ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Παππάς Αλέξανδρος

Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτ. Εργασίας: Καθηγητής Δημήτρης Πλεξουσάκης

Παρασκευή, 10/2/2017, 16:00

Αίθουσα B108, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

"Εξερεύνηση μέτρων σημαντικότητας για δημιουργία συνόψεων σε βάσεις δεδομένων γράφων"

ΠΕΡΙΛΗΨΗ

Ο πραγματικός κόσμος είναι πλούσια διασυνδεδεμένος. Ως εκ τούτου οι φυσικές ιδιότητες των γραφημάτων, τα καθιστούν εξαιρετικά χρήσιμα στη μοντελοποίηση του πραγματικού κόσμου και την κατανόηση μια ευρείας ποικιλίας συνόλων δεδομένων, προσφέροντας παράλληλα εφαρμόσιμες λύσεις σε διάφορους τομείς της βιομηχανίας. Μια βάση δεδομένων γραφημάτων, είναι ένα επιχειρησιακό σύστημα διαχείρισης βάσεων δεδομένων, το οποίο μπορεί να εκτελέσει μεθόδους δημιουργίας, ανάγνωσης, ενημέρωσης και διαγραφής, οι οποίες εκθέτουν ένα μοντέλο δεδομένων γράφου. Διαφέροντας από τις παραδοσιακές σχεσιακές βάσεις δεδομένων, οι βάσεις δεδομένων γραφημάτων έχουν βελτιστοποιηθεί και σχεδιαστεί κυρίως για διεργασίες πάνω σε δεδομένα γράφων, αποδοτικότερη διάσχιση των δεδομένων και εκτέλεση αλγορίθμων γράφων σε πολύπλοκες ιεραρχικές δομές.

Με δεδομένη την εκθετική αύξηση στο μέγεθος και την πολυπλοκότητα των δεδομένων του διαδικτύου, εκτιμάται ότι μέχρι το τέλος του 2018, το 70% των κορυφαίων οργανισμών θα αξιοποιεί μία ή περισσότερες βάσεις δεδομένων γραφημάτων. Οι τριπλέτες αποθήκευσης αποτελούν μια υποκατηγορία των βάσεων δεδομένων γραφημάτων, η οποία διαμορφώθηκε και μοντελοποιήθηκε βασισμένη στις προδιαγραφές του Resource Description Framework (RDF) και σχεδιάστηκε ως ένας επισημασμένος, κατευθυνόμενος, πολυγράφος.

Προς αυτή την κατεύθυνση, υπάρχει τώρα περισσότερο από ποτέ ανάγκη για την ανάπτυξη μεθόδων και εργαλείων, προκειμένου να διευκολυνθεί η κατανόηση και η εξερεύνηση των RDF γνωσιακών βάσεων δεδομένων. Λαμβάνοντας υπόψη το γεγονός ότι ο ανθρώπινος εγκέφαλος μπορεί να ερμηνεύσει μόνο μερικές εκατοντάδες κόμβους σε ένα γράφημα, τότε είναι προφανές ότι το μέγεθος των σημερινών δεδομένων και η πολυπλοκότητα του σχήματος είναι εκτός των δυνατοτήτων εξερεύνησης που μπορούν να προφέρουν οι μέθοδοι αυτοματοποιημένων σχεδιασμών.

