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.2.2

Εξαγωγή δεδομένων, αντιστοίχιση τιμών και δομών, και ολοκλήρωση πληροφορίας

Σεπτέμβριος 2015

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Δράση 2	**Ολοκλήρωση παραδοσιακών και μη δεδομένων, πλοήγηση και αναζήτηση**				
Ομάδα	Ερ. Ομάδα 2	**Έναρξη**	01/06/2012	**Λήξη**	30/11/2015

Συντονιστής ΕΟ2	Τ. Σελλής (ΙΠΣΥ - ΕΚ «Αθηνά» & RMIT)	
Υποδράση: ΥΔ 2.2	Εξαγωγή δεδομένων, αντιστοίχιση τιμών και δομών, και ολοκλήρωση πληροφορίας	
Συμμετέχοντες	*Μέλη ΚΕΟ*	Τ. Σελλής (ΙΠΣΥ - ΕΚ «Αθηνά» & RMIT), D. Pfoser (ΙΠΣΥ - ΕΚ «Αθηνά»), Β. Βασσάλος (ΟΠΑ), Γ. Κούτρικα (Μετακαλούμενη - IBM Almaden), Θ. Δαλαμάγκας (ΙΠΣΥ - ΕΚ «Αθηνά»),
	Μέλη ΟΕΣ	Γ.Παπαδάκης (ΙΠΣΥ - ΕΚ «Αθηνά»), Κ. Μακρυνιώτη (ΟΠΑ), Γ. Παπαστεφανάτος (ΙΠΣΥ - ΕΚ «Αθηνά»), M. Reczko (Ε.ΚΕ.Β.Ε. Α. Φλέμινγκ)
Σύντομη Περιγραφή	Η Υποδράση ΥΔ2.2 στοχεύει στον προσδιορισμό ενός συνόλου από αλγορίθμους και τεχνικές για την εξαγωγή δεδομένων από ετερογενείς πηγές, στην εύρεση αντιστοιχήσεων μεταξύ τιμών και δομών δεδομένων μέσω εννοιολογικά πλούσιας μεταπληροφορίας και στην οργάνωση - ολοκλήρωσή τους σε ένα ενιαίο υπερχώρο δεδομένων.	
Παραδοτέο	Π.2.2 Εξαγωγή δεδομένων, αντιστοίχιση τιμών και δομών, και ολοκλήρωση πληροφορίας	
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 2 «Ολοκλήρωση παραδοσιακών και μη δεδομένων, πλοήγηση και αναζήτηση» παρέχει τεχνικές για τον ορισμό υπερχώρων δεδομένων και την αξιοποίηση παραδοσιακών και μη δεδομένων σε τέτοια περιβάλλοντα. Η Δράση οργανώνεται σε τρεις θεμελιώδεις δράσεις, εκ των οποίων η πρώτη αφορά στον ορισμό του εννοιολογικού μοντέλου αναπαράστασης υπερχώρων, η δεύτερη την περιγραφή του μηχανισμού ενσωμάτωσης νέων πηγών σε έναν υπερχώρο και την έυρεση αντιστοιχήσεων μεταξύ ετερογενών πηγών δεδομένων και η τρίτη την αρχιτεκτονική και τους μηχανισμούς που θα πρέπει να διαθέτει ένα σύστημα υποστήριξης υπερχώρων δεδομένων για την εφαρμογή επερωτήσεων και την ανάκτηση πληροφορίας από αυτό.

Το παρόν Παραδοτέο Π.2.2 περιλαμβάνει τα αποτελέσματα της υποδράσης ΥΔ2.2. Στην ενότητα 1 παρουσιάζουμε το γενικότερο πλαίσιο του προβλήματος. Στις ενότητες 2 και 3 περιγράφουμε τεχνικές για την ταυτοποίηση οντοτήτων, που στόχο έχει να εντοπίσει και να ενοποιήσει διαφορετικές περιγραφές οντοτήτων που περιγράφουν το ίδιο αντικείμενο στον πραγματικό κόσμο. Στην ενότητα 4 περιγράφουμε τη διαδικασία της ανάλυσης συναισθήματος στο κοινωνικό δίκτυο Twitter[1].

