ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Δημιουργία Περιλήψεων Διαδικτυακών Συζητήσεων μέσω Εξαγωγής Θεμάτων Συζήτησης και Εκτίμησης Διαφωνιών

Θοδωρής Ε. Γεωργίου
Εμμανουήλ Ι. Καρβούνης

Επιβλέποντες: **Ιωάννης Ιωαννίδης,** Καθηγητής

ΑΘΗΝΑ

ΙΑΝΟΥΑΡΙΟΣ 2011

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Δημιουργία Περιλήψεων Διαδικτυακών Συζητήσεων μέσω Εξαγωγής Θεμάτων Συζήτησης και Εκτίμησης Διαφωνιών

Θοδωρής Ε. Γεωργίου

Α.Μ.: 1115200500181

Εμμανουήλ Ι. Καρβούνης

Α.Μ.: 1115200500059

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Ιωάννης Ιωαννίδης,** Καθηγητής

ΠΕΡΙΛΗΨΗ

Σε αυτή την εργασία, προτείνουμε μια νέα μέθοδο για την εξόρυξη απόψεων και τη δημιουργία περιλήψεων σε συζητήσεις διαδικτυακών φόρουμ (Web Forums / Web Discussion Boards). Πιο συγκεκριμένα, δημιουργούμε περιλήψεις των συζητήσεων εξάγοντας ένα μικρό ποσοστό των αναρτήσεων, έτσι ώστε να επιτυγχάνεται η μέγιστη κάλυψη του θέματος και να παρουσιάζονται όλες οι διαφορετικές απόψεις των συμμετεχόντων. Για να πετύχουμε τα παραπάνω, εργαζόμαστε σε δύο βασικούς άξονες. Πρώτον, εντοπίζουμε όλα τα επιμέρους θέματα της συζήτησης και τη χωρίζουμε σε υπο-συζητήσεις. Αυτό μας δίνει τη δυνατότητα να ομαδοποιούμε τις αναρτήσεις σε ομάδες που καλύπτουν όλα τα θέματα που συζητήθηκαν. Δεύτερον, δημιουργούμε ομάδες από χρήστες που συμφωνούν μεταξύ τους και προσδιορίζουμε τις πιθανές διαφωνίες μεταξύ των ομάδων αυτών. Αυτό μας δίνει τη δυνατότητα να δημιουργούμε περιλήψεις που μπορούν να χρησιμοποιηθούν για τη γρήγορη και αποτελεσματική αναγνώριση ομάδων χρηστών, των απόψεών τους, των επιχειρημάτων τους και φυσικά των σημείων τριβής μεταξύ τους. Όλες οι περιλήψεις περιλαμβάνουν μετα-πληροφορίες που μπορούν να χρησιμοποιηθούν σε αναζήτηση, με χρήση παραδοσιακών αλγορίθμων αντιστοίχισης λέξεων-κλειδιών. Αρχικά πειράματα με τυχαία επιλεγμένα διαδικτυακά φόρουμ και με ανθρώπινη αξιολόγηση έδωσαν πολύ ενθαρρυντικά αποτελέσματα και δείχνουν τις μεγάλες δυνατότητες της συνολικής προσέγγισης και των αλγορίθμων μας.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Εξόρυξη Απόψεων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: περιλήψεις απόψεων, χώροι διαδικτυακών συζητήσεων, εξόρυξη θεμάτων συζήτησης, εκτίμηση διαφωνιών, κοινωνικά δίκτυα

ABSTRACT

In this thesis, we propose a novel method for opinion mining and summarization of online web forum discussions that contain debates. We create summaries of the discussions by extracting a small percentage of the posts, aiming at maximizing topic coverage and bringing out the different viewpoints of the participants. To achieve this, we proceed in two main steps. First, we identify all the sub-topics of the discussion and divide it in sub-discussions. This enables us to cluster posts in groups that cover all the topics discussed. Second, we create groups of agreeing users and identify disagreements between them. This enables us to create summaries that can be used to quickly and efficiently identify the different groups of users, their opinions, their arguments and the point of friction between them. All summaries include metadata information that can be used for searching by traditional keyword matching algorithms. Initial experiments with randomly chosen web forum discussions and human evaluators have given very encouraging results and indicate the great potential of the overall approach and the specific algorithms.

SUBJECT AREA: Opinion Mining

KEYWORDS: opinion summarization, web forums, topic extraction, disagreement estimation

Αφιερώνουμε την παρούσα πτυχιακή εργασία στους γονείς μας για τη συνεχή στήριξή τους στο δύσκολο δρόμο που αποφασίσαμε να ακολουθήσουμε

Θοδωρής, Μάνος

ΕΥΧΑΡΙΣΤΙΕΣ

Για τη διεκπεραίωση της παρούσας Πτυχιακής Εργασίας, θα θέλαμε να ευχαριστήσουμε πρωτίστως τον επιβλέποντά μας καθ. Γιάννη Ιωαννίδη για την πολύτιμη συμβολή του στην ολοκλήρωσή της και την άψογη συνεργασία μας. Με την εμπειρία του, αλλά και με τη γενικότερη στάση του ως επιστήμονας και ως άνθρωπος, κατάφερε να μας διδάξει πολλά περισσότερα απ' όσα μπορούν να αποτυπωθούν με αυτή την πτυχιακή. Σίγουρα χρωστάμε μεγάλο μέρος της τωρινής και μελλοντικής επιτυχίας μας σε αυτόν. Σας ευχαριστούμε!

Προσωπικά, θέλω ακόμα να ευχαριστήσω τον συμφοιτητή μου Μάνο για την άψογη συνεργασία του καθ' όλη τη διάρκεια της εργασίας αυτής, από το ξεκίνημα της σαν μια περίληψη για poster στο συνέδριο του SIGMOD μέχρι και την υποβολή της σαν πλήρες επιστημονικό άρθρο. Θέλω επίσης να ευχαριστήσω τη Μένη που ήταν πάντα κοντά μου και συμμεριζόταν υπομονετικά τους προβληματισμούς μου και με εμψύχωνε. Το Μανόλη που το μεράκι του, οι συμβουλές του και η υποδειγματική σχετική δουλειά του αποτέλεσε βάση και στόχο σε αυτά που προσπαθώ να πετύχω. Τέλος ένα μεγάλο ευχαριστώ σε όλους τους φίλους μου στο Τμήμα Πληροφορικής του Πανεπιστημίου Αθηνών που βοηθήσανε στο να περάσει αυτός ο χρόνος ευχάριστα ώστε να κοιτάω πίσω και να βλέπω ωραίες στιγμές.

Θοδωρής

Καταρχάς, θα ήθελα να ευχαριστήσω το φίλο μου και συνάδελφο Θοδωρή γιατί αποδείχθηκε άψογος συνεργάτης στην δύσκολη πορεία που χρειάστηκε να ακολουθήσουμε μέχρι η εργασία μας να φτάσει στην τωρινή της μορφή. Οι τεράστιες δυνατότητές του και η προσωπικότητά του ήταν το τέλειο συμπλήρωμα για την ομάδα μας. Επίσης, θα ήθελα να ευχαριστήσω τον επιστήθιο φίλο μου Παναγιώτη για την πολύτιμη και καθοριστική συμβολή του σε πολλά στάδια της εργασίας, από τη συζήτηση των αλγορίθμων μέχρι τη βοήθεια κατά την περίοδο της αξιολόγησης του τελικού συστήματος. Αλλά κυρίως γιατί, από τα παιδικά μας χρόνια μέχρι και σήμερα, δεν σταμάτησε ποτέ να νοιάζεται για μένα και να με βοηθάει, με την ιδιαίτερη οξυδέρκειά του, σε ό,τι χρειαζόμουν. Επιπλέον, θα ήθελα να ευχαριστήσω ιδιαίτερα την Κωνσταντίνα, που αποτελεί, απλά και μόνο με την παρουσία της, πηγή έμπνευσης, στήριξης και ηρεμίας στη ζωή μου. Της οφείλω πολύ περισσότερα απ' όσα μπορούν να ειπωθούν σε μερικές γραμμές. Τέλος, θα ήθελα να ευχαριστήσω ολόψυχα τους γονείς μου, Γιάννη και Ματίνα, που με την συνεχή αγάπη και προστασία τους μου επέτρεψαν να αναπτύξω απρόσκοπτα το χαρακτήρα μου και να αναζητήσω τη γνώση σε ένα περιβάλλον στήριξης και κατανόησης. Ό,τι είμαι σήμερα και ό,τι γίνω στο μέλλον θα το οφείλω πάντα σε αυτούς.

Μάνος

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η εργασία αυτή πρακτικά ξεκίνησε το Φεβρουάριο του 2010 με την υποβολή μιας περίληψης των αρχικών μας ιδεών στον προπτυχιακό διαγωνισμό του συνεδρίου ACM Sigmod (Special Interest Group on Management Of Data). Η περίληψή αυτή, που αναφερόταν σε ένα κομμάτι της παρούσας πτυχιακής εργασίας το οποίο όμως πλέον έχει εξελιχθεί αρκετά, έγινε δεκτή, κάτι που αποτελούσε κι από μόνο του μεγάλη επιτυχία, μιας και ήταν μόλις η δεύτερη φορά από καταβολής του διαγωνισμού που μια εργασία από φοιτητές ελληνικών Πανεπιστημίων γινόταν δεκτή. Ως αποτέλεσμα, κληθήκαμε να ταξιδέψαμε το καλοκαίρι στις Ηνωμένες Πολιτείες, στην πόλη Indianapolis, για να παρουσιάσουμε τη δουλεία μας στα πλαίσια του συνεδρίου SIGMOD 2010 (http://www.sigmod2010.org/ugposter_list.shtml).

Δεδομένου ότι η δουλειά μας ήταν σε πολύ πρώιμο στάδιο τη στιγμή που γίναμε δεκτοί, ακολούθησαν μερικοί μήνες πολύ σκληρής εργασίας ώστε να έχουμε στα χέρια μας κάτι το οποίο να είναι παρουσιάσιμο και επιστημονικά τεκμηριωμένο. Τελικά, τον Ιούνιο ήταν η στιγμή του μεγάλου ταξιδιού! Η εμπειρία που αποκομίσαμε από τη συμμετοχή μας σε ένα τόσο κορυφαίο συνέδριο καθώς και τα πολύτιμα σχόλια που λάβαμε πάνω στην εργασίας μας από διακεκριμένους επιστήμονες, σίγουρα μας έδωσαν μια τεράστια ώθηση για να φτάσουμε στο αποτέλεσμα που παρουσιάζουμε στην παρούσα εργασία. Επίσης, στο τέλος του συνεδρίου, είχε προγραμματιστεί και μια βράβευση για την καλύτερη εργασία του διαγωνισμού. Η χαρά μας ήταν απερίγραπτη όταν, μετά από ένα αρχικό σοκ, συνειδητοποιήσαμε ότι είχαμε βραβευτεί ως η καλύτερη προπτυχιακή παρουσίαση του συνεδρίου (http://www.sigmod.org/sigmod-awards/sigmod-awards#undergraduate)!

Παράλληλα με τα προηγούμενα, εργαστήκαμε στην συγγραφή της εργασίας και σε μορφή επιστημονικού άρθρου με σκοπό να την υποβάλουμε στο συνέδριο Eureka 2010 (http://eureka.hpclab.ceid.upatras.gr/eureka2010). Η εργασίας μας έγινε δεκτή και βραβεύτηκε με το “Best Paper Award” του συνεδρίου!

Τέλος, αξίζει να σημειωθεί ότι το περιεχόμενο της παρούσας εργασίας αποτελεί την πιο πρόσφατη και ολοκληρωμένη έκδοση της δουλειάς μας, όπως την γράψαμε για να την υποβάλουμε ως επιστημονικό άρθρο στο επόμενο συνέδριο του SIGMOD (2011) που θα γίνει στην Αθήνα.

1. ΕΙΣΑΓΩΓΗ

Ο Παγκόσμιος Ιστός έχει εξελιχθεί σε μια πολύτιμη, αλλά ιδιαίτερα αδόμητη και δύσκολη στην πλοήγηση, πηγή πληροφοριών. Από τη σημερινή της μορφή σαν μια «εγκυκλοπαίδεια» πληροφοριών, όλο και περισσότερο μετατρέπεται σε μια συνεχώς αυξανόμενη πηγή απόψεων των χρηστών που συμβάλουν στο περιεχόμενό της. Και μόνο από τον αριθμό του όγκου τους, οι εν λόγω απόψεις μπορεί να έχουν σπουδαίο ρόλο σε κάθε διαδικασία λήψης αποφάσεων, όμως δεν είναι εύκολα προσβάσιμες ή κατάλληλα παρουσιασμένες προς τον ενδιαφερόμενο χρήστη για την αποτελεσματική του πλοήγηση.