Ως προς την επίλυση αυτού του προβλήματος, οι μέθοδοι συνόψισης επιδιώκουν την παραγωγή μιας συνοπτικής έκδοσης της αρχικής πηγής δεδομένων, αναδεικνύοντας τις πιο αντιπροσωπευτικές έννοιες. Βασικά ερωτήματα για την παραγωγή μιας συνόψισης είναι: το πως θα προσδιοριστούν οι σημαντικότεροι κόμβοι ενός συνόλου και εν συνεχεία το πώς θα συνδεθούν προκειμένου να παράγει έναν έγκυρο υπογράφο. Σε αυτή την εργασία, προσπαθούμε να απαντήσουμε το πρώτο ερώτημα με την χρήση και την προσαρμογή σε γνωσιακές βάσεις δεδομένων, μέτρα σημαντικότητας τα οποία έχουν ήδη ερευνηθεί στο παρελθόν, ώστε να καλύψουν ένα ευρύ φάσμα διαφορετικών δεδομένων για την επιλογή των πιο σημαντικών κόμβων. Έπειτα μοντελοποιούμε το πρόβλημα της διασύνδεσης των κόμβων ως ένα Δέντρο Στάινερ σε γράφημα, το οποίο ανήκει σε προβλήματα συνδυαστικής βελτιστοποίησης, με κοινό ζητούμενο να βρεθεί η συντομότερη διασύνδεση για ένα ορισμένο σύνολο κόμβων. Δεδομένου ότι το πρόβλημα αυτό ανήκει στην κατηγορία των δυσεπίλυτων προβλημάτων, διερευνήσαμε τρεις προσεγγιστικούς αλγορίθμους, χρησιμοποιώντας ευρηστικά τεχνάσματα τα οποία επιταχύνουν την εκτέλεση τους, για την επίλυση του προβλήματος σε πολυωνυμικό χρόνο. Μέσω της διεξαγωγής λεπτομερών πειραμάτων εμφανίζουμε την προστιθέμενη αξία της προσέγγισης μας, δεδομένου ότι α) οι προσαρμογές μας ξεπερνούν τις τρέχουσες τεχνικές υψηλού επιπέδου μέτρων σημαντικότητας για την επιλογή των πιο σημαντικών κόμβων και β) η παραγόμενη σύνοψη έχει καλύτερη ποιότητα, εισάγοντας μικρότερο αριθμό πρόσθετων κόμβων, καθώς οι προσεγγιστικοί αλγόριθμοι του Δέντρου Στάινερ αποδίδουν καλύτερα από τις μεθόδους οι οποίες έχουν χρησιμοποιηθεί στο παρελθόν.

Pappas Alexandros
M.Sc. Thesis
Computer Science Department
University of Crete
Master's Thesis Supervisor: Professor Dimitris Pleksousakis

Friday, 27/1/2017, 16:00
Room B108, Computer Science dept., University of Crete

"Exploring Importance Measures for Summarization on Graph Databases"

ABSTRACT

The real world is richly interconnected. As such the natural properties of graphs, render them extremely useful in modeling real world, understanding a wide diversity of data-sets and offering applied solutions in different fields of industry. A graph database is an on-line, operational database management system with Create, Read, Update, and Delete (CRUD) methods that expose a graph data model. Alternative to traditional relational databases, graph databases are being optimized and designed predominantly for graph workloads, traversal performance and executing graph algorithms on complex hierarchical structures.

Given the explosive growth in the size and the complexity of the Data Web, it is estimated that by the end of 2018, 70% of leading organizations will have one or more utilizing graph databases. Triple stores are a subcategory of graph databases, modeled around the Resource Description Framework (RDF) specifications and designed as labeled, directed multi-graphs.

To this direction, there is now more than ever, an increasing need to develop methods and tools in order to facilitate the understanding and exploration of RDF/S Knowledge Bases (KBs). Given the fact that the human brain can only interpret at most a few hundred nodes in one chart it becomes obvious that current data size and schema complexity are far beyond the exploration capability that any automated layout can provide.

Summarization approaches try to produce an abridged version of the original data source, highlighting the most representative concepts. Central questions to summarization are: how to identify the most important nodes and then how to link them in order to produce a valid sub-schema graph. In this thesis, we try to answer the first question by revisiting several measures covering a wide range of alternatives for selecting the most important nodes and adapting them for RDF/S KBs. Then, we proceed further to model the problem of linking those nodes as a graph Steiner-Tree problem (GSTP). Since the GSTP is NP-complete, we explore three approximations (SDIST, CHINS and HEUM) employing heuristics to speed up the execution of the respective algorithms. Our detailed experiments show the added value of our approach since a) our adaptations outperform current state of the art measures for selecting the most important nodes and b) the constructed summary has a better quality in terms of the additional nodes introduced to the generated summary as GSTP approximations outperform past approaches.