1.1 Πλαίσιο έρευνας

Ο σκοπός της Υποδράσης 2.2 είναι να παρέχει τεχνικές για την αναζήτηση και εξαγωγή δεδομένων από ετερογενείς πηγές, την εύρεση αντιστοιχήσεων μεταξύ τιμών και δομών δεδομένων και την οργάνωση - ολοκλήρωσή τους σε ένα ενιαίο

[1] https://twitter.com/

υπερχώρο δεδομένων. Κομβικό στοιχείο της προσέγγισης μας είναι η σταδιακή δημιουργία αντιστοιχήσεων μεταξύ των μοντέλων δεδομένων όσο ο χρήστης προχωράει σε μεγαλύτερα επίπεδα ολοκλήρωσης σε αντίθεση με τις παραδοσιακές μεθόδους ολοκλήρωσης δεδομένων, που προϋποθέτουν την ύπαρξη ενός πλήρους συνόλου σημασιολογικών και δομικών αντιστοιχήσεων.

Ένα θεμελιώδες πρόβλημα στην αναζήτηση και επεξεργασία ετερογενούς πληροφορίας σχετίζεται με την ικανότητα των συστημάτων να προσδιορίζουν δομικές και σημασιολογικές συσχετίσεις δεδομένων ή/και μεταδεδομένων που προέρχονται από διαφορετικές πηγές, με απώτερο στόχο την ολοκλήρωσή τους. Η δυνατότητα των χρηστών να πραγματοποιούν αναζητήσεις με ενιαίο και ομοιόμορφο τρόπο προϋποθέτει (α) την αναζήτηση και εξαγωγή σχετικών με τα κριτήρια αναζήτησης δεδομένων μέσα από μια πληθώρα διαφορετικών πηγών, (β) τον προσδιορισμό των ακριβών συσχετίσεων μεταξύ δεδομένων και αντικειμένων προερχόμενων από ετερογενείς πηγές, μια διαδικασία γνωστή ως δημιουργία αντιστοιχήσεων (mapping generation) και τέλος (γ) την εναρμόνιση και ολοκλήρωση των δεδομένων σε ένα σημασιολογικά χρήσιμο αποτέλεσμα για τους χρήστες. Η εξεύρεση, λοιπόν, των σωστών αντιστοιχήσεων είναι κομβικής σημασίας για όλη τη διαδικασία ολοκλήρωσης και δυστυχώς ο προσδιορισμός τους είναι μια αρκετά χρονοβόρα διαδικασία και με πολλές πιθανότητες δημιουργίας λανθασμένων αντιστοιχήσεων, καθώς η πρωταρχική πληροφορία είναι ετερογενής, έχει σχεδιαστεί και παραχθεί ανεξάρτητα μέσα από διαφορετικές εφαρμογές, είναι αποθηκευμένη σε διαφορετικές μορφές και εμπεριέχει αρκετή αβεβαιότητα ως προς τη σημασιολογία της. Στην συγκεκριμένη Υποδράση θα προσεγγίσουμε το πρόβλημα της αντιστοίχισης τιμών και δομών και ολοκλήρωσης πληροφορίας σε υπερχώρους δεδομένων μέσω σταδιακής δημιουργίας αντιστοιχήσεων μεταξύ των μοντέλων δεδομένων. Επιπλέον, θα μελετήσουμε τεχνικές δημιουργίας αντιστοιχήσεων μεταξύ μοντέλων δεδομένων σε πραγματικό χρόνο, οι οποίες θα εμπλουτίζονται σημασιολογικά και θα εξειδικεύονται με τη συμμετοχή του χρήστη.

1.2 Κίνητρα της έρευνας και κεντρική ιδέα

Στο πλαίσιο της Υποδράσης 2.2 σχεδιάσαμε και υλοποιήσαμε αλγορίθμους που επιτρέπουν την εξαγωγή δεδομένων από ετερογενείς πηγές, την εύρεση

αντιστοιχήσεων μεταξύ τιμών και δομών δεδομένων και την οργάνωση - ολοκλήρωσή τους σε ένα ενιαίο υπερχώρο δεδομένων.