Η Εξόρυξη Απόψεων είναι ένας τομέας που ενδιαφέρεται ιδιαίτερα στον εντοπισμό, την εξαγωγή, τον συνοψισμό, και τον συμπερασμό των απόψεων που εμφανίζονται με αδόμητη μορφή σε έγγραφα του ιστού. Η μέχρι τώρα έρευνα έχει κυρίως επικεντρωθεί σε μία μόνο μορφή των γνωμών αυτών, αυτών που έχουν να κάνουν με κριτικές και σχολιασμούς ταινιών, προϊόντων, κ.λπ. Άλλες μορφές όμως γνωμών και απόψεων που μπορούν να βρεθούν στο Διαδίκτυο είναι εξίσου σημαντικές, όπως απόψεις για τρέχοντα γεγονότα, για θέματα της επικαιρότητας, καθώς και άλλα ζητήματα που είναι ανοικτά για συζήτηση, για τα οποία έχει δοθεί ελάχιστη προσοχή από την ερευνητική κοινότητα.

Η εν λόγω πληροφορία μπορεί να βρεθεί σε πολλά σημεία του διαδικτύου, αλλά το σημαντικότερο και αυτό που συγκεκριμένα δημιουργήθηκε για να εξυπηρετήσει το σκοπό αυτό (παράθεση και εναλλαγή απόψεων), είναι τα διαδικτυακά φόρουμ συζητήσεων γνωστά και ως διαδικτυακοί πίνακες συζητήσεων. Σε αυτούς τους ιστοχώρους μακρές και ποικίλες συνομιλίες λαμβάνουν χώρα, και ενδιαφέρουσες γνώμες για πληθώρα θεμάτων συνεχώς ανταλλάσσονται μεταξύ των χρηστών. Ένα βασικό χαρακτηριστικό των διαδικτυακών συζητήσεων είναι ότι η ανταλλαγή απόψεων γίνεται με μια «χρήστη-προς-χρήστη συνολική ανωνυμία», πράγμα που σημαίνει ότι όλοι οι χρήστες έχουν ισότιμο ρόλο στις συζητήσεις (σε αντίθεση με τα blogs, για παράδειγμα, όπου ο συγγραφές του blog ορίζει τη συζήτηση) και μπορούν να αναρτήσουν τη γνώμη τους χωρίς να αποκαλύψουν την πραγματική τους ταυτότητα. Αυτό ενθαρρύνει τις ανοικτές συζητήσεις, όπου ένα πλούσιο δίκτυο διαφωνιών και συμφωνιών σχηματίζεται μεταξύ των χρηστών. Επιπλέον, αυτή η μη περιορίσιμη δομή της συζήτησης αποτελεί και μια μεγάλη ευκαιρία για μακρές συζητήσεις που μπορούν να εκτείνονται σε πολλά επιμέρους θέματα, που συνήθως όλα είναι εξίσου σημαντικά για την πλήρη κατανόηση της συζήτησης.

Σε αυτή την εργασία, επικεντρωνόμαστε στο να γίνουν οι διάφορες απόψεις που βρίσκονται σε διαδικτυακά φόρουμ, εύκολα προσβάσιμες σε όλους τους συμμετέχοντες του φόρουμ, αλλά και σε κάθε εξωτερικό ενδιαφερόμενο, χωρίς να απαιτείται η χρονοβόρα λεπτομερής ανάγνωση μιας συζήτησης στο σύνολό της. Πιο συγκεκριμένα, δημιουργούμε μια περίληψη της συζήτησης εξάγοντας ένα μικρό ποσοστό των αναρτήσεων (οι αναρτήσεις είναι το δομικό στοιχείο του νήματος – κάθε νήμα αναπαριστά μια συζήτηση), με τρόπο ώστε να επιτυγχάνεται η μέγιστη κάλυψη του θέματος και να παρουσιάζονται όλες οι διαφορετικές απόψεις των συμμετεχόντων.

Για παράδειγμα, ας μελετήσουμε μια συζήτηση σχετικά με τα ηθικά ζητήματα που αφορούν την άμβλωση. Καθώς η συζήτηση προχωρεί, διάφορες υπο-συζητήσεις αρχίζουν να προκύπτουν, όσον αφορά τη σύνδεση της νομοθεσίας για τις αμβλώσεις με την ηθική του θέματος, τα διάφορα επιχειρήματα σχετικά με το πότε αρχίζει η ζωή και πότε ένα έμβρυο γίνεται άνθρωπος κλπ. Με την εφαρμογή αλγορίθμων εξαγωγής πληροφορίας τους οποίους παρουσιάζουμε σε αυτήν την εργασία, η συζήτηση χωρίζεται σε επιμέρους συζητήσεις και κάθε μία σχετίζεται με τις λέξεις-κλειδιά που περιγράφουν το θέμα της. Παράλληλα, αναγνωρίζονται οι συμφωνίες και οι διαφωνίες μεταξύ των συμμετεχόντων και στη συνέχεια σχηματίζονται ομάδες από χρήστες που συμφωνούν μεταξύ τους. Τέλος, όλες οι παραπάνω πληροφορίες συνδυάζονται και εξάγεται μια μικρή περίληψη για κάθε επιμέρους συζήτηση, η οποία περιέχει και τις πιο χαρακτηριστικές και σημαντικές αναρτήσεις ώστε να τονίζονται οι διάφορες και διαφορετικές απόψεις των χρηστών. Η μέθοδός μας αναπαριστάται γραφικά στο σχήμα 1 παρακάτω.

Σχήμα 1: Αφαιρετικό διάγραμμα ροής της μεθόδου μας

Αρχικά, δημιουργείται ο γράφος απαντήσεων, ο οποίος παρέχει πληροφορίες σχετικά με τους στόχους των απόψεων κάθε χρήστη, δηλαδή, ποιος απευθύνεται σε ποιον όταν γίνεται μια νέα ανάρτηση (post). Στη συνέχεια ο γράφος αυτός χρησιμοποιείται στα δύο, ως επί το πλείστον ανεξάρτητα, κομμάτια της μεθόδου που ακολουθούν, δηλαδή την εξόρυξη του θέματος και τον εντοπισμό διαφωνιών.

Στο κομμάτι της εξόρυξης του θέματος αναγνωρίζονται οι επιμέρους συζητήσεις μέσα σε μία μεγαλύτερη και γενικότερη συζήτηση, με βάση την παρατήρηση ότι οι αναρτήσεις που αποτελούν απάντηση από ένα χρήστη σε έναν άλλο (και δεν είναι γενικές τοποθετήσεις χωρίς συγκεκριμένο παραλήπτη) είναι πολύ πιθανό να αφορούν περίπου το ίδιο θέμα με την ανάρτηση στην οποία απαντάνε. Αυτό μας δίνει τη δυνατότητα να μπορούμε να προσδιορίσουμε τοπικές ομάδες αναρτήσεων μέσα σε μια συζήτηση που αφορούν το ίδιο υπο-θέμα (θεματικό πλαίσιο ανάρτησης – thematic context), το οποία περιγράφεται από λέξεις-κλειδιά τα οποία εξάγουμε με χρήση παραδοσιακών αλγορίθμων εξαγωγής λέξεων-κλειδιών. Τέλος, γίνεται συγχώνευση όσων ομάδων αναρτήσεων τυχαίνει να έχουν κοινές λέξεις-κλειδιά, καθώς αυτό σημαίνει ότι αναφέρονται στο ίδιο θέμα. Οι τελικές ομάδες αναρτήσεων που μένουν ορίζουν τις

διάφορες επιμέρους συζητήσεις της συνομιλίας (που αφορούν και διαφορετικά υπο-θέματα).

Στο κομμάτι του εντοπισμού διαφωνιών δημιουργούνται ομάδες χρηστών οι οποίοι συμφωνούν σε αυτά που λένε και στη συνέχεια προσδιορίζεται ποιες από αυτές τις ομάδες έχουν διαφορετική άποψη πάνω στο ίδιο (υπο-)θέμα. Συμφωνίες και διαφωνίες εντοπίζονται με βάση τη συμπεριφορά των συμμετεχόντων σε μια συζήτηση και τον τρόπο που αλληλεπιδρούν μεταξύ τους. Η ανάλυση αυτή εφαρμόζεται μόνο σε τμήματα του γράφου απαντήσεων που προσδιορίζονται ως πολύ πιθανό να περιέχουν πλούσια αλληλεπίδραση μεταξύ των συμμετεχόντων (φάσεις "βρασμού" της συζήτησης). Είναι σημαντικό να σημειωθεί ότι δεν γίνεται χρήση μεθόδων επεξεργασίας φυσικής γλώσσας στο κομμάτι αυτό.

Οι επιμέρους συζητήσεις που εντοπίζουμε και οι αντίστοιχες λέξεις-κλειδιά τους, οι ομάδες συμφωνούντων χρηστών και οι διαφωνίες μεταξύ τους, καθώς και οι περιλήψεις που εν τέλει δημιουργούνται, αποθηκεύονται σε μια βάση δεδομένων. Όλη αυτή η πληροφορία μπορεί στη συνέχεια να χρησιμοποιηθεί με πολλούς και χρήσιμους τρόπους:

1. Από τους χρήστες του φόρουμ ώστε να μπορούν να καταλάβουν εύκολα και γρήγορα τι ακριβώς συζητιέται, και σε τι επίπεδο, σε ένα νήμα (thread).

2. Από συστήματα συστάσεων για τον εντοπισμό συζητήσεων που μπορεί να ενδιαφέρουν ένα συγκεκριμένο χρήστη του φόρουμ, για προτάσεις φιλίας μεταξύ χρηστών κλπ.

3. Από πράκτορες που θα μπορούν να κάνουν επερωτήσεις στη βάση δεδομένων μας χρησιμοποιώντας την παραδοσιακή αναζήτηση λέξεων-κλειδιών, και να ανακτούν περιλήψεις της γνώμης του κοινού πάνω σε διάφορα θέματα που μπορεί να τους ενδιαφέρουν. Οι περιλήψεις αυτές μπορούν να χρησιμοποιηθούν και για την υποστήριξη διαφόρων διαδικασιών λήψης αποφάσεων, να βοηθήσουν στον προσδιορισμό τρεχουσών τάσεων, στην ενίσχυση της έρευνας με διάφορα κοινωνικά πειράματα που είναι αναγκαίος μεγάλος όγκος απόψεων κλπ.

Ελπίζουμε ότι με τα αποτελέσματα της παρούσας εργασίας, θα μπορέσουμε να προσφέρουμε μια στέρεη βάση για τη μελλοντική έρευνα, και επίσης, να βοηθήσουμε να επισημανθεί η χρησιμότητα και μοναδικότητα, των απόψεων που βρίσκονται σε χώρους συζητήσεων του Παγκόσμιου Ιστού.

2. Σχετική Βιβλιογραφία και Προγενέστερη Δουλειά

Η εργασία μας αποτελείται από τρία κύρια μέρη: την ανάλυση της ροής της συζήτησης, την εκτίμηση διαφωνιών μεταξύ των χρηστών, και την περίληψη των απόψεων. Αν και υπάρχει αρκετή και σημαντική δουλειά σε όλους αυτούς τους τομείς, δεν έχουν, μέχρι τώρα, συνδυαστεί και εφαρμοστεί σε διαδικτυακά φόρουμ με σκοπό την περιληπτική παρουσίαση των απόψεων.

Η μέχρι τώρα έρευνα πάνω στην ανάλυση της ροής διαδικτυακών συζητήσεων βασίζεται κυρίως σε γλωσσολογικές μεθόδους, όπως η λεκτική ομοιότητα και η θεματική απόσταση για την ταξινόμηση των αναρτήσεων σε διάφορα θέματα [14, 15]. Η ιδέα της χρήσης δομικών στοιχείων μιας συζήτησης προτείνεται στις εργασίες [12, 13]. Στην [12], η ώρα της δημοσίευσης μιας ανάρτησης, το χρονικό διάστημα μεταξύ των δημοσιεύσεων των αναρτήσεων, καθώς και τα ονόματα των χρηστών που συμμετέχουν στη συζήτηση χρησιμοποιούνται για την εξαγωγή σχέσεων που θα μπορούσαν ενδεχομένως να αντιστοιχούν σε θεματικές ομοιότητες. Στην [13], ο μηχανισμός παραθέσεων χρησιμοποιείται για να αποφασιστεί εάν μια ανάρτηση έχει σχέση με το αρχικό θέμα της συζήτησης, αλλά τα αποτελέσματα δεν χρησιμοποιούνται για την ανάλυση της ροής του θέματος. Τέλος, μια μέθοδος που χρησιμοποιεί τα ειδικά δομικά χαρακτηριστικά της δομής των διαδικτυακών φόρουμ για την ανάλυση της ροής της συζήτησης προτείνεται στην [11]. Η προσέγγισή μας διαφέρει από τις υπάρχουσες εργασίες σε δύο βασικά σημεία: 1) Υποθέτουμε μηδενική εκ των προτέρων γνώση του θέματος που συζητιέται και 2) επικεντρωνόμαστε στη δομική και όχι λεκτική πληροφορία των συζητήσεων.

