Αξιοποίηση της συσχέτισης μεταξύ λέξεων για τη βελτίωση του προσεγγιστικού φιλτραρίσματος πληροφορίας

Σε ένα σύστημα φιλτραρίσματος πληροφορίας, ή αλλιώς σύστημα έκδοσης/συνδρομής, οι χρήστες εγγράφονται σε πηγές πληροφορίας (εκδότες) και ενημερώνονται όταν εκδίδονται νέα έγγραφα που τους ενδιαφέρουν. Λόγω του τεράστιου όγκου των ενημερώσεων που προωθούνται από τους εκδότες στους συνδρομητές του συστήματος, ένα σενάριο προσεγγιστικού φιλτραρίσματος πληροφορίας κρίνεται απαραίτητο. Στο προσεγγιστικό φιλτράρισμα πληροφορίας, οι συνδρομητές επιλέγουν και παρακολουθούν συγκεκριμένους εκδότες, οι οποίοι αποθηκεύουν τοπικά το συνεχές ερώτημα (συνδρομή) του χρήστη. Οι εκδότες που επιλέγονται για κάθε συνδρομητή και μόνο αυτοί προωθούν ενημερώσεις όταν νέα έγγραφα ταιριάζουν με τη συνδρομή του.

Η επιλογή κατάλληλων εκδοτών για κάποιο συνεχές ερώτημα που υποβάλλεται γίνεται με χρήση στατιστικών μεταδεδομένων. Τα στατιστικά στοιχεία που αποθηκεύονται σε καταλόγους του συστήματος, σχετίζονται με το καλύτερο σύνολο όρων και όχι εγγράφων ανά εκδότη. Το γεγονός αυτό, σε συνδυασμό με την ανά όρο οργάνωση των καταλόγων, όπου δε λαμβάνονται υπόψη πιθανές συσχετίσεις μεταξύ των λέξεων, οδηγεί σε χαμηλά επίπεδα ανάκλησης.

Για παράδειγμα, ας θεωρήσουμε έναν χρήστη, ο οποίος υποβάλλει το συνεχές ερώτημα q = {Ιωάννινα, λίμνη} στο σύστημα. Το τυπικό σύστημα φιλτραρίσματος πληροφορίας προωθεί στο χρήστη τα έγγραφα που είναι συναφή με το ερώτημα και προέρχονται από κάθε εκδότη του συστήματος. Στο αντίστοιχο κατά προσέγγιση σύστημα, το συνεχές ερώτημα διαχωρίζεται σε δύο όρους και με χρήση των στατιστικών υπολογίζεται ένα συναθροιστικό αποτέλεσμα για κάθε όρο και εκδότη. Η προσέγγιση αυτή για την επιλογή κατάλληλου εκδότη μπορεί να οδηγήσει σε μειωμένη ανάκληση. Αυτό συμβαίνει γιατί οι υψηλά διατεταγμένοι εκδότες για κάθε όρο μπορεί να μην έχουν την ίδια καλή διάταξη για το συνεχές ερώτημα εξ' ολοκλήρου.

Για τη διατήρηση και διάδοση στατιστικών στοιχείων των εκδοτών χρησιμοποιούνται τεχνικές για τη σύνοψη του τοπικού περιεχομένου τους. Τα Hash Sketches και οι

KMV Synopses αποτελούν γνωστές τεχνικές που χρησιμοποιούνται για να υπολογίσουν τον πληθάριθμο ενός πολυσυνόλου *S*. Τα Hash Sketches [1] αποτελούν δομές που βασίζονται στη χρήση του κατακερματισμού. Οι τιμές εισόδου μέσω μιας συνάρτησης κατακερματισμού αντιστοιχίζονται ομοιόμορφα σε ένα σύνολο τιμών εξόδου. Η βασική ιδιότητα των Hash Sketches έγκειται στην ευκολία τους να συνδυάζονται. Για παράδειγμα, για να υπολογίσουμε το Hash Sketch της ένωσης δύο πολυσυνόλων *A* και *B* αρκεί να πάρουμε το λογικό OR των Hash Sketches των επιμέρους πολυσυνόλων. Ο πληθάριθμος της τομής υπολογίζεται με χρήση του γνωστού τύπου $|A \cap B| = |A| + |B| - |A \cup B|$ από τη θεωρία συνόλων. Ο συνδυασμός μεγάλου αριθμού Hash Sketches οδηγεί σε μη ακριβή υπολογισμό των διαφορετικών τιμών. Επιπλέον, πρόβλημα αποτελεί το μεγάλο κόστος για τον υπολογισμό τους.