Τα παραδοσιακά συστήματα ανταλλαγής και ολοκλήρωσης δεδομένων στοχεύουν να παρέχουν υπηρεσίες συμβατές με τα περισσότερα συστήματα βάσεων δεδομένων. Θεωρούν ένα συγκεκριμένο σύνολο πηγών δεδομένων για τις οποίες είναι γνωστά εκ των προτέρων τόσο το σχήμα όσο και η σημασιολογία των δεδομένων που περιέχονται σε αυτές. Οι δομικές και σημασιολογικές αντιστοιχήσεις μεταξύ των δεδομένων δημιουργούνται σε μια αρχική στιγμή με βάση τα μεταδεδομένα των πηγών και στη συνέχεια πραγματοποιείται η εναρμόνιση και ολοκλήρωση των δεδομένων με βάση αυτές τις αντιστοιχήσεις. Αυτό επιβαρύνει χρονικά τόσο την αρχική εγκατάσταση ενός συστήματος ανταλλαγής και ολοκλήρωσης δεδομένων όσο και την ενσωμάτωση μιας νέας πηγής δεδομένων σε αυτό.

Δυστυχώς η πρότερη γνώση της δομικής και σημασιολογικής πληροφορίας σε περιβάλλοντα υπερχώρων δεδομένων δεν είναι πάντα εφικτή. Η δομική και σημασιολογική πληροφορία που προέρχεται από πηγές δεδομένων που απαρτίζουν τους υπερχώρους μπορεί να μην είναι διαθέσιμη παρά μόνο κατά τη φάση της αναζήτησης (π.χ., δεδομένα αισθητήρων, αρχεία κειμένου), είτε να εξελίσσεται ταχύτατα (π.χ., web logs) εμποδίζοντας την υιοθέτηση σταθερών αντιστοιχήσεων μεταξύ τους. Οι υπερχώροι δεδομένων, λοιπόν, θα πρέπει να αντιμετωπιστούν περισσότερο ως χώροι συνύπαρξης ετερογενούς πληροφορίας η οποία είναι διαθέσιμη προς επεξεργασία και ολοκλήρωση και λιγότερο ως συστήματα ολοκλήρωσης δεδομένων. Σε αυτή τη βάση, προτείνεται η υιοθέτηση τεχνικών δυναμικής δημιουργίας αντιστοιχήσεων μεταξύ των μοντέλων δεδομένων που απαρτίζουν τον υπερχώρο δεδομένων και σταδιακού εμπλουτισμού τους με βάση το είδος των πηγών και τα επίπεδα ολοκλήρωσης που ο χρήστης επιθυμεί. Για παράδειγμα, η αναζήτηση μέσω λέξεων κλειδιών χωρίς καμιά παραπάνω επεξεργασία των αποτελεσμάτων μπορεί να παρέχεται στην πλειοψηφία των πηγών ως βασική λειτουργία εξαγωγής, αντιστοίχησης και ολοκλήρωσης δεδομένων. Τα δεδομένα που είναι αποθηκευμένα σε διάφορες μορφές θα πρέπει να περιέχουν μια λέξη κλειδί χωρίς να απαιτείται καμιά επιπλέον πληροφορία για τη σημασιολογία ή τη δομή τους. Στην περίπτωση των πιο σύνθετων ερωτήσεων, επιπλέον σημασιολογική και δομική

πληροφορία θα απαιτείται, η οποία θα επιβαρύνει το χρήστη με το κόστος να την περιγράψει και να την παρέχει. Θα πρέπει να σημειωθεί ότι η επιβάρυνση του χρήστη δεν προϋποθέτει ότι θα πρέπει να γνωρίζει απαραίτητα το πλήθος, το είδος ή τη δομή των δεδομένων που αναζητάει, αλλά ότι θα διαμορφώνει σταδιακά τις σημασιολογικές αντιστοιχήσεις και θα φιλτράρει το πλήθος των αποτελεσμάτων μέσα από μια διαδικασία πλοήγησης σε αυτά.

Επίσης, η πρωτοτυπία της Υποδράσης βρίσκεται στη δημιουργία αντιστοιχήσεων για μη παραδοσιακά δεδομένα, όπως είναι οι ροές δεδομένων, (π.χ. tweets), με βάση μεταπληροφορία που θα παρέχεται σε πραγματικό χρόνο. Σε αυτές τις περιπτώσεις, οι διαδικασίες ταιριάσματος και αντιστοίχησης γίνονται σε πραγματικό χρόνο, χωρίς να υπάρχει η δυνατότητα επανεπεξεργασίας των δεδομένων ή αποθήκευσής τους για εξαγωγή αντιστοιχήσεων εκ των υστέρων.