Έρευνα πάνω στην αναγνώριση συμφωνιών και διαφωνιών σε διαδικτυακές συνομιλίες έχει πραγματοποιηθεί στις εργασίες [1, 2, 3, 4]. Κοινό χαρακτηριστικό αυτών των εργασιών είναι ότι χρησιμοποιούν δεδομένα από ζωντανές συναντήσεις και εφαρμόζουν τις μεθόδους τους στην καταγραφή των συνομιλιών. Στην [4], κυρίως χρησιμοποιούνται λεξιλογικά και προσωδιακά χαρακτηριστικά (π.χ. παύσεις) σε μια μη εποπτευόμενη προσέγγιση μηχανικής μάθησης. Οι εργασίες που ακολούθησαν βελτίωσαν τη δουλειά σε αυτόν τον τομέα, με τη δοκιμή και την εφαρμογή διαφόρων αλγορίθμων μηχανικής μάθησης. Ιδιαιτέρου ενδιαφέροντος χρήζει η [2], όπου εξετάζεται το πρόβλημα προσδιορισμού του στόχου μιας γνώμης που εκφράζεται από κάποιο συνομιλητή. Η εργασία μας διαφέρει πολύ από τις προσεγγίσεις αυτές, δεδομένου ότι δεν γίνεται χρήση μεθόδων επεξεργασίας φυσικής γλώσσας. Αντ' αυτού, ακολουθούμε μια νέα προσέγγιση όπου αναλύουμε την ίδια τη συζήτηση, και στη συνέχεια δημιουργούμε

χαρακτηριστικά βάσει στατιστικών και αναλύσεων συμπεριφοράς πάνω στις αλληλεπιδράσεις μεταξύ των συνομιλητών.

Σχετικά με την περιλήψεων απόψεων, το μεγαλύτερο κομμάτι της τρέχουσας έρευνας έχει επικεντρωθεί στις περιλήψεις διαφόρων ειδών κριτικών και τοποθετήσεων που βρίσκονται στον παγκόσμιο ιστό [5, 6, 7]. Δουλειές περισσότερο σχετικές με την παρούσα εργασία μπορούν να βρεθούν στις [8, 9]. Οι εργασίες αυτές επικεντρώνονται στην περιληπτική παρουσίαση θέσεων σε ιστολόγια και των σχολίων που γίνονται σε αυτά, και προσπαθούν να αναδείξουν απόψεις με διαφορετική πολικότητα. Για τον προσδιορισμό της πολικότητας μιας γνώμης (αρνητική, θετική ή ουδέτερη) αλγόριθμοι ανάλυσης συναισθημάτων χρησιμοποιούνται οι οποίοι κάνουν χρήση προκαθορισμένων βαθμολογιών συναισθήματος για διάφορες λέξεις που εκφράζουν γνώμη. Για παράδειγμα, η λέξη «όμορφη» έχει ένα θετικό συναίσθημα, και οι απόψεις που την περιέχουν μπορούν να θεωρηθούν θετικές. Ωστόσο, είναι πολύ εύκολο να σκεφτούμε ρεαλιστικά παραδείγματα όπου η πολικότητα μιας γνώμης, όπως προσδιορίζεται από την ανάλυση συναισθήματος, θα μπορούσε να είναι διαφορετική για δύο πολύ παρόμοιες απόψεις, είτε η ίδια για δύο διαφορετικές απόψεις. Για παράδειγμα, υποθέτουμε τα εξής δύο σχόλια από ένα ιστολόγιο, σχετικά με τις προηγούμενες προεδρικές εκλογές στις ΗΠΑ:

1. Συμφωνώ, ο Μπάρακ Ομπάμα έδωσε μια πολύ εμπνευσμένη ομιλία χθες, και φαίνεται ότι πάει κατευθείαν για την Προεδρία.

2. Συγκρίνοντας όλες τις ομιλίες που έχω ακούσει και από τους δύο, μπορώ να σας πω ότι ο Μακέιν είναι απλά απαίσιος στο να παρουσιάζει τις ιδέες του, αντιθέτως με τον Ομπάμα.

Στην εργασία μας, δεν χρησιμοποιούμε μεθόδους ανάλυσης συναισθήματος για να βρούμε αντίθετες απόψεις. Αντ' αυτού, εξετάζουμε την αλληλεπίδραση των συνομιλητών για να βρούμε χαρακτηριστικά πρότυπα συμπεριφοράς τα οποία να αποκαλύπτουν τη διαφωνία τους, και στη συνέχεια ομαδοποιούμε τους χρήστες που συμφωνούν για την αποφυγή επανάληψης ίδιων απόψεων.

Μια πρόσφατη και πολύ ενδιαφέρουσα εργασία σχετικά με περίληψη απόψεων μπορεί να βρεθεί στην [10]. Παρά το γεγονός ότι γίνεται ξανά χρήση της πολικότητας των απόψεων, και η μέθοδος έχει δοκιμαστεί σε σχόλια και κριτικές, το θεωρητικό πλαίσιο που παρουσιάζεται μπορεί να είναι χρήσιμο για πιο γενικές εργασίες περίληψης και εξαγωγής απόψεων.

3. Γράφος Απαντήσεων

Η καινοτομία της προτεινόμενης μεθόδου μας έγκειται στο γεγονός ότι αναλύουμε κυρίως τις δομικές πληροφορίες των διαδικτυακών φόρουμ για την επίτευξη των επιμέρους εργασιών που περιγράφηκαν στην εισαγωγή. Ο τρόπος που οι συνομιλητές απαντάνε μεταξύ τους σε ένα νήμα προσφέρει πολύ σημαντικές πληροφορίες σχετικά με τη ροή των θεμάτων που συζητήθηκαν και τις σχέσεις συμφωνίας ή διαφωνίας μεταξύ τους. Για την εξαγωγή της δομικής αυτής πληροφορίας, παριστάνουμε μια συζήτηση - νήμα ως ένα γράφο με αναρτήσεις για κόμβους και σχέσεις που υποδεικνύουν ότι μια ανάρτηση απαντάει σε μια άλλη σαν ακμές. Μια τέτοια σχέση μεταξύ δύο αναρτήσεων (σχέση απάντησης) δηλώνει ότι ο συγγραφέας της δεύτερης ανάρτησης απαντάει στο συγγραφέα της πρώτης, και αυτό υποδεικνύεται από την κατεύθυνση των ακμών. Οι σχέσεις αυτές μπορεί να είναι είτε άμεσες είτε έμμεσες.

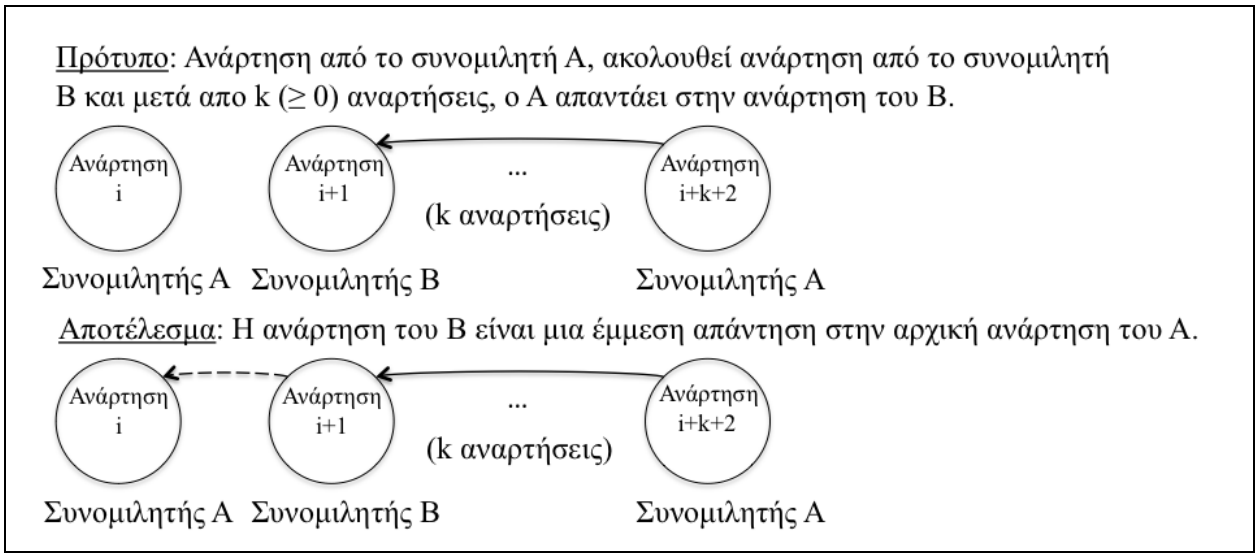
Ο τρόπος που σχηματίζονται άμεσες σχέσεις μεταξύ αναρτήσεων, είναι μέσω του μηχανισμού παραθέσεων. Ο μηχανισμός αυτός δηλώνει σε ποια ανάρτηση απαντάει μια άλλη. Είναι πολύ παρόμοιος με το μηχανισμό παραθέσεων του ηλεκτρονικού ταχυδρομείου όπου ένας παραθέτει το τμήμα του κειμένου στο οποίο θέλει να απαντήσει. Ομοίως, τα διαδικτυακά φόρουμ προσφέρουν αυτό το βασικό εργαλείο στους χρήστες τους ώστε να είναι οι συζητήσεις σαφείς και κατανοητές. Χρησιμοποιούμε εκτεταμένα αυτήν την πληροφορία για τη δημιουργία του γράφου απαντήσεων. Ας σημειωθεί εδώ ότι μία ανάρτηση μπορεί να απαντήσει σε πολλές άλλες ή να έχει σχέσεις απάντησης με πολλές αναρτήσεις που την ακολουθούν.

Πέρα από τον εύκολα αναγνωρίσιμο μηχανισμό παραθέσεων, υπάρχουν και αρκετοί άλλοι, πιο έμμεσοι τρόποι με τον οποίο οι συνομιλητές απαντάνε ο ένας στον άλλο. Έχουμε εντοπίσει αρκετά τέτοια πρότυπα απαντήσεων και τα χρησιμοποιήσαμε για να αναγνωρίζουμε πολύ περισσότερες περιπτώσεις απαντήσεων. Αυτά τα πρότυπα αναλύονται διεξοδικά στο επόμενο υποκεφάλαιο.

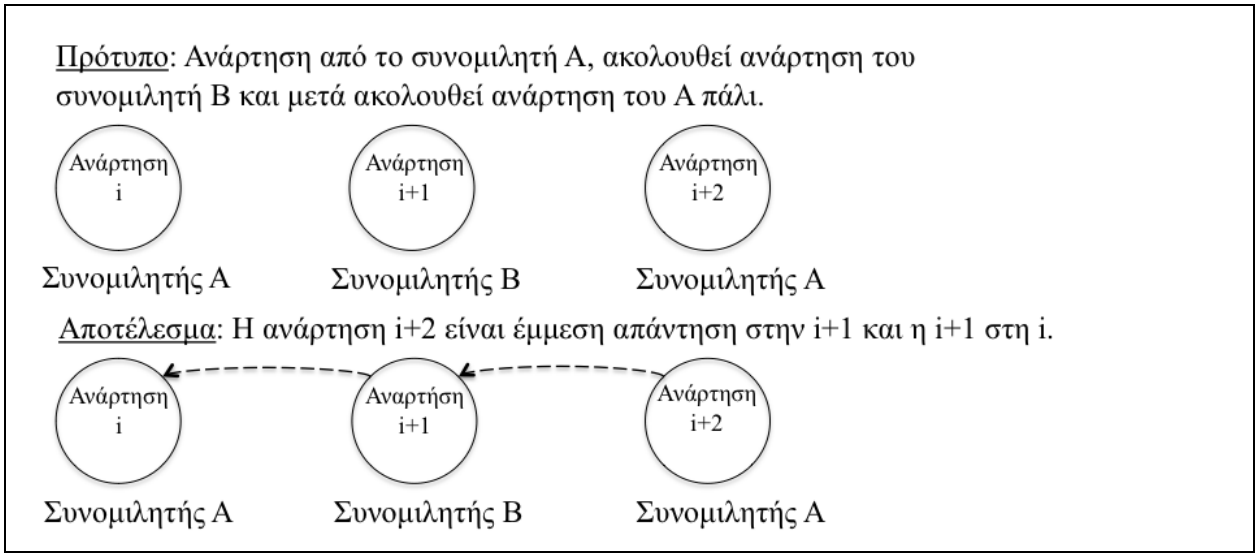
3.1 Πρότυπα Έμμεσων Απαντήσεων

Τα πρότυπα απαντήσεων που έχουμε εντοπίσει περιγράφονται καλύτερα με γραφική απεικόνιση. Στα γραφήματα των σχημάτων 2 έως 7, οι κόμβοι αντιστοιχούν σε αναρτήσεις και επισημαίνονται με έναν κωδικό αριθμό ενδεικτικό της θέσης τους στη συζήτηση. Κάτω από κάθε κόμβο υπάρχει το όνομα του συγγραφέα της ανάρτησης. Οι στερεές ακμές υποδεικνύουν υπάρχουσες σχέσεις απάντησης, δηλαδή εκείνες που εξάγονται από τον μηχανισμό παραθέσεων ή από προηγούμενες εφαρμογές των

