον χρήστη) μπορεί να αυξηθεί αρκετά παρουσιάζουμε στην εργασία αυτή έναν αλγόριθμο που εξάγει τα τοπ-κ σημαντικότερα χαρακτηριστικά.



(a) Database schema.

```
SELECT M.title, M.year, G.genre
FROM D, M2D, M, G
WHERE D.name = "M. Scorsese"
      AND D.directorid = M2D.directorid
      AND M2D.movieid = M.movieid
      AND M.movieid = G.movieid
```

(b) Example query.

M.title	M.year	G.genre
The Aviator	2004	Biography
Gangs of New York	2002	Drama
Goodfellas	1990	Biography
Casino	1995	Drama
Shutter Island	2004	Thriller
M. Jackson: Video Greatest Hits	1995	Drama
The Last Waltz	1978	Biography
Raging Bull	1980	Documentary

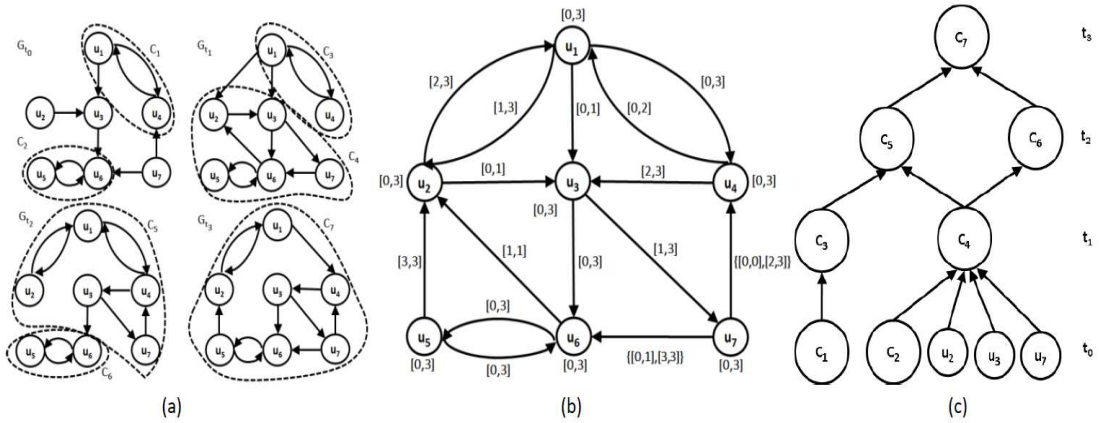
(c) Query result.

Σχήμα 2(a) σχήμα βάσης, (b) παράδειγμα ερωτήματος και (c) το αποτέλεσμα του ερωτήματος.

3 Χρονικό Ευρετήριο για γρήγορη Ιστορική πλοήγηση ενός γραφήματος.

Δεδομένου ότι τα περισσότερα γραφήματα εξελίσσονται με την πάροδο του χρόνου, είναι χρήσιμο να υπάρχει η δυνατότητα να διερευνήσουμε την ιστορία τους. Στην εργασία [2], θεωρούμε ιστορικά ερωτήματα που ζητούν την ύπαρξη μιας διαδρομής μεταξύ δύο κόμβων σε κάποιο χρονικό διάστημα στο παρελθόν, είτε σε όλη τη διάρκεια του διαστήματος (συνδετικά ερωτήματα), ή τουλάχιστον σε μία χρονική στιγμή (διαζευκτικά ερωτήματα) στο διάστημα.

Μελετάμε δύο εναλλακτικές λύσεις, πρώτον, την αποθήκευση της πλήρης μεταβατική κλειστότητα «Transitive Closure» του εξελισσόμενου γραφήματος και δεύτερον, την απευθείας διάσχιση του. Στη συνέχεια, προτείνουμε ένα ευρετήριο προσβασιμότητας, που ονομάζεται ευρετήριο *TimeReach*, το οποίο εκμεταλλεύεται το γεγονός ότι τα περισσότερα πραγματικά γραφήματα περιέχουν μεγάλες ισχυρές συνδεδεμένες συνιστώσες.



Σχήμα 3 (a) Εξελισσόμενο γράφημα, (b) το αντίστοιχο έκδοση γράφημα (version graph) και (c) η εξέλιξη των συνδεδεμένων συνιστωσών.