Οι KMV Synopses [2] έχουν χαμηλότερο υπολογιστικό κόστος και μεγαλύτερη ακρίβεια αποτελεσμάτων σε σχέση με τα Hash Sketches. Βασίζονται στην ιδέα των DV εκτιμητών. Ας υποθέσουμε ότι *D* σημεία κατανέμονται ομοιόμορφα σε ένα διάστημα τιμών. Η αναμενόμενη απόσταση δύο γειτονικών σημείων είναι $1/(D+1) \approx 1/D$, οπότε η αναμενόμενη τιμή του *k* μικρότερου σημείου θα είναι $E[U_k] = k/D$. Επομένως $D = k/E[U_k]$ και ο βασικός DV εκτιμητής για το πλήθος των σημείων θα είναι $D_k = k/U_k$. Χρησιμοποιώντας την ιδέα του βασικού εκτιμητή, η KMV Synopsis ενός πολυσυνόλου *S* με πεδίο ορισμού το $\theta(S)$ δημιουργείται εφαρμόζοντας μια συνάρτηση κατακερματισμού σε κάθε τιμή του $\theta(S)$. Έπειτα καταγράφονται οι *k* μικρότερες κατακερματισμένες τιμές. Η σύνοψη αυτή του πολυσυνόλου *S* ονομάζεται KMV Synopsis για *k* μικρότερες τιμές. Χρησιμοποιώντας την επέκταση του βασικού εκτιμητή $D_k = (k-1)/U_k$, καταλήγουμε σε εκτιμητές των διαφορετικών τιμών της ένωσης και της τομής δύο πολυσυνόλων *A* και *B*.

Όσον αφορά την αρχιτεκτονική του συστήματος φιλτραρίσματος πληροφορίας, αυτό αποτελείται από τους εκδότες, τους συνδρομητές και τους καταλόγους. Οι εκδότες ενημερώνουν το σύστημα για στατιστικά στοιχεία σχετικά με έγγραφα που υπάρχουν στο τοπικό περιεχόμενό τους. Τα στατιστικά περιέχουν για κάθε όρο μια ανεστραμμένη λίστα εγγράφων, το μέγεθός της, καθώς και πρόσθετες πληροφορίες (π.χ. συχνότητες) που βοηθούν τους συνδρομητές να διατάξουν τους εκδότες. Οι εκδότες αποθηκεύουν τα στατιστικά στοιχεία με τη χρήση Hash Sketches ή KMV Synopses και τα διανέμουν περιοδικά στους κατάλληλους καταλόγους του συστήματος.

Ο ρόλος των καταλόγων του συστήματος είναι η αποθήκευση των στατιστικών και η προώθηση τους στους συνδρομητές που τα ζητούν. Οι κατάλογοι διατηρούνται από υπερ-κόμβους και είναι οργανωμένοι σε ένα Chord DHT [3]. Η συνάρτηση κατακερματισμού διαμερίζει το σύνολο των όρων του συστήματος έτσι ώστε για κάθε όρο να υπάρχει μοναδικός κατάλογος για τον οποίο είναι υπεύθυνος.

Κάθε συνδρομητής χρησιμοποιεί στατιστικά στοιχεία των καταλόγων για να διατάξει κατάλληλα τους εκδότες. Οι υψηλά διατεταγμένοι εκδότες θα δημοσιεύσουν με μεγάλη πιθανότητα στο μέλλον σχετικά έγγραφα με το συνεχές ερώτημα του χρήστη, το οποίο αποθηκεύουν τοπικά. Εκεί, το ερώτημα ταιριάζει με κάθε νέο έγγραφο που δημοσιεύεται. Για την διάταξη των εκδοτών, ο συνδρομητής λαμβάνει υπόψη το τοπικό περιεχόμενο τους και επιπλέον χρησιμοποιεί τεχνικές πρόβλεψης της μελλοντικής συμπεριφοράς τους. Η διαδικασία της διάταξης εκτελείται περιοδικά λόγω του δυναμικού χαρακτήρα των δημοσιεύσεων.