2 Μετα-ομαδοποίηση: Οδηγώντας την ταυτοποίηση οντοτήτων στο επόμενο επίπεδο

Η παρούσα δημοσίευση ασχολείται επίσης με το θέμα της Ταυτοποίησης Οντοτήτων (Entity Resolution), που αποτελεί θεμελιώδη διαδικασία για την ενοποίηση αλληλεπικαλυπτόμενων πηγών πληροφορίας. Όπως ήδη αναφέρθηκε, η διαδικασία αυτή έχει τετραγωνική πολυπλοκότητα, αφού κάθε οντότητα πρέπει να συγκριθεί με όλες τις άλλες. Για να βελτιωθεί η αποδοτικότητά της (efficiency), συνήθως χρησιμοποιούνται τεχνικές blocking που περιορίζουν τις συγκρίσεις μεταξύ όμοιων αντικειμένων. Στα πλαίσια θορυβωδών ετερογενών δεδομένων (noisy, heterogeneous data), οι τεχνικές αυτές χρησιμοποιούν πλεονασμό (redundancy) για να επιτύχουν υψηλό recall. Τοποθετούν, δηλαδή, κάθε οντότητα σε πολλά blocks. Όμως, αυτό έχει σαν αποτέλεσμα να προκύπτουν πολλές επαναλαμβανόμενες συγκρίσεις καθώς επίσης και περιττές (superfluous) συγκρίσεις μεταξύ αταίριαστων (non-matching) οντοτήτων. Στόχος της παρούσας δημοσίευσης είναι να προτείνει το meta-blocking σαν μια καινούρια προσέγγιση για την βελτίωση της ακρίβειας (precision) των τεχνικών blocking, χωρίς να επιφέρει σημαντική μείωση στο recall. Για το σκοπό αυτό, μετατρέπει ένα σύνολο από blocks σε γράφο, οι

κόμβοι του οποίου αντιστοιχούν σε οντότητες, ενώ οι ακμές του συνδέουν τις οντότητες που συγκρίνονται τουλάχιστον μια φορά στα blocks. Κατά τη δημιουργία του γράφου, όλες οι επαναλαμβανόμενες συγκρίσεις εξαλείφονται, αφού δεν δημιουργούνται παράλληλες ακμές. Στη συνέχεια, κάθε ακμή παίρνει ένα βάρος που είναι ανάλογο με την πιθανότητα που έχουν οι προσκείμενες οντότητες να ταιριάζουν (δηλαδή, να αντιστοιχούν στο ίδιο αντικείμενο του πραγματικού κόσμου). Έτσι, οι ακμές με χαμηλό βάρος μπορούν να περικοπούν ώστε να διαγραφεί ένα μεγάλο μέρος των περιττών συγκρίσεων. Στο τέλος, δημιουργείται ένα καινούριο block για κάθε ακμή που δεν περικόπηκε και ο κλαδεμένος γράφος μετατρέπεται σε ένα καινούριο σύνολο από blocks. Η καλή απόδοση της συγκεκριμένης προσέγγισης αποδεικνύεται με εκτενή πειράματα πάνω σε μεγάλα σύνολα θορυβωδών δεδομένων.

Τα αποτελέσματά μας δημοσιεύθηκαν στο περιοδικό IEEE Transactions on Knowledge and Data Engineering [PKPN14].

3 Μετα-ομαδοποίηση με επίβλεψη

Στόχος της Ταυτοποίησης Οντοτήτων (Entity Resolution) είναι να εντοπίσει τις διαφορετικές περιγραφές οντοτήτων που ουσιαστικά περιγράφουν το ίδιο αντικείμενο στον πραγματικό κόσμο. Αποτελεί μια κοστοβόρα διαδικασία τετραγωνικής πολυπλοκότητας, η οποία συνήθως εφαρμόζεται σε μεγάλα δεδομένα μέσω τεχνικών blocking. Αυτές οι μέθοδοι ομαδοποιούν παρόμοιες οντότητες σε blocks και εκτελούν συγκρίσεις μόνο μέσα στα blocks. Το πρόβλημα αυτής της προσέγγισης είναι ότι περιλαμβάνει μεγάλο αριθμό περιττών συγκρίσεων. Η ακρίβεια της όμως μπορεί να βελτιωθεί με τη βοήθεια του meta-blocking. Η τεχνική αυτή αναδιαμορφώνει ένα σύνολο από blocks, αφαιρώντας ένα μεγάλο μέρος των περιττών συγκρίσεων. Οι υπάρχουσες μέθοδοι για meta-blocking είναι unsupervised με αποτέλεσμα να χρησιμοποιούν απλούς κανόνες περιορισμένης ακρίβειας για την περικοπή των αχρείαστων συγκρίσεων. Στην παρούσα δημοσίευση, προτείνουμε supervised τεχνικές για meta-blocking, οι οποίες μαθαίνουν σύνθετους κανόνες για τον εντοπισμό των περιττών συγκρίσεων με τη βοήθεια μηχανικής μάθησης. Για τη μοντελοποίηση του προβλήματος, χρησιμοποιούμε ένα περιορισμένο αριθμό χαρακτηριστικών (features) που εξάγονται με μικρό κόστος και παίρνουν τιμές υψηλής