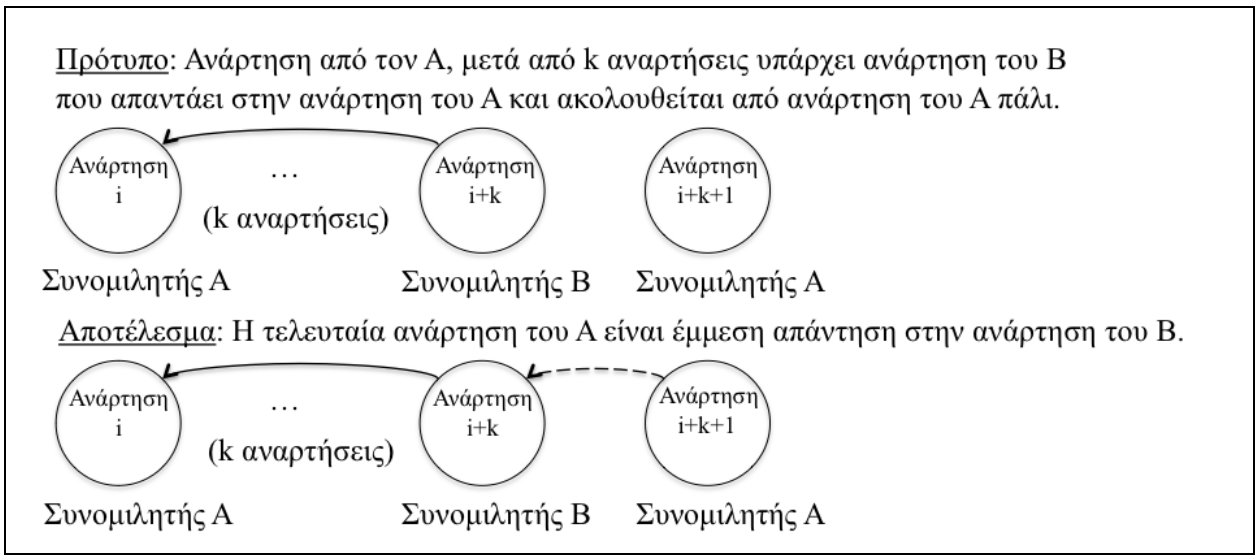
προτύπων έμμεσων απαντήσεων, ενώ οι διακεκομμένες ακμές προσδιορίζουν τις έμμεσες σχέσεις απάντησης που θα αναγνωριστούν και δημιουργηθούν με την εφαρμογή του εκάστοτε προτύπου. Η διαδικασία ξεκινάει με ένα γράφο απαντήσεων που περιλαμβάνει μόνο τις σχέσεις απαντήσεων από παραθέσεις (άμεσες) και τελειώνει με την ολοκλήρωση του γράφου ώστε να περιέχει και τις έμμεσες σχέσεις απαντήσεων που προσδιορίζονται από τα πρότυπα. Σε κάθε βήμα, οι έμμεσες σχέσεις που έχουν βρεθεί μέχρι εκείνο το σημείο μπορούν να χρησιμοποιηθούν. Ξεκινώντας από την πρώτη ανάρτηση του νήματος και διατρέχοντας τις αναρτήσεις κατά αύξουσα σειρά αναγνωριστικού αριθμού διασφαλίζουμε ότι αναγνωρίζονται όλα τα πρότυπα που υπάρχουν ή που σχηματίζονται ενώ τρέχει ο αλγόριθμος εντοπισμού τους.



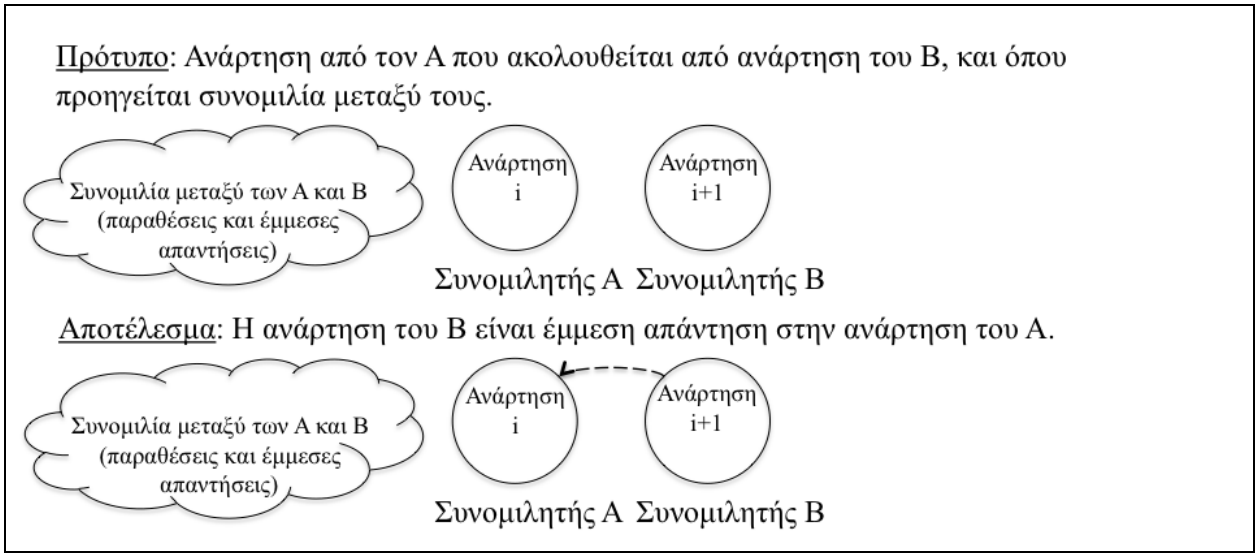
Σχήμα 2: Πρότυπο έμμεσων απαντήσεων #1



Σχήμα 3: Πρότυπο έμμεσων απαντήσεων #2

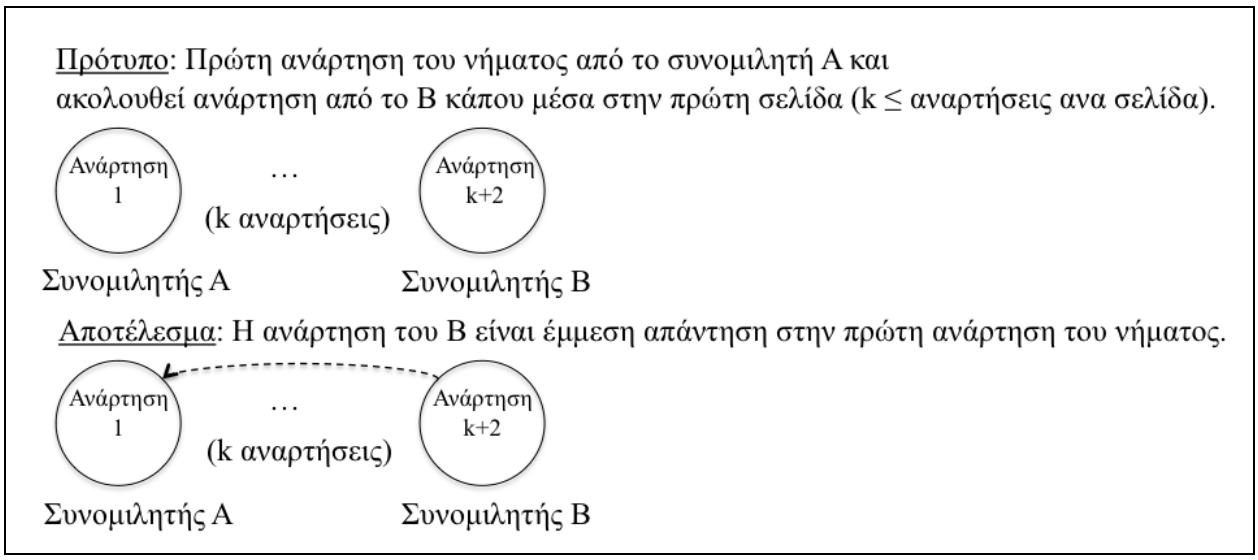


Σχήμα 4: Πρότυπο μη-ρητών απαντήσεων #3



Σχήμα 5: Πρότυπο έμμεσων απαντήσεων #4

Όλα τα παραπάνω πρότυπα βασίζονται κυρίως στην υπόθεση ότι, όταν δύο συνομιλητές αλληλεπιδρούν σε κάποιο μέρος της συζήτησης, τότε δύο αναρτήσεις τους οι οποίες και είναι κοντά μέσα στο νήμα (αλλά χωρίς να περιέχουν παραθέσεις) θα μπορούσαν με ασφάλεια να συνδεθούν με μια ακμή, δεδομένου ότι είναι πολύ πιθανό να απαντάνε η μία στην άλλη.

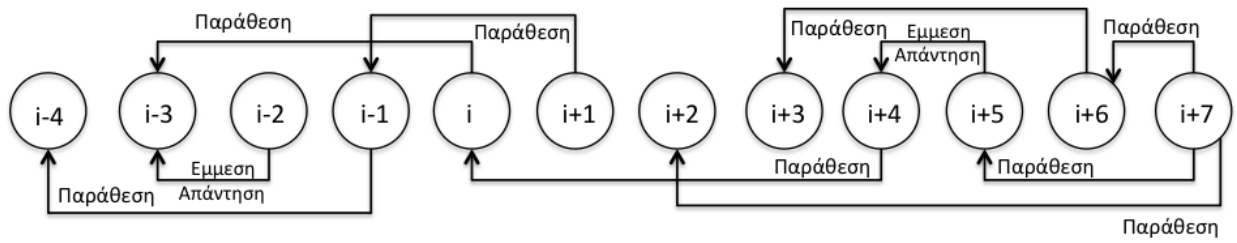


Σχήμα 6: Πρότυπο έμμεσων απαντήσεων #5

Το τελευταίο πρότυπο (# 5), βασίζεται στην παρατήρηση ότι στην πρώτη σελίδα ενός νήματος, οι συνομιλητές πολύ συχνά απαντάνε χωρίς να χρησιμοποιούν το μηχανισμό παράθεσης γιατί είναι σαφές ότι απαντάνε στο συγγραφέα της πρώτης ανάρτησης (εκκινητής του νήματος). Γενικά, σε όλες τις περιπτώσεις που περιγράφονται εδώ, οι συνομιλητές επέλεξαν να μην χρησιμοποιήσουν το μηχανισμό παραθέσεων διότι είναι σαφές σε κάθε περίπτωση ότι απαντάνε στην αμέσως προηγούμενη ανάρτηση (ή την πρώτη του νήματος) και συνεπώς μια παράθεση δεν είναι απαραίτητη.

3.2 Ακμές Γράφου και Βάρη

Στο τέλος, ο γράφος απαντήσεων περιέχει όλες τις αναρτήσεις της συζήτησης ως κόμβους και όλες τις σχέσεις απάντησης ως ακμές, όπως αυτές εξήχθησαν χρησιμοποιώντας το μηχανισμό παραθέσεων και τα πρότυπα εμμέσων απαντήσεων. Οι ακμές κατευθύνονται από μια ανάρτηση προς αυτές στις οποίες απαντάει. Ένα παράδειγμα δίνεται στο σχήμα 7. Κάθε ακμή συνοδεύεται με την πληροφορία του αν η αντίστοιχη σχέση απάντησης προέκυψε από παράθεση ή σαν έμμεση απάντηση.



Σχήμα 7: Υποθετικός γράφος απαντήσεων μια συζήτησης

Για το σκοπό της εξαγωγής του θέματος (κεφάλαιο 4), κάθε ακμή του γράφου έχει ένα βάρος αναλόγως του πόσο μπορεί το θέμα να διατηρείται αμετάβλητο μεταξύ των δύο

συνδεδεμένων αναρτήσεων. Με άλλα λόγια, πόσο πιθανό είναι οι δύο αναρτήσεις να αφορούν το ίδιο ακριβώς θέμα. Υπάρχουν τρεις μετρικές που χρησιμοποιούμε για την εκτίμηση αυτή:

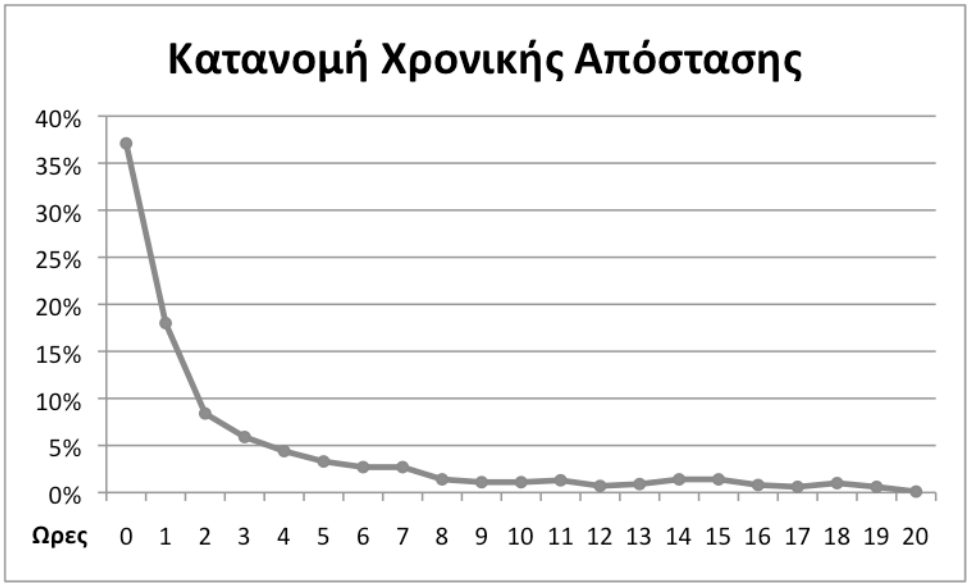
1. Βεβαιότητα Απάντησης: Για τις σχέσεις απάντησης που έχουν προσδιοριστεί μέσω του μηχανισμού παραθέσεων, αυτή η μετρική παίρνει τιμή 1, υποδεικνύοντας μηδενική αβεβαιότητα αφού οι συνομιλητές παραθέτουν μόνο αναρτήσεις που θέλουν να σχολιάσουν ή στις οποίες θέλουν να αναφερθούνε, οπότε το θέμα είναι πολύ συχνά το ίδιο. Για τις σχέσεις που ορίζονται μέσα από τα 3 πρώτα πρότυπα εμμέσων απαντήσεων η μετρική παίρνει τιμή 0,8, υποδεικνύοντας έναν πολύ μικρό βαθμό αβεβαιότητας, ενώ οι σχέσεις απάντησης που ορίζονται από τα δύο τελευταία πρότυπα δίνουν τιμή 0,6. Οι αριθμοί αυτοί έχουν εμπειρικά εκτιμηθεί μετά από την εξέταση πολλών συζητήσεων και τη μέτρηση της αναλογίας των λάθος κατηγοριοποιήσεων που προκύπτουν σε κάθε απάντηση-πρότυπο.