Για την αύξηση των επιπέδων ανάκλησης του συστήματος προτείνεται η αξιοποίηση πιθανής συσχέτισης μεταξύ όρων σε σύνολα όρων. Η υπό συνθήκη πιθανότητα ότι ένα τυχαίο έγγραφο μιας συλλογής περιέχει τον όρο *a*, δεδομένου ότι περιέχει τον όρο *b* είναι:

$$P(A|B)=\frac{df(ab)/|D|}{df(b)/|D|}=\frac{df(ab)}{df(b)}$$

όπου *df(ab)* είναι το πλήθος των εγγράφων που περιέχουν το ζεύγος όρων *ab*, *df(b)* το πλήθος των εγγράφων που περιέχουν τον όρο *b* και $|D|$ το πλήθος των εγγράφων της συλλογής.

Για να ανακαλύψουμε ενδιαφέροντα σύνολα όρων για τα οποία θα κρατήσουμε στατιστικά στοιχεία μια προσέγγιση είναι να επιλέγουμε για σύνολα όρων συνεχή ερωτήματα, τα οποία εκτελούνται διαρκώς στο σύστημα. Καλή ιδέα ακόμη είναι να κρατάμε σύνολα όρων τα οποία παρουσιάζουν μεγάλη συχνότητα εμφάνισης.

Διαισθητικά, θεωρούμε ενδιαφέροντα σύνολα όρων αυτά των οποίων οι όροι είναι ασυσχέτιστοι (uncorrelated) ή αρνητικά συσχετισμένοι (negative correlated). Για παράδειγμα, για ένα ζεύγος ασυσχέτιστων όρων *ab*, υπάρχει ένας μικρό πλήθος εγγράφων που το περιέχει. Με τη χρήση στατιστικών στοιχείων μεμονωμένα για τον κάθε όρο και τη συνάθροιση των αποτελεσμάτων που προκύπτουν δεν μπορούμε να ανακαλύψουμε εκδότες, οι οποίοι στο μέλλον θα δώσουν έγγραφα που περιέχουν το ζεύγος των όρων. Τελικά, το ζεύγος *ab* θεωρείται ενδιαφέρον αν οι πιθανότητες *P(A/B)* και *P(B/A)* είναι μικρότερες από κάποιο κατώφλι β.

Για την αξιοποίηση πιθανής συσχέτισης μεταξύ όρων, προτείνονται οι αλγόριθμοι USS και CSS.

Ο αλγόριθμος USS (Unique Synopses Storage) χρησιμοποιεί στατιστικά ανά όρο για να εκτιμήσει τη μελλοντική συμπεριφορά των πηγών πληροφορίας για σύνολα όρων. Πιο συγκεκριμένα, έστω $q = \{k_1, k_2, \ldots, k_n\}$ το συνεχές ερώτημα κάποιου συνδρομητή. Ο αλγόριθμος διαχωρίζει το q σε n ανεξάρτητους όρους. Υπολογίζει, βάσει των στατιστικών που υπάρχουν στους υπεύθυνους καταλόγους του συστήματος, τους εκδότες που εμφανίζονται σε όλα τα στατιστικά. Στη συνέχεια χρησιμοποιώντας τα συναθροιστικά αποτελέσματα που προκύπτουν και κάποιες τεχνικές πρόβλεψης διατάσσει τους εκδότες. Τελικά επιλέγει για το q τους υψηλότερα διατεταγμένους εκδότες. Η παραπάνω διαδικασία εκτελείται περιοδικά στο σύστημα λόγω της άφιξης νέων εγγράφων.

Λόγω των προβλημάτων που αντιμετωπίζει ο αλγόριθμος USS, όπως ο υψηλός δικτυακός φόρτος, η μειωμένη ακρίβεια και τα σφάλματα πρόβλεψης προτείνεται ο αλγόριθμος CSS. Ο CSS χρησιμοποιεί στατιστικά στοιχεία που διατηρούνται για επιλεγμένα σύνολα όρων. Έστω $S = \{k_1, k_2, \ldots, k_n\}$ ένα σύνολο όρων. Με τη χρήση κάποιας ντετερμινιστικής συνάρτησης επιλέγεται ένας κατάλογος $d(S)$, που είναι υπεύθυνος για το σύνολο S. Ο $d(S)$ επικοινωνεί με τους υπεύθυνους καταλόγους για κάθε όρο k_j του S και υπολογίζει την ένωση των συνόψεων όλων των εκδοτών που περιέχουν τον k_j. Τελικά, ο κατάλογος $d(S)$ υπολογίζει τον πληθάριθμο του συνόλου των εγγράφων που περιέχουν το S και στη συνέχεια τις υπό συνθήκη πιθανότητες για κάθε όρο k_j. Στόχος του αλγορίθμου είναι να ανακαλύψει ενδιαφέροντα σύνολα όρων, των οποίων οι όροι έχουν μικρές υπό συνθήκη πιθανότητες. Για τα σύνολα αυτά διατηρεί στατιστικά στοιχεία.