Συγκεκριμένα στο σχήμα 3(a) δείχνουμε την εξέλιξη ενός γραφήματος τις χρονικές στιγμές t_0, t_1, t_2, t_3, την αναπαράσταση που χρησιμοποιούμε για να απεικονίσουμε την εξέλιξη του γραφήματος 3(b) και τέλος 3(c) την εξέλιξη των ισχυρών συνιστωσών. Δείξαμε ότι είναι χρονικά απαγορευτικό να υπολογίσουμε τη πλήρη μεταβατική κλειστότητα ακόμη και για μερικές χιλιάδες κόμβους. Ωστόσο αν μπορούμε να υπολογίσουμε την μεταβατική κλειστότητα σε μικρά γραφήματα, η εύρεση αν ένα ζευγάρι κόμβων συνδέεται με ένα μονοπάτι κάποια χρονική στιγμή γίνεται σε σταθερό χρόνο. Στην περίπτωση της απλής διάσχισης χρησιμοποιούμε έναν τροποποιημένο BFS αλγόριθμο που διασχίζει τον "version graph" ώστε να βρει αν υπάρχει σύνδεση μεταξύ δύο κόμβων. Αντίστοιχα στην περίπτωση του TimeReach αν δύο κόμβοι ανήκουν σε κάθε χρονική στιγμή στην ίδια ισχυρή συνιστώσα τότε από τον ορισμό των ισχυρών συνιστωσών οι συγκεκριμένοι κόμβοι συνδέονται.

Επεκτείνουμε το TimeReach κρατώντας τα γραφήματα συνιστωσών για κάθε χρονική στιγμή ώστε στη περίπτωση που οι κόμβοι u, v ανήκουν σε διαφορετικές συνιστώσες, ο u κόμβος συνδέεται με ένα μονοπάτι από την συνιστώσα του στην συνιστώσα του v κόμβου. Έτσι όταν δύο κόμβοι ανήκουν σε διαφορετικές ισχυρές συνιστώσες, διασχίζουμε τα γραφήματα των συνιστωσών. Για να αποφύγουμε την αποθήκευση όλης της παραπάνω πληροφορίας, παρουσιάζουμε το συνοπτικό TimeReach (Condensed TimeReach) ευρετήριο που μετατάσσει τα id των συνιστωσών ώστε ο αριθμός τους να ελαχιστοποιείται θέτοντας νέα id συνιστωσών στις περιπτώσεις που ένας κόμβος ανήκει σε διαφορετική ισχυρή συνιστώσα σε σύγκριση με τις

προηγούμενες χρονικές στιγμές. Τέλος, χρησιμοποιούμε από την υπάρχουσα βιβλιογραφία τη προσέγγιση 2hop στο Condensed TiemReach, που απαντάει γρήγορα αν δύο κόμβοι συνδέονται χωρίς να χρειαστεί να κάνουμε κάποια διάσχιση.

4 Σύστημα επεξεργασίας OLAP ερωτήσεων σε Βάση Δεδομένων.

Στην παρούσα εργασία [3], εξετάζουμε πώς μπορούμε να εκμεταλλευτούμε την ύπαρξη ενός σχήματος αστέρα (star schema), προκειμένου να απαντηθούν OLAP ερωτήματα των χρηστών με CineCube movies. Για να παραχθεί μια ταινία (movie), χρειαζόμαστε κείμενο ήχο και τα "επεισόδια" που θα την απαρτίζουν. Για το σκοπό αυτό η μέθοδος που υλοποιήθηκε σε ένα πραγματικό σύστημα, περιλαμβάνει τα παρακάτω βήματα. Ο χρήστης υποβάλλει ένα OLAP ερώτημα σε ένα υπάρχον σχήμα αστέρα. Λαμβάνοντας αυτό το ερώτημα ως είσοδο, το σύστημα παράγει ένα σύνολο από ερωτήματα που συμπληρώνουν το περιεχόμενο των πληροφοριών του αρχικού ερωτήματος, και τα εκτελεί. Στη συνέχεια, το σύστημα οπτικοποιεί τα αποτελέσματα του κάθε ερωτήματος και συνοδεύει την παρουσίαση τους με κείμενο το οποίο σχολιάζει τα σημαντικά μέρη των αποτελεσμάτων. Επιπλέον, μέσω ενός συστήματος μετατροπής κειμένου σε ήχο, το σύστημα μας παράγει αυτόματα ήχο για το κείμενο που δημιουργούμε. Κάθε συνδυασμός της απεικόνισης, του κειμένου και του ήχου αποτελεί ουσιαστικά μία CineCube movie, η οποία υλοποιείται ως μια παρουσίαση του PowerPoint και επιστρέφεται στον χρήστη.