2. Απόσταση Αναρτήσεων: Αυτή η μετρική είναι ίση με τον αριθμό των αναρτήσεων που υπάρχουν στο νήμα μεταξύ δύο συνδεδεμένων αναρτήσεων στο γράφο. Όσο μεγαλύτερη είναι η απόσταση, τόσο λιγότερο πιθανό οι συνδεδεμένες αναρτήσεις να αναφέρονται στο ίδιο θέμα (δεδομένου ότι το θέμα μιας συζητητής εν γένει μεταβάλλεται).

3. Χρονική Απόσταση: Η μετρική αυτή είναι ίση με τον χρόνο (σε λεπτά) μεταξύ των δημοσιεύσεων δύο αναρτήσεων. Και πάλι, όσο περισσότερος είναι ο χρόνος, τόσο το λιγότερο πιθανό οι αναρτήσεις αυτές να αναφέρονται στο ίδιο θέμα.

Στα σχήματα 8 και 9 εμφανίζεται η κατανομή του χρόνου και της απόστασης αναρτήσεων μεταξύ των συνδεδεμένων αναρτήσεων ενός νήματος. Εξετάσαμε 3000 αναρτήσεις από ένα διαδικτυακό φόρουμ, όπου τα νήματα περιέχουν 15 αναρτήσεις ανά σελίδα. Να σημειωθεί ότι στο σχήμα 9 (χρονική απόσταση), ο οριζόντιος άξονας έχει μονάδα μέτρησης ώρες και όχι λεπτά που χρησιμοποιούμε στη μετρική χρονικής απόστασης.

Σχήμα 8: Κατανομή απόστασης αναρτήσεων



Σχήμα 9: Κατανομή χρονικής απόστασης

Είναι προφανές ότι οι περιπτώσεις που κυριαρχούν είναι εκείνες όπου οι αναρτήσεις απαντούν στην ακριβώς τελευταία ανάρτηση του νήματος (ή την προηγούμενη) και σε πολύ σύντομο χρονικό διάστημα (1-2 ώρες). Επίσης, είναι σαφές ότι οι χρήστες των διαδικτυακών φόρουμ τείνουν να απαντάνε σε αναρτήσεις των οποίων η δημοσίευση δεν έγινε παρά 7 ώρες πριν, το πολύ (δηλαδή είναι σχετικά πρόσφατες). Επίσης, εν γένει απαντάνε μεταξύ των 15 τελευταίων αναρτήσεων και όχι στις παλαιότερες (να σημειωθεί ότι 15 είναι επίσης ο αριθμός των αναρτήσεων ανά σελίδα στο διαδικτυακό φόρουμ που αναλύσαμε). Θέτουμε τους αριθμούς αυτούς των αναρτήσεων (15) και του χρόνου (7 ώρες) ως όρια πέραν των οποίων οι αναρτήσεις που συνδέονται μεταξύ τους είναι πολύ πιθανό πλέον να αφορούν διαφορετικά θέματα.

Τα όρια αυτά τα χρησιμοποιούμε στον τύπο που έχουμε ορίσει για την ανάθεση βαρών σε ακμές (Σχήμα 10) και προσδιορίζουν πότε οι αποστάσεις αναρτήσεων και χρόνου θα πρέπει να χρησιμοποιούνται για τον υπολογισμό του βάρους. Αυτό σημαίνει ότι εντός των παραπάνω ορίων ο υπολογισμός γίνεται με διαφορετικό τρόπο αφού η συμπεριφορά των χρηστών είναι και διαφορετική. Στις περιπτώσεις όπου οι αποστάσεις αναρτήσεων και χρόνου δεν υπερβαίνουν αυτά τα όρια, μόνον ο παράγοντας βεβαιότητας απάντησης καθορίζει την ομοιότητα του θέματος δύο συνδεδεμένων αναρτήσεων (περισσότερες λεπτομέρειες υπάρχουν στο υποκεφάλαιο 3.3).

$$\text{βάρος} = \begin{cases} \text{αν time\_distance} > 7\text{ώρες } \text{Η΄ post\_distance} > \text{ppp:} \\ \dfrac{[\alpha * \text{post\_distance} + \beta * \text{time\_distance}]}{\alpha + \beta} * \text{certainty\_factor} \\ \text{αλλιώς:} \\ \quad \text{certainty\_factor} \end{cases}$$

Σχήμα 10: Βάρος ακμής (i, j)

Στον τύπο του βάρος στο σχήμα 10, οι σταθερές α και β είναι αριθμοί που εξομαλύνει τις τιμές των αποστάσεων, δεδομένου ότι χρησιμοποιούν διαφορετικές μονάδες (αναρτήσεις και τα λεπτά αντίστοιχα). Η μεταβλητή ppp («posts per page») περιέχει τον αριθμό των αναρτήσεων ανά σελίδα (στο παράδειγμά μας είναι ίσος με 15). Οι μεταβλητές time_distance (χρονική απόσταση), post_distance (απόσταση αναρτήσεων) και certainty_factor (παράγοντας βεβαιότητας) παίρνουν τιμές που αντιστοιχούν στις μετρικές μεταξύ των κόμβων i και j.

3.3 Θεματικό Πλαίσιο Ανάρτησης

Τέλος, είναι σημαντικό να εισαχθεί ο όρος «θεματικό πλαίσιο ανάρτησης» (στο οποίο θα αναφερόμαστε και σαν «θεματικό πλαίσιο»), που χρησιμοποιούμε στις μεθόδους μας για την ανάλυση της ροής του θέματος και τη δημιουργία ομάδων συμφωνούντων χρηστών. Κάθε ανάρτηση περιβάλλεται από πολλές άλλες, μέσα σε ένα νήμα. Ωστόσο, σε πολλές περιπτώσεις υπάρχουν κοντινές αναρτήσεις, οι οποίες όμως δεν αναφέρονται στο ίδιο ακριβώς θέμα. Για παράδειγμα, όταν υπάρχει μια ξαφνική αλλαγή του θέματος θα βρεθούν αναρτήσεις που είναι πολύ κοντά, αλλά με διαφορετικό θέμα. Στην περίπτωση αυτή, το θεματικό πλαίσιο της κάθε ανάρτησης θα πρέπει να περιλαμβάνει μόνο τις αναρτήσεις που έχουν παρόμοιο θέμα, και δεν αφορούν την νέα κατεύθυνση που πήρε η συζήτηση. Η μέθοδος που χρησιμοποιούμε για να

καθορίσουμε αυτά τα θεματικά πλαίσια περιγράφεται στο επόμενο κεφάλαιο (υποκεφάλαιο 4.1).

4. Ανάλυση Ροής Συζήτησης (Thread Extraction)

Σε αυτό το κεφάλαιο, περιγράφουμε τη μέθοδό μας για την ανάλυση της ροής μιας συζήτησης. Η ανάλυση αυτή έχει να κάνει με τον εντοπισμό αλλαγών στη ροή του θέματος μιας συζήτησης και παρουσιάζουμε αλγορίθμους που το επιτυγχάνουν με την ομαδοποίηση αναρτήσεων που αφορούν το ίδιο θέμα. Η κύρια ιδέα είναι η δημιουργία θεματικών λέξεων-κλειδιών (παράγραφος 4.2) για κάθε ανάρτηση ενός νήματος και στη συνέχεια η ομαδοποίηση αυτών σε υπο-συζητήσεις (παράγραφος 4.3). Είναι σημαντικό εδώ να τονίσουμε ότι η συζήτηση είναι όλο το νήμα, ενώ μια υπο-συζήτηση είναι ένα μέρος της συζήτησης που έχει ένα συγκεκριμένο και "σταθερό" θέμα σε όλες τις αναρτήσεις της.

Η σημαντικότερη πρόκληση που έπρεπε να αντιμετωπίσουμε σε αυτό το έργο ήταν ότι οι περισσότερες αναρτήσεις δεν περιέχουν επαρκή θεματική πληροφορία (προκύπτουν πολύ λίγες ή και καθόλου λέξεις-κλειδιά). Για να βρίσκουμε επιτυχώς λέξεις-κλειδιά για κάθε ανάρτηση, εκτός του ίδιου του περιεχομένου της ανάρτησης επεκτείνουμε την ανάλυση και στο θεματικό πλαίσιό της (δηλαδή στις σχετικές αναρτήσεις που αφορούν το ίδιο θέμα).

4.1 Γειτονίες Αναρτήσεων

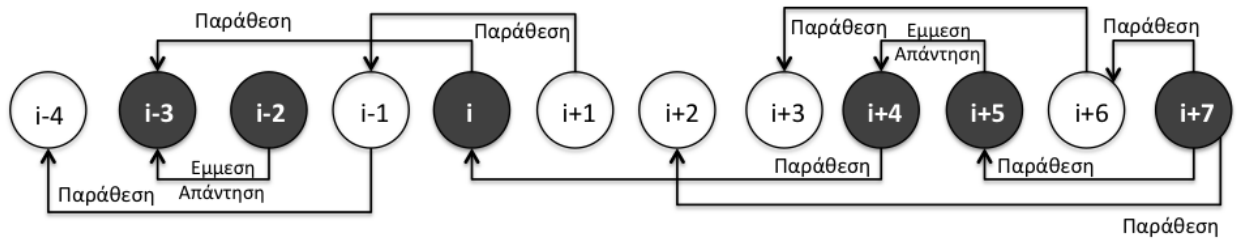
Για κάθε ανάρτηση στο γράφο απαντήσεων είμαστε σε θέση να δημιουργήσουμε μια ομάδα από άλλες αναρτήσεις που σχετίζονται με αυτή μέσα από σχέσεις απαντήσεων. Αυτή η ομάδα χρησιμοποιείται ως το θεματικό πλαίσιο της ανάρτησης και οι αναρτήσεις που περιέχονται είναι πολύ πιθανό να αφορούν το ίδιο ακριβώς θέμα.

Για να βρούμε τους γείτονες μιας ανάρτησης στο γράφο απαντήσεων (π.χ., για την ανάρτηση με αναγνωριστικό αριθμό i) συλλέγουμε όλες τις αναρτήσεις που συνδέονται με αυτήν μέσω ακμών απαντήσεων. Συνεχίζουμε να συλλέγουμε κόμβους που συνδέονται με τον i, όσο το βάρος του μονοπατιού που τους συνδέει δεν είναι μεγαλύτερο από μια τιμή w. Όταν δεν υπάρχουν άλλοι κόμβοι για να επισκεφτούμε (όλοι όσοι απομένουν βρίσκονται πιο μακριά από την απόσταση w) η επανάληψη τερματίζει. Η τιμή του w ορίζεται σαν η μέγιστη απόσταση αναρτήσεων και χρόνου γύρω από τον κόμβο i που το θέμα είναι συγκεκριμένο και "σταθερό" (και η τιμή αυτή είναι μεταβλητή ανάλογα το φόρουμ, το νήμα και τη συμμετοχή των χρηστών).

Για να εκτιμήσουμε τη τιμή του w, αναλύσαμε αρκετά νήματα και διαπιστώσαμε ότι πρέπει να είναι ανάλογο με την πυκνότητα της συζήτησης. Αυτό σημαίνει ότι όταν μια συζήτηση είναι πολύ αραιή, στο w πρέπει να ανατεθεί μια μεγάλη τιμή ώστε να

υπάρχουν επαρκείς αριθμοί γειτόνων για κάθε ανάρτηση. Η πυκνότητα ενός νήματος υπολογίζεται διαιρώντας τον αριθμό των αναρτήσεων με τη διαφορά της ώρας ανάμεσα στη δημοσίευση της πρώτης και τελευταίας ανάρτησης.