διακριτικότητας (distinctiveness). Τα πειράματα μας αποδεικνύουν ότι η συγκεκριμένη προσέγγιση μπορεί να επιτύχει υψηλή ακρίβεια ανεξαρτήτως του αλγορίθμου μάθησης, ακόμα και όταν χρησιμοποιούμε ένα μικρό σύνολο δεδομένων για την εκμάθηση των σύνθετων κανόνων. Επίσης, αποδεικνύουν ότι η προσέγγισή μας επιτυγχάνει καλύτερη απόδοση από τις υπάρχουσες unsupervised τεχνικές.

Τα αποτελέσματά μας δημοσιεύθηκαν στο περιοδικό Proceedings of the VLDB Endowment [PaPK14]. Επίσης, τα αποτελέσματα παρουσιάσθηκαν με ομιλία στο 13ο Ελληνικό Συμπόσιο Διαχείρισης Δεδομένων [PaPK15].

4 Εξαγωγή Συναισθήματος από Tweets: Προκλήσεις που προκύπτουν από την Πολυγλωσσία

Καθημερινά χρήστες των κοινωνικών δικτύων και των υπηρεσιών microblogging μοιράζονται τις απόψεις τους για προϊόντα, εταιρείες, ταινίες και γενικότερα τα συναισθήματα τους για ποικιλία θεμάτων. Καθώς οι πλατφόρμες κοινωνικής δικτύωσης και microblogging γίνονται όλο και περισσότερο δημοφιλείς, η ανάγκη να εξάγουμε και να αναλύσουμε το περιεχόμενο τους μεγαλώνει. Στην παρούσα δημοσίευση μελετάμε τη διαδικασία της ανάλυσης συναισθήματος (Sentiment Analysis) στο γνωστό κοινωνικό δίκτυο Twitter, της οποίας στόχος είναι να αναγνωρίσει αν ένα κείμενο εκφράζει θετική ή αρνητική άποψη, ή αν είναι αντικειμενικό και περιγράφει ένα γεγονός. Παρουσιάζουμε μια μελέτη εστιασμένη σε ελληνικά tweets και προτείνουμε μια αποτελεσματική μέθοδο η οποία τα κατηγοριοποιεί σε θετικά, αρνητικά και ουδέτερα σύμφωνα με το συναίσθημα που εκφράζουν. Επιβεβαιώνουμε την αποτελεσματικότητα της μεθόδου και σε ελληνικά και σε αγγλικά δεδομένα με σκοπό να ελέγξουμε την ευρωστία της σε πολυγλωσσικές προκλήσεις, και παρουσιάζουμε την πρώτη πολυγλωσσική συγκριτική μελέτη με τρεις προϋπάρχουσες σύγχρονες τεχνικές για εξαγωγή συναίσθηματος σε αγγλικά tweets. Τέλος, εξετάζουμε τη σημασία και την επιρροή διάφορων τεχνικών προεπεξεργασίας σε διαφορετικές γλώσσες. Η μέθοδος μας έχει καλύτερη επίδοση σε σχέση με δύο από τις τρεις μεθόδους με τις οποίες συγκρίναμε και βρίσκεται στο ίδιο επίπεδο με την

καλύτερη από αυτές, με τη διαφορά ότι απαιτεί αισθητά λιγότερο χρόνο για πρόβλεψη και εκπαίδευση.

Τα αποτελέσματά μας δημοσιεύθηκαν στο άρθρο [MaVa15] που παρουσιάστηκε στο 17th International Conference on Big Data Analytics and Knowledge Discovery (DaWak 2015).