Η ενημέρωση στατιστικών στοιχείων για σύνολα όρων που έχουν χαρακτηριστεί ως ενδιαφέροντα δεν επιφέρει επιπλέον φόρτο στο δίκτυο. Αυτό συμβαίνει γιατί οποτεδήποτε ένας συνδρομητής ενημερώνει τα στατιστικά του υπεύθυνου καταλόγου d(S) για κάποιο όρο k_j, μπορεί να ενημερώσει τα στατιστικά στοιχεία για το σύνολο όρων S.

Όσον αφορά την καλύτερη αξιοποίηση των στατιστικών στοιχείων για πολλαπλούς όρους, μια ιδέα είναι το σύστημα να χρησιμοποιεί στατιστικά για σύνολα όρων που είναι υποσύνολα του συνεχούς ερωτήματος κάποιου χρήστη. Για την επιλογή των κατάλληλων υποσυνόλων, ο αλγόριθμος CSS χρησιμοποιεί τα maximal υποσύνολα ανάμεσα στα διαθέσιμα υποσύνολα πολλαπλών όρων.

Με βάση αυτά τα ενδιαφέροντα υποσύνολα, ο υπολογισμός της τιμής διάταξης (score) κάποιου εκδότη *p*, δίνεται από τον τύπο:

$$\text{score}_s(p) = \sum_{S_i \subseteq S} |S_i| \cdot \text{predScore}_{s_i}(p)$$

Το $\text{predScore}_{Si}(p)$ δηλώνει την πιθανότητα που έχει ο εκδότης *p* να δημοσιεύσει στο μέλλον ένα έγγραφο που περιέχει το σύνολο όρων S_i.

Για την πειραματική αξιολόγηση χρησιμοποιήθηκε ένα αναγνωρισμένο σύστημα μετρήσεων (benchmark), σχεδιασμένο για την αξιολόγηση κατανεμημένων συστημάτων ανάκτησης πληροφορίας. Το σύστημα μετρήσεων αποτελείται από περισσότερα από 800,000 έγγραφα από τη Wikipedia και από έναν αλγόριθμο που κατανέμει τα έγγραφα σε 1,000 εκδότες με ελεγχόμενη επικάλυψη. Η μετρική που χρησιμοποιήθηκε για την αξιολόγηση είναι η μέση ανάκληση, δηλαδή ο μέσος λόγος του συνολικού αριθμού των ενημερώσεων που λαμβάνονται από τους συνδρομητές προς το σύνολο των εγγράφων που ταιριάζουν με τις συνδρομές. Τα ερωτήματα που τέθηκαν στο σύστημα είναι δύο, τριών και τεσσάρων διαστάσεων.

Στα πειράματα έγινε σύγκριση μεταξύ του βασικού αλγορίθμου (baseline) που δε χρησιμοποιεί συνόψεις και τεχνικές πρόβλεψη συμπεριφοράς των εκδοτών με τον αλγόριθμο USS που χρησιμοποιεί Hash Sketches, τον USS με KMV Synopses και τον αλγόριθμο CSS. Ο CSS υπερτερεί των άλλων αλγορίθμων δίνοντας τα μεγαλύτερα κέρδη ανάκλησης. Ακόμη, ο USS με χρήση KMV Synopses είναι καλύτερος από τον USS που χρησιμοποιεί Hash Sketches, ιδιαίτερα για ερωτήματα με περισσότερους από δύο όρους. Τέλος, η χρήση συνόλων όρων που είναι ασυσχέτιστοι μεταξύ τους δίνουν το μεγαλύτερο κέρδος σε ανάκληση για τον αλγόριθμο CSS.

Αναφορές

[1] Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar and L. Trevisan. Counting Distinct Elements in a Data Stream. In RANDOM 2002.

[2] K. S. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis and R. Gemulla. On synopses for Distinct-Value Estimation Under Multiset Operations. In SIGMOD, 2007.

[3] I. Stoica, R. Morris, D. R. Karger, M. F. Kaashoek and H. Balakrishnan. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. In SIGCOMM, 2001.