Για τη συλλογή όλων των γειτόνων μιας ανάρτησης χρησιμοποιούμε έναν αλγόριθμο αναζήτησης «πρώτα κατά πλάτος» όπου επισκεπτόμαστε όλους τους κόμβους ξεκινώντας από την ανάρτηση i και συνεχίζουμε όσο το βάθος της αναζήτησης (συνολικό βάρος των μονοπατιών που ακολουθούμε) δεν υπερβαίνει το w. Στο τέλος, κάθε ανάρτηση της συζήτησης είναι πλέον συνδεδεμένη με την αντίστοιχη ομάδα από γειτονικές αναρτήσεις. Αυτές οι ομάδες μπορεί να είναι επικαλυπτόμενες (ή ακόμη και ίδιες) για αναρτήσεις που είναι πολύ κοντά στο γράφο. Ένα παράδειγμα γειτονικών αναρτήσεων για την ανάρτηση i δίνεται στο Σχήμα 11, όπου οι γειτονικοί κόμβοι εμφανίζονται με σκούρο χρώμα. Αυτοί οι γείτονες όπως είπαμε, συνθέτουν το θεματικό πλαίσιο της ανάρτησης i.



Σχήμα 11: Παράδειγμα γειτόνων μιας ανάρτησης i στο γράφο απαντήσεων

4.2 Θεματικές Λέξεις-Κλειδιά

Το σύνολο των θεμάτων που σχετίζονται με μια ανάρτηση αναπαριστάται από ένα σύνολο λέξεων-κλειδιών. Οι θεματικές λέξεις-κλειδιά μπορούν να είναι είτε μεμονωμένες λέξεις είτε μικρές φράσεις. Το σύνολο των λέξεων αυτών για κάθε ανάρτηση εξάγεται σε δύο στάδια:

1. Γλωσσική ανάλυση του περιεχομένου της ανάρτησης

Ένας αλγόριθμος εξόρυξης θεματικών λέξεων-κλειδιών [16] εφαρμόζεται σε κάθε ανάρτηση της συζήτησης και παράγει λέξεις-κλειδιά καθώς και τις αντίστοιχες βαθμολογίες τους (πόσο σχετικές είναι με το θέμα της ανάρτησης). Όπως αναφέρθηκε προηγουμένως όμως, υπάρχουν αναρτήσεις με μη επαρκές περιεχόμενο που δεν παράγουν χρήσιμα αποτελέσματα. Για να λάβουμε πραγματικά χρήσιμα αποτελέσματα κατά τη διάρκεια αυτού του βήματος, οι αναρτήσεις πρέπει να περιέχουν τουλάχιστον 500 λέξεις κάτι το οποίο δεν συμβαίνει αρκετά συχνά. Πειραματική ανάλυση που κάναμε οδήγησε σε μηδενικό αριθμό θεματικών λέξεων-κλειδιών για περισσότερες από το 10%

των αναρτήσεων και σε μία μόνο λέξη-κλειδί για περισσότερες από το 15%. Αυτό αθροίζει στο 25% των αναρτήσεων να μην περιέχουν συγκεκριμένη θεματική πληροφορία.

2. Μεταβατική συσχέτιση των λέξεων-κλειδιών με χρήση των γειτονικών αναρτήσεων

Σε αυτό το βήμα συγχωνεύουμε όλες τις λέξεις-κλειδιά μιας ανάρτησης με τις λέξεις-κλειδιά των γειτονικών αναρτήσεών της και κρατάμε τις 5 πιο συχνά εμφανιζόμενες. Η ανάλυση συχνοτήτων κάθε λέξης γίνεται με βάση τη ρίζα της. Ορίζουμε ως συχνότητα ενός όρου το άθροισμα των επί μέρους βαθμολογιών του σε κάθε ανάρτηση στην οποία εμφανίζεται. Όλες οι λέξεις-κλειδιά τυγχάνουν ίσης μεταχείρισης. Με αυτόν τον τρόπο, ακόμη και για τις αναρτήσεις με φτωχό περιεχόμενο, συλλέγουμε λέξεις-κλειδιά από τις αναρτήσεις του θεματικού τους πλαισίου.

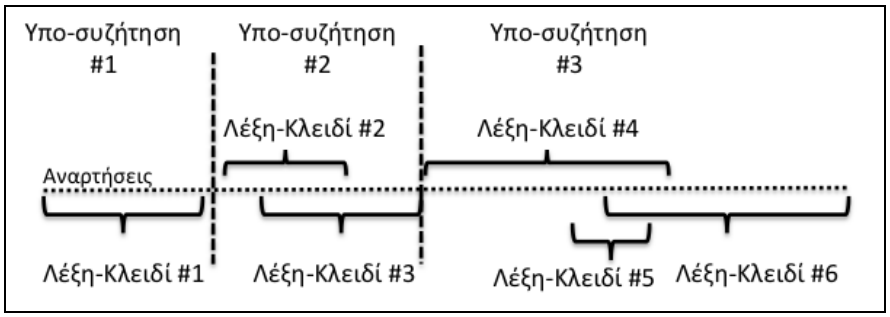
Αναρτήσεις που παραμένουν χωρίς θέμα (καμία ή μία θεματικές λέξεις-κλειδιά) ακόμα και μετά την παραπάνω διαδικασία, είναι αυτές με τόσο φτωχό περιεχόμενο όσο και θεματικό πλαίσιο. Ευτυχώς, τέτοιες περιπτώσεις φαίνεται να αποτελούν λιγότερο από το 5% του συνόλου, γεγονός που σημαίνει ότι πάνω από το 95% των αναρτήσεων καταλήγουν με 5 συγκεκριμένες και κατανοητές θεματικές λέξεις-κλειδιά.

4.3 Ομαδοποίηση Συζητήσεων σε Υπο-συζητήσεις

Στο τέλος, ομαδοποιούμε τις διαδοχικές, μέσα στο νήμα, αναρτήσεις που περιέχουν κοινές θεματικές λέξεις-κλειδιά (topic keywords) (τις οποίες βρίσκουμε με την μέθοδο που περιγράφηκε στην προηγούμενη ενότητα). Για παράδειγμα, αν μια συγκεκριμένη θεματική λέξη-κλειδί υπάρχει στις αναρτήσεις 10 έως 35, 37, 40 έως 48 και 80 έως 106, δύο ομάδες θα δημιουργηθούν και θα περιέχουν τις αναρτήσεις με αναγνωριστικούς αριθμούς από 10 έως 48 και από 80 έως 106. Να σημειωθεί ότι μια ανοχή ίση με ‘tol’ χρησιμοποιείται κατά την αναζήτηση για διαδοχικές αναρτήσεις που θα ομαδοποιηθούν (στο παραπάνω παράδειγμα το ‘tol’ είναι ίσο με 3 – το ‘tol’ είναι πάλι ανάλογο της πυκνότητας της συζήτησης). Με αυτό τον τρόπο δημιουργούμε μεγαλύτερες και πιο ουσιαστικές ομάδες. Τέλος, συγχωνεύουμε όλες τις ομάδες που επικαλύπτονται και οι νέες ομάδες που προκύπτουν αποτελούν τις υπο-συζητήσεις που εξάγονται σαν αποτέλεσμα της μεθόδου.

Στο Σχήμα 12 φαίνεται ένα παράδειγμα ενός νήματος που χωρίζεται σε 3 επιμέρους υπο-συζητήσεις. Η διακεκομμένη γραμμή αντιπροσωπεύει τις αναρτήσεις του νήματος και οι αγκύλες σηματοδοτούν τα όρια των ομάδων αναρτήσεων που μοιράζονται κοινές

λέξεις-κλειδιά. Υπο-συζητήσεις παράγονται στη συνέχεια, από τη συγχώνευση επικαλυπτόμενων συνόλων αναρτήσεων.



Σχήμα 12: Ανάλυση ενός νήματος σε υπο-συζητήσεις

5. Δημιουργία Ομάδων Χρηστών

Η δεύτερη φάση της μεθόδου που προτείνουμε για εξόρυξη απόψεων αποτελείται από τη δημιουργία ομάδων συμφωνούντων χρηστών και την αναγνώριση δεσμών διαφωνίας μεταξύ τους. Είναι σημαντικό να σημειωθεί ότι αυτό το στάδιο είναι εν πολλοίς ανεξάρτητο από τη μέθοδο εξόρυξης θεμάτων συζήτησης που παρουσιάστηκε στην προηγούμενη ενότητα. Όμως, τα αποτελέσματα αυτών των δύο μεθόδων ενώνονται στα πλαίσια του αλγορίθμου δημιουργίας περιλήψεων που θα παρουσιαστεί στην ενότητα 6.

5.1 Ανάλυση Αντιπαραθετικών Συζητήσεων

5.1.1 Βασικές Παρατηρήσεις

Η μέθοδός μας για την ανίχνευση συμφωνίας / διαφωνίας δε βασίζεται σε τεχνικές επεξεργασίας και κατανόησης φυσικής γλώσσας. Στην πραγματικότητα, το κείμενο των αναρτήσεων δεν χρησιμοποιούνται καθ 'οιονδήποτε τρόπο, κάτι που καθιστά αυτή τη μέθοδο καθολικά εφαρμόσιμη σε οποιοδήποτε χώρο διαδικτυακής συζήτησης, ανεξάρτητα της γλώσσας που χρησιμοποιείται. Αντ’ αυτού, στηρίζεται σε διάφορα τυπικά χαρακτηριστικά των συζητήσεων που αναπτύσσονται στα πλαίσια ενός διαδικτυακού χώρου συζήτησης, όπως αυτά παρατηρήθηκαν εμπειρικά μετά από εξέταση μιας ευρεία κλίμακας τέτοιων συζητήσεων. Οι πιο θεμελιώδεις από αυτές τις παρατηρήσεις είναι ότι κάθε αντιπαραθετική συζήτηση περιλαμβάνει εν γένει τα εξής στάδια:

1. Μια φάση θέρμανσης, όπου το θέμα εισάγεται και εκφράζονται οι πρώτες γνώμες. Συνήθως, οι συμμετέχοντες δεν εκφράζουν έντονη διαφωνία ακόμα, αλλά προσπαθούν να κατανοήσουν τις γενικές απόψεις των άλλων συμμετεχόντων. Αυτή η φάση είναι συνήθως μικρής διάρκειας, ιδιαίτερα σε ανεπίσημες και ομότιμες αντιπαραθετικές συνομιλίες, και μπορεί ακόμη και να απουσιάζει εξ' ολοκλήρου αν οι συμμετέχοντες γνωρίζονται μεταξύ τους καλά.
2. Μια φάση βρασμού, όπου η ανταλλαγή απόψεων είναι πιο συχνή και επικεντρωμένη. Το γενικό θέμα της συζήτησης έχει ήδη καθοριστεί και η συζήτηση δεν έχει πλέον διερευνητικό χαρακτήρα. Ο κύριος σκοπός της ανταλλαγής απόψεων γίνεται τώρα η υποστήριξη της εγκυρότητας των απόψεων που εκφράζονται, και η αντίκρουση των επιχειρημάτων της άλλης πλευράς. Είναι σημαντικό να τονιστεί ότι σε αυτή τη φάση μπορεί να δημιουργηθούν νέα υπο-θέματα για συζήτηση.

3. Μια φάση ψύξης, όπου η συζήτηση ηρεμεί, επειδή δεν υπάρχουν νέες απόψεις να συζητηθούν, ή επειδή καθίσταται σαφές ότι η συναίνεση δεν πρόκειται ποτέ να επιτευχθεί. Σε αυτό το σημείο η συζήτηση μπορεί να υποστεί είτε εκ νέου θέρμανση και να επιστρέψει σε ένα σημείο βρασμού, πχ με την εισαγωγή νέων απόψεων προς διαβούλευση, ή να λήξει εντελώς.

Για την ανακάλυψη των συμφωνιών και διαφωνιών που αναπτύσσονται σε αυτές τις συζητήσεις, επικεντρωνόμαστε κυρίως στον προσδιορισμό των διαφόρων σταδίων βρασμού της συζήτησης. Πιο συγκεκριμένα, κατά τη διάρκεια του των φάσεων βρασμού η ακόλουθη παρατήρηση είναι γενικά αληθής:

Οι διαφωνίες συνεχώς τροφοδοτούνται από την αδυναμία συναίνεσης και την ανάγκη να αντικρουστούν τα επιχειρήματα της αντίπαλης πλευράς. Αυτό σημαίνει ότι η ανταλλαγή απόψεων μεταξύ των ομάδων που διαφωνούν γίνεται με πολύ μεγαλύτερη συχνότητα από ό, τι εντός των ομάδων, όπου τα μέλη τους συμφωνούν μεταξύ τους.

5.1.2 Εντοπισμός Φάσεων Βρασμού

Για να μπορέσουμε να επωφεληθούμε από την προηγούμενη παρατήρηση, πρέπει πρώτα να προσδιοριστούν τα μέρη της συζήτησης που αντιστοιχούν σε φάσεις βρασμού. Όπως αναφέρθηκε στην προηγούμενη ενότητα, ένας αξιόπιστος τρόπος για τον εντοπισμό αυτών των φάσεων είναι η αναγνώριση της υψηλής συχνότητας της ανταλλαγής απόψεων, ιδίως σε σύγκριση με εκείνη των φάσεων θέρμανσης και της ψύξης.