5 Ανακεφαλαίωση

Το παρόν παραδοτέο Π2.2 παρουσιάζει τα αποτελέσματα της υποδράσης ΥΔ2.2 του έργου ΕΙΚΟΣ. Ο στόχος της υποδράσης ΥΔ2.2 ήταν να σχεδιάσουμε και να υλοποιήσουμε αλγορίθμους που επιτρέπουν την εξαγωγή δεδομένων από ετερογενείς πηγές, την εύρεση αντιστοιχήσεων μεταξύ τιμών και δομών δεδομένων και την οργάνωση - ολοκλήρωσή τους σε ένα ενιαίο υπερχώρο δεδομένων.

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

1. Προτείναμε μεθόδους για την εύρεση αντιστοιχήσεων μεταξύ ετερογενών πηγών δεδομένων. Συγκεκριμένα, εστιάσαμε στη βελτίωση της αποδοτικότητας της Ταυτοποίησης Οντοτήτων αναβαθμίζοντας τη λειτουργία των τεχνικών blocking. Τον σκοπό αυτό εξυπηραιτεί αποτελεσματικά το meta-blocking, κάνοντας πλέον εφικτή την εφαρμογή της Ταυτοποίησης Οντοτήτων σε δεδομένα μεγάλης κλίμακας και υψηλού θορύβου χωρίς να έχουμε σημαντικές απώλειες ούτε στο recall ούτε στο precision.
2. Επιπλέον, βελτιώσαμε την απόδοση των τεχνικών meta-blocking έτσι ώστε να εντοπίζονται και να διαγράφονται περιττές συγκρίσεις με μεγαλύτερη ακρίβεια. Αυτό έχει σαν αποτέλεσμα να βελτιώνεται η λειτουργία των τεχνικών blocking σε ακόμα μεγαλύτερο βαθμό και έτσι να αυξάνεται η κλιμακωσιμότητα (scalability) της Ταυτοποίησης Οντοτήτων, μειώνοντας ακόμα περισσότερο τον χρόνο που χρειάζεται η Ταυτοποίηση Οντοτήτων σε δεδομένα μεγάλης κλίμακας και υψηλού θορύβου που προέρχονται από το Web.
3. Συμπληρωματικά, προτείναμε μεθόδους για την εξαγωγή σύντομων κειμένων, όπως είναι τα tweets που αποτελούν μη δομημένο τύπο

δεδομένων, και τον εμπλουτισμό τους με την πληροφορία του συναισθήματος που εκφράζουν. Για τη διαδικασία αυτή χρησιμοποιούνται μέθοδοι επεξεργασίας φυσικής γλώσσας και αλγόριθμοι μηχανικής μάθησης. Ιδιαίτερα στην περίπτωση των προσωπικών υπερχώρων δεδομένων, το συναίσθημα των δεδομένων κειμένου ενός χρήστη μπορεί να αποτελέσει χρήσιμο μεταδεδομένο στη βελτίωση της εξατομίκευσης περιεχόμενου και να συμβάλει στην εξέλιξη των προσωποκεντρικών συστημάτων.

Δημοσιεύσεις

[PKPN14] George Papadakis, Georgia Koutrika, Themis Palpanas, Wolfgang Nejl. Meta-Blocking: Taking Entity Resolution to the Next Level. IEEE Transactions on Knowledge and Data Engineering, Volume 26, Number 8, p. 1946-1960, August 2014.

[PaPK14] George Papadakis, George Papastefanatos, Georgia Koutrika. Supervised Meta-blocking. Proceedings of the VLDB Endowment, Volume 7, Number 14, p. 1929-1940, October 2014.

[PaPK15] George Papadakis, George Papastefanatos, Georgia Koutrika. Supervised Meta-blocking. Παρουσιάσθηκε στο 13ο Ελληνικό Συμπόσιο Διαχείρισης Δεδομένων (ΕΣΔΔ 2015), Αθήνα, 30-31 Ιουλίου, 2015.

[MaVa15] Nantia Makrynioti, Vasilis Vassalos. Sentiment Extraction from Tweets: Multilingual Challenges. In Proceedings 17th International Conference on Big Data Analytics and Knowledge Discovery (DaWak 2015), Valencia, Spain, September 1-4, 2015.

Παράρτημα