5 Ανακεφαλαίωση

Το παρόν παραδοτέο Π3.1 παρουσιάζει τα αποτελέσματα της υποδράσης ΥΔ3.1 του έργου ΕΙΚΟΣ. Ο στόχος της υποδράσης ΥΔ3.1 ήταν η ανάπτυξη μοντέλων και αλγοριθμικών μεθόδων που επιτρέπουν την αποδοτική αποθήκευση και τη γενικευμένη αναπαράσταση της συμπεριφοράς και των επιλογών του χρήστη στον υπερχώρο δεδομένων, καθώς και για την επεξεργασία των καταγεγραμμένων δεδομένων. που οδηγεί στην εξαγωγή ποικίλων στατιστικών και προτύπων συμπεριφοράς χρήστη.

Στα πλαίσια της διερεύνησής μας, λοιπόν, επιτύχαμε να ανταποκριθούμε στο στόχο της υποδράσης με τους ακόλουθους τρόπους:

1. Κατασκευάσαμε ένα πρωτότυπο τρόπο διεπαφής του χρήστη με ένα πληροφοριακό σύστημα που επιτρέπει στο χρήστη τη πλοήγηση στη βάση δεδομένων. Με αυτόν τον τρόπο, ξεκινώντας από ένα βασικό ερώτημα, ο χρήστης με τη βοήθεια του συστήματος συστάσεων, μπορεί να διατυπώνει ερωτήματα με βάση τις προσωπικές του προτιμήσεις και τις ατομικές πληροφοριακές του ανάγκες. Το προτεινόμενο σύστημα συστάσεων δημοσιεύθηκε στο VLDBJ (2013).
2. Κατασκευάσαμε ένα ευρετήριο το οποίο επιτρέπει τη γρήγορη πλοήγηση της ιστορίας ενός γραφήματος. Συγκεκριμένα το ευρετήριο μπορεί γρήγορα να εξακριβώσει αν υπήρχε μια σύνδεση μεταξύ δύο κόμβων μέσω ενός μονοπατιού σε κάποιο χρονικό διάστημα στο παρελθόν είτε σε όλη τη διάρκεια του διαστήματος. Με αυτόν τον τρόπο, ο χρήστης μπορεί να διατυπώνει ερωτήματα που αφορούν οποιοδήποτε χρονικό διάστημα με βάσει τις προτιμήσεις του και να ανακτήσει την ιστορία του γραφήματος και να ερευνήσει το περιβάλλον που διαδραματιζόταν γύρω από αυτόν την εκάστοτε χρονική στιγμή. Τα αποτελέσματα μας δημοσιεύθηκαν στο EDBT (2015).
3. Παρείχαμε μια αρχική θεμελίωση για πρότυπα πρότασης συμπερασμάτων προς το χρήστη, τα οποία στο μέλλον μπορεί να βασίζονται σε πρότυπα συμπεριφοράς ή προτιμήσεων. Η παρούσα εργασία συνεισφέρει και στην υλοποίηση εργαλείων διαχείρισης της εξέλιξης καθώς τα αποτελέσματα εντάχθηκαν στο εργαλείο Cinecubes. Η παρούσα εργασία είναι το αρχικό βήμα για την συνέχιση της ερευνητικής εργασίας σε οικοσυστήματα υπερχώρων. Τα αποτελέσματα μας δημοσιεύθηκαν στο DOLAP (2013).

Δημοσιεύσεις

[1] M. Drosou and E. Pitoura. YMALDB: exploring relational databases via result-driven recommendations. In VLDB Journal, volume 22, pages 849–874, 2013.

[2] K. Semertzidis, K. Lillis, and E. Pitoura. Timereach: Historical

reachability queries on evolving graphs. In EDBT , pages 121–132, 2015.

[3] D. Gkesoulis, P. Vassiliadis, and P. Manousis. CineCubes: aiding data workers gain insights from OLAP queries. In Information Systems, pages 60–86 , 2015.

Παράρτημα