Για να επιτευχθεί αυτό, θα δημιουργήσουμε πρώτα το γράφο απαντήσεων από όλη τη συζήτηση, με τη μέθοδο που αναφέρεται στο κεφάλαιο 3. Αυτός ο γράφος συνήθως δεν είναι συνδεδεμένος, αλλά αποτελείται από συνδεδεμένους υπο-γράφους, κάποιοι από τους οποίους έχουν σημαντικά μεγαλύτερη πυκνότητα συνδέσεων από τους υπόλοιπους. Αυτά τα συγκεκριμένα κομμάτια του γράφου δείχνουν συχνές και μακροσκελείς ανταλλαγές απόψεων, κάτι που δείχνει μια έντονη συζήτηση και, κατά συνέπεια, φάσεις βρασμού. Αυτά τα κομμάτια του γράφου αναγνωρίζονται χρησιμοποιώντας ένα όριο στον αριθμό των αναρτήσεων που περιέχουν (t_{bp}), και κάθε στοιχείο του γράφου που περιέχει περισσότερες αναρτήσεις από αυτό το όριο εξάγεται ως φάση βρασμού της συζήτησης.

Για την επίτευξη ανάκτησης μεγαλύτερων κομματιών της συζήτησης κατά τη φάση της δημιουργίας περίληψης και επίσης, για να υπάρξει μια ανοχή στα αναπόφευκτα λάθη κατά την αναγνώριση μη ρητών απαντήσεων, προσπαθούμε να συνενώσουμε ορισμένα

μέρη του γράφου. Πιο συγκεκριμένα, αν δύο μέρη του γράφου έχουν ένα σημαντικό ποσοστό των συζητητών τους κοινό, και η τελευταία ανάρτηση του πρώτου τμήματος είναι κοντά στην πρώτη ανάθεση του δεύτερου (t_{CPET}), τότε τα δύο αυτά μέρη εξάγονται σαν μια ενιαία φάση βρασμού της συζήτησης. Πρακτικά, αυτό σημαίνει ότι οι αναρτήσεις τους θα συνυπολογιστούν εναντίον του ορίου για την εξαγωγή φάσεων βρασμού, και εάν το συνολικό πλήθος αναρτήσεων τους ξεπερνά αυτό το όριο, τότε και οι δύο θα εξαχθούν ως μία ενιαία φάση βρασμού.

```
Για κάθε συνδεδεμένο υπογράφο cp1
    Για κάθε άλλο συνδεδεμένο υπογράφο cp2
        αν ποσοστό_ίδιων_αντικειμένων(cp1.συζητητές, cp2. συζητητές) >
        th_cpet και cp2.πρώτη_ανάρτηση.χρόνος_δημοσίευσης –
        cp1.τελευταία_ανάρτηση.χρόνος_δημοσίευσης < time-distance
            πρόσθεσε cp1 και cp2 σαν έναν ενιαίο υπογράφο στην
            component_list
            σταμάτησε την επανάληψη
    αν cp1 δεν συνενώθηκε με κανένα υπογράφο
            πρόσθεσε cp1 στην component_list
Για κάθε υπογράφο στην component-list
    θέσε posts ίσο με το πλήθος των αναρτήσεων εντός της cp
    αν posts > th_bp
        εξήγαγε cp σαν φάση βρασμού
```

Σχήμα 13: Ο αλγόριθμος που χρησιμοποιείται για την εξαγωγή φάσεων βρασμού

Στις ενότητες που ακολουθούν, παρουσιάζουμε τη μέθοδο που χρησιμοποιήθηκε για την εκτίμηση των συμφωνιών και των διαφωνιών μεταξύ των συζητητών. Για αυτό το σκοπό θα χρησιμοποιήσουμε μόνο τα τμήματα του γράφου απαντήσεων που εξήχθησαν ως φάσεις βρασμού. Κάνουμε την παραδοχή ότι οι σχέσεις συμφωνίας και διαφωνίας που θα εντοπίσουμε μέσα σε αυτές τις φάσεις θα ισχύουν σε όλη τη συζήτηση και, κατά συνέπεια, σε κάθε επιμέρους υπο-συζήτηση έχουμε εντοπίσει. Με άλλα λόγια, κάνουμε την παραδοχή ότι οι διαφωνίες που σχηματίζονται σε ένα μέρος της συζήτησης δεν αλλάζουν σε συμφωνίες σε ένα άλλο μέρος, και αντιστρόφως. Είναι σημαντικό να σημειωθεί ωστόσο, ότι η πιθανότητα να υπάρχουν σημαντικές σχέσεις συμφωνίας/διαφωνίας σε τμήματα του γράφου που δεν είναι φάσεις βρασμού είναι αμελητέα, δεδομένου ότι είναι πολύ σπάνιο τέτοιες σχέσεις να δημιουργούνται στις φάσεις θέρμανσης ή ψύξης.

5.2 Ανάλυση της Αλληλεπίδρασης Ομάδας-Συζητητή

Στην ενότητα αυτή, θα παρουσιάσουμε τα χαρακτηριστικά που χρησιμοποιούνται για τον προσδιορισμό του εάν ένας συζητητής D συμφωνεί ή διαφωνεί με την ομάδα G. Όσον αφορά την G, θεωρούμε ότι γνωρίζουμε όλους τους συζητητές που περιέχει (που συμφωνούν μεταξύ τους), καθώς και τους δεσμούς διαφωνίας της G με άλλες ομάδες συζητητών.

Για να προσδιορίσουμε εάν ο D συμφωνεί ή διαφωνεί με την G, χρησιμοποιούμε τις παρατηρήσεις που παρουσιάσθηκαν στην προηγούμενη ενότητα. Πιο συγκεκριμένα, αναλύουμε τα μέρη του γράφου απαντήσεων που εξήχθησαν ως βάσεις βρασμού και τα οποία περιέχουν αναρτήσεις μεταξύ των D και G, ώστε να υπολογίσουμε τρεις βασικές τιμές: 1) *πιθανές αλληλεπιδράσεις,* δηλαδή ένα μέτρο των ευκαιριών που είχαν οι D και G να ανταλλάξουν απόψεις, 2) *την ένταση της αλληλεπίδρασης,* δηλαδή τον αριθμό των φορών που μια πραγματική ανταλλαγή απόψεων συνέβη μεταξύ των D και G, 3) *πλήθος κοινών εχθρών,* δηλαδή τον αριθμό των φορών που οι D και G αντάλλαξαν απόψεις με ένα κοινό τρίτο συζητητή D'.

Όπως αναφέρθηκε στην προηγούμενη ενότητα, μέσα σε μια φάση βρασμού, η συχνότητα ανταλλαγής απόψεων μεταξύ των δύο συζητητών που διαφωνούν είναι πολύ υψηλότερη απ' ότι μεταξύ δύο που συμφωνούν. Η γενική ιδέα που εισάγουμε με τις μεθόδους που ακολουθούν, είναι ότι εάν ο D διαφωνεί με την G, τότε η *ένταση της αλληλεπίδρασης* μεταξύ τους θα έχει μεγάλη τιμή. Από την άλλη πλευρά, εάν συμφωνούν, η ένταση της αλληλεπίδρασης θα έχει χαμηλή τιμή, αλλά θα υπάρχει μεγάλη τιμή στο *πλήθος κοινών εχθρών*. Και στις δύο περιπτώσεις, μία υψηλή τιμή στο *πλήθος πιθανών αλληλεπιδράσεων* είναι αναγκαία, προκειμένου να διασφαλίσουμε ότι ο D και οι συζητητές στην G είχαν πολλές ευκαιρίες να ανταλλάξουν απόψεις μεταξύ τους, και είτε επέλεξαν να το πράξουν *(ένταση αλληλεπίδρασης),* είτε επέλεξαν να εστιάσουν την προσοχή τους προς άλλους, κοινούς, συζητητές *(κοινοί εχθροί).*

Για να γίνουν κατανοητά τα παραπάνω σημεία και να ενισχύσουμε την εγκυρότητά τους, παρουσιάζουμε ένα παράδειγμα που συνδυάζει τις προηγούμενες ιδέες.

Ας υποθέσουμε ότι ο D συμμετέχει σε μια συζήτηση όπου έχουν δημιουργηθεί δύο ομάδες συμφωνούντων συζητητών, οι G1 και G2, και ότι η G1 διαφωνεί με την G2. Ένας συζητητής D1 από την G1 εκφράζει τη γνώμη του και αμέσως ένας συζητητής D2 από την G2 απαντάει σε αυτόν. Ας υποθέσουμε ότι ο D απαντά στον D2. Αυτό σημαίνει ότι το D είχε μια πιθανή αλληλεπίδραση και με την G1 και με την G2, αφού

εξέφρασε την άποψή κοντά σε μια απόψεις που εκφράστηκαν τόσο από την G1 όσο και από την G2. Όμως, επέλεξε να αλληλεπιδράσει μόνο με την G2, ενώ είχε και έναν κοινό στόχο με την G1, τον D2. Εάν αυτό συμβαίνει συχνά, δηλαδή ο D να απαντάει σε συζητητές που ανήκουν στην G2, αν και θα μπορούσε να έχει επιλέξει να απαντήσει σε αναρτήσεις συζητητών της G1 οι οποίες έγιναν και αυτές σε κοντινή χρονική στιγμή και στα πλαίσια της ίδιας ακριβώς συζήτησης, τότε ο D είναι πολύ πιθανό να συμφωνήσει με την G1 και θα πρέπει να τον εντάξουμε ως μέλος της.

Στο υπόλοιπο της παρούσας ενότητας, περιγράφουμε τα προηγούμενα τρία χαρακτηριστικά με μεγαλύτερη ακρίβεια και με περισσότερες λεπτομέρειες.

Πιθανές Αλληλεπιδράσεις

Για να μετρήσουμε τις πιθανές αλληλεπιδράσεις μεταξύ των D και G χρησιμοποιήσουμε την ιδέα των θεματικών πλαισίων που παρουσιάστηκε στο κεφάλαιο 3. Το πλαίσιο μιας ανάρτησης είναι ένα σύνολο αναρτήσεων οι οποίες είναι πολύ πιθανό να συζητούν ακριβώς το ίδιο θέμα. Χρησιμοποιούμε αυτό το πλαίσιο ως ένδειξη μιας σειράς απόψεων που εκφράστηκαν κατά τη διάρκεια ενός μικρού χρονικού διαστήματος.

Πιο συγκεκριμένα, οι πιθανές αλληλεπιδράσεις ορίζονται με τον ακόλουθο αλγόριθμο:

Για κάθε ανάρτηση *P* της D

βρες όλες τις αναρτήσεις *PS* που είναι μέρος του θεματικού πλαισίου της P

συνολικό_πλήθος_αναρτήσεων_της_D++

αν *PS* περιέχει τουλάχιστον μια ανάρτηση από συζητητή που ανήκει στην G

πλήθος_πιθανών_αλληλεπιδράσεων++

$$\text{πιθανές\_αλληλεπιδράσεις}=\frac{\textit{πλήθος\_πιθανών\_αλληλεπιδράσεων}}{\textit{συνολικό\_πλήθος\_αναρτήσεων\_της\_D}}$$

Σχήμα 14: Ο αλγόριθμος που χρησιμοποιείται για τον προσδιορισμό των πιθανών αλληλεπιδράσεων μεταξύ του συζητητή D και της ομάδας G

Είναι προφανές ότι η τιμή αυτού του χαρακτηριστικού είναι μεταξύ 0 και 1.Υψηλότερες τιμές δείχνουν ότι οι D και G έχουν εκφράσει πολλές απόψεις κοντά μεταξύ τους και πάνω στο ίδιο θέμα, και επομένως είναι πολύ καλοί υποψήφιοι για την εξαγωγή μιας σχέσης συμφωνίας ή διαφωνίας. Από την άλλη πλευρά, χαμηλότερες τιμές δείχνουν ότι ο D δεν ήταν ενεργός όταν συζητητές του G εξέφρασαν τις απόψεις τους, και κατά συνέπεια δεν θα ήταν ασφαλές να εξαχθούν σχέσεις μεταξύ τους.

Ένταση Αλληλεπίδρασης

Για τη μέτρηση της *έντασης της αλληλεπίδρασης* μεταξύ D και G χρησιμοποιούμε το γράφο απαντήσεων για να εντοπίσουμε αναρτήσεις του D που συνδέονται άμεσα με αναρτήσεις συζητητών που ανήκουν στην G. Αυτές οι αναρτήσεις θα είναι τότε απαντήσεις, ρητά ή όχι, η μία στην άλλη, οπότε οι δύο συζητητές ανταλλάσσουν άμεσα απόψεις μεταξύ τους.

Πιο συγκεκριμένα, η *ένταση της αλληλεπίδρασης* καθορίζεται χρησιμοποιώντας τον ακόλουθο αλγόριθμο:

Για κάθε ανάρτηση *P* της D

αν P συνδέεται με ακμή στον πίνακα απαντήσεων με ανάρτηση P2 από συζητητή που ανήκει στην G

αλληλεπιδράσεις++

συνολικό_πλήθος_αναρτήσεων_της_D ++

$$\text{ένταση\_αλληλεπίδρασης}=\frac{\textit{αλληλεπιδράσεις}}{\textit{συνολικό\_πλήθος\_αναρτήσεων\_της\_D}}$$

Σχήμα 15: Ο αλγόριθμος που χρησιμοποιείται για τον προσδιορισμό της έντασης της αλληλεπίδρασης μεταξύ του συζητητή D και της ομάδας G

Οι τιμές κοντά στο 1 σημαίνουν ότι ο D έχει αφιερώσει μεγάλο μέρος της συμβολής του στη συζήτηση σε ανταλλαγή απόψεων με την G. Ως εκ τούτου, είναι πολύ πιθανό να διαφωνούν. Από την άλλη πλευρά, οι τιμές που προσεγγίζουν 0 σημαίνουν ότι ο D έχει εστιάσει την προσοχή στην αλληλεπίδραση με άλλους συζητητές, κάτι που αποτελεί μια καλή ένδειξη συμφωνίας, ιδίως εάν αυτές οι τρίτοι συζητητές είναι τα κοινοί για τους D και G.

Κοινοί Εχθροί

Για τη μέτρηση των κοινών εχθρών των D και G θα πρέπει να αξιοποιήσουμε τους δεσμούς διαφωνίας μεταξύ της G και των άλλων ομάδων. Αυτό μας δίνει τη δυνατότητα να προσδιορίσουμε όλους τους συζητητές που διαφωνούν με την G, και μπορούμε να υπολογίσουμε στη συνέχεια τις αλληλεπιδράσεις του D με αυτές.

Πιο συγκεκριμένα, το πλήθος *κοινών εχθρών* προσδιορίζεται με τον ακόλουθο αλγόριθμο:

Για κάθε δεσμό διαφωνίας L της G

έστω G` η ομάδα που συνδέεται με την G μέσω του L

πρόσθεσε όλους τους συζητητές της G` στην *enemies-list*

Για κάθε συζητητή D` στην enemies-list

αν υπάρχει ανάρτηση P από τον D η οποία συνδέεται στο γράφο απαντήσεων με ανάρτηση P` από τον D`

αλληλεπιδράσεις++

$$\text{κοινοί\_εχθροί} = \frac{\textit{αλληλεπιδράσεις}}{\textit{συνολικό\_πλήθος\_αναρτήσεων\_της\_D}}$$

Σχήμα 16: Ο αλγόριθμος που χρησιμοποιείται για τον προσδιορισμό της τιμής του χαρακτηριστικού των κοινών εχθρών μεταξύ συζητητή D και ομάδας G

Υψηλές τιμές σημαίνουν ότι ο D έχει ανταλλάξει πολλές απόψεις με συζητητές με τους οποίους η G διαφωνεί. Αυτό αποτελεί μια πολύ καλή ένδειξη της συμφωνίας μεταξύ D και G. Από την άλλη πλευρά, οι τιμές που προσεγγίζουν το 0 είναι μια ένδειξη πιθανής διαφωνίας, ιδίως εάν η ένταση της αλληλεπίδρασης μεταξύ των D και G είναι υψηλή. Σημειώστε ότι σε ορισμένες, αρκετά σπάνιες, περιπτώσεις αυτό το χαρακτηριστικό μπορεί να έχει τιμές που υπερβαίνουν το 1.Όταν συμβεί αυτό, θα μετατρέψουμε τις τιμές αυτές σε 1, προκειμένου να έχουμε μια κοινή κλίμακα για όλα τα χαρακτηριστικά.

Από την αξία των τριών αυτών χαρακτηριστικών δημιουργείται ένα διάνυσμα, το οποίο στη συνέχεια κανονικοποιείται ώστε να παράσχει μεγαλύτερη ανοχή κατά τη διαδικασία ανίχνευσης συμφωνιών/διαφωνιών. Πιο συγκεκριμένα, κάθε τιμή αντιστοιχίζεται σε ένα σύνολο αριθμών {1,2,3,4,5}, που κυμαίνεται από εξαιρετικά ανεπαρκής (1) ως εξαιρετικά επαρκής (5). Ο Πίνακας 1 παρουσιάζει τη σημασιολογία της κάθε κανονικοποιημένης τιμής ανά χαρακτηριστικό.

Πίνακας 1. Η σημασιολογία των κανονικοποιημένων τιμών ανά χαρακτηριστικό

Τιμή	Σημασιολογία ανά Χαρακτηριστικών
1	Πιθανές Αλληλεπιδράσεις: Πολύ λίγες ευκαιρίες αλληλεπίδρασης, δεν μπορούν να εξαχθούν σχέσεις με ασφάλεια
	Ένταση Αλληλεπίδρασης: Πολύ λίγες αλληλεπιδράσεις, πολύ καλή ένδειξη συμφωνίας
	Κοινοί εχθροί: Πολύ λίγοι κοινοί εχθροί, πολύ καλή ένδειξη διαφωνίας
2	Λίγες ευκαιρίες αλληλεπίδρασης, μπορεί να χρησιμοποιηθεί για την εξαγωγή σχέσεων μόνο αν τα δύο άλλα χαρακτηριστικά έχουν πολύ ευνοϊκές τιμές
	Λίγες αλληλεπιδράσεις, καλή ένδειξη διαφωνίας
	Λίγοι κοινοί εχθροί, καλή ένδειξη διαφωνίας
3	Ικανοποιητικό πλήθος πιθανών αλληλεπιδράσεων, οι σχέσεις μπορούν να εξαχθούν με ασφάλεια
	Ένα ουδέτερο πλήθος αλληλεπιδράσεων, ανεπαρκές για διαφωνία, αλλά και αρκετά μεγαλύτερο απ' ότι θα μπορούσε να υποστηρίξει μια συμφωνία
	Ένα ουδέτερο πλήθος κοινών εχθρών, ανεπαρκές για συμφωνία, αλλά και αρκετά μεγαλύτερο απ' ότι θα μπορούσε να υποστηρίξει μια διαφωνία
4	Καλό πλήθος πιθανών αλληλεπιδράσεων, οι σχέσεις μπορούν να εξαχθούν με ιδιαίτερη ασφάλεια
	Πολλές αλληλεπιδράσεις, καλή ένδειξη διαφωνίας
	Πολύ κοινοί εχθροί, καλή ένδειξη συμφωνίας

5	Πάρα πολλές πιθανές αλληλεπιδράσεις, οι σχέσεις μπορούν να εξαχθούν με μεγάλη βεβαιότητα
	Πάρα πολλές αλληλεπιδράσεις, πολύ καλή ένδειξη διαφωνίας
	Πάρα πολύ μεγάλο πλήθος κοινών εχθρών, πολύ καλή ένδειξη συμφωνίας

Αφότου το διάνυσμα έχει κανονικοποιηθεί, προσδιορίζουμε δύο τιμές βεβαιότητας για το κατά πόσο το διάνυσμα αυτό υποδηλώνει συμφωνία ή διαφωνία χρησιμοποιώντας τον ακόλουθο απλό αλγόριθμο:

Βεβαιότητα Συμφωνίας	**Βεβαιότητα Διαφωνίας**
αν πιθανές_αλληλεπιδράσεις<2 **επίστρεψε** -1 **επίστρεψε** πιθανές_αλληλεπιδράσεις + (6 – ένταση_αλληλεπίδρασης) + κοινοί_εχθροί	**αν** πιθανές_αλληλεπιδράσεις<2 **επίστρεψε** -1 **επίστρεψε** πιθανές_αλληλεπιδράσεις + ένταση_αλληλεπίδρασης + (6 – κοινοί_εχθροί)

Εάν για μία από αυτές τις τιμές η βεβαιότητα είναι πάνω από ένα χειροκίνητα προσδιοριζόμενο όριο (t_{conf}), τότε εξάγεται η αντίστοιχη σχέση μεταξύ του τρέχοντος συζητητή και της ομάδας. Σε αυτό το σημείο πρέπει να σημειωθεί ότι η ελάχιστη τιμή για το όριο αυτό είναι ίση με 12, έτσι ώστε να είναι αδύνατο να εξαχθεί ταυτόχρονα συμφωνία και διαφωνία μεταξύ των D και G.

5.3 Αλγόριθμος Δημιουργίας Ομάδων

Στην ενότητα αυτή, παρουσιάζουμε τον αλγόριθμο για τη δημιουργία των ομάδων συζητητών και σύνδεσή τους με σχέσεις συμφωνίας και διαφωνίας. Μέσα στην ίδια ομάδα, οι συζητητές έχουν παρόμοιες απόψεις πάνω στο θέμα της συζήτησης, ενώ οι συζητητές σε ομάδες που συνδέονται με έναν σύνδεσμο διαφωνίας εκφράζουν διαφορετικές απόψεις.

Αρχικά, σε ένα στάδιο προεπεξεργασίας, αφαιρούμε όλους τους συζητητές που έχουν αριθμό αναρτήσεων μικρότερο από ένα όριο (t_{SNP}), αφού η συνεισφορά τους στη συζήτηση είναι ασήμαντη.

Έπειτα, αναζητούμε τον συζητητή που έχει το μεγαλύτερο άθροισμα αναρτήσεων και απαντήσεων στις αναρτήσεις του. Μια αρχική ομάδα δημιουργείται και αυτός ο χρήστης εντάσσεται σε αυτή. Έπειτα συνεχίζουμε με τον κεντρικό αλγόριθμο για τη δημιουργία ομάδων και σχέσεων μεταξύ τους όπως φαίνεται παρακάτω:

```
Για κάθε συζητητή D που δεν ανήκει σε καμία ομάδα
    Για κάθε ομάδα G ήδη δημιουργημένη
      rel_sum = (αναρτήσεις της D που απαντούν σε αναρτήσεις
    της G) + (αναρτήσεις της G που απαντούν σε αναρτήσεις του
    D)
έστω D ο συζητητής με το μεγαλύτερο rel_sum
δημιούργησε ομάδα GD και πρόσθεσε τον D σε αυτή
```

```
Για κάθε ήδη υπάρχουσα ομάδα G εκτός της GD
    δημιούργησε το διάνυσμα χαρακτηριστικών μεταξύ D και G
    αν βεβαίοτητα_συμφωνίας(διάνυσμα χαρακτηριστικών) > th_rsex
      δημιούργησε σχέση συμφωνίας μεταξύ GD και D
    αν βεβαίοτητα_διαφωνίας(διάνυσμα χαρακτηριστικών) > th_rsex
      δημιούργησε σχέση διαφωνίας μεταξύ GD και D
```

Σχήμα 17: Ο αλγόριθμος δημιουργίας ομάδων και σχέσεων μεταξύ τους

Σημειώστε ότι για κάθε σχέση συμφωνίας και διαφωνίας αποθηκεύουμε έναν αριθμό που υποδηλώνει τη δύναμή της. Αυτή η δύναμη είναι ίση με τη βεβαιότητα του διανύσματος χαρακτηριστικών που χρησιμοποιήθηκαν για την εξαγωγή αυτού του συνδέσμου.

Ο αλγόριθμος που φαίνεται στο Σχήμα 16 εκτελείται επαναληπτικά έως ότου όλοι οι συζητητές έχουν υποστεί επεξεργασία. Στο τέλος κάθε επανάληψης, λαμβάνει χώρα ένα βήμα συγχώνευσης που μας δίνει τη δυνατότητα να χρησιμοποιηθούν όσες σχέσεις συμφωνίας ανακαλύφθηκαν για τη συγχώνευση της ομάδας που μόλις δημιουργήθηκε *(GD)* με άλλες ομάδες που δημιουργήθηκαν από προηγούμενες επαναλήψεις του αλγορίθμου. Αυτό μας επιτρέπει να δημιουργήσουμε μεγαλύτερες ομάδες συμφωνούντων συζητητών, κάτι πολύ χρήσιμο για τη δημιουργία των περιλήψεων που θα ακολουθήσει.

Πριν από την παρουσίαση του αλγορίθμου συγχώνευσης, πρέπει να ορίσουμε το πότε η συγχώνευση δύο ομάδων είναι ασυνεπής:

Η συγχώνευση δύο ομάδων είναι ασυνεπής αν και μόνο αν η ομάδα που θα προκύψει από τη διαδικασία συγχώνευσης οδηγεί σε σχέση διαφωνίας προς την ίδια την ομάδα, ή σε σχέσεις συμφωνίας και διαφωνίας προς μια τρίτη ομάδα.

Τώρα είμαστε έτοιμοι για την παρουσίαση του αλγορίθμου συγχώνευσης. Αξίζει να σημειωθεί ότι η πρώτη ομάδα που εξετάζετε για συγχώνευση θα είναι η *GD,* δηλαδή η ομάδα που δημιουργήθηκε κατά την τελευταία επανάληψη του αλγορίθμου δημιουργίας ομάδων. Σε κάθε επόμενο βήμα, η ομάδα που ελέγχεται για δυνατότητα συγχώνευσης θα είναι αυτή που η GD συγχωνεύθηκε μαζί της, και ούτω καθεξής, μέχρι ότου δεν υπάρξουν άλλες συγχωνεύσεις.

```
έστω GL η ομάδα που δημιουργήθηκε πιο πρόσφατα

Για κάθε δεσμό συμφωνίας L της GL σε φθίνουσα σειρά δύναμης
    έστω GM η ομάδα που συνδέεται με την GL μέσω του L
    αν συνεπής_συνένωση(GL, GM)
      δημιούργησε νέα ομάδα G
      πρόσθεσε όλους τους συζητητές των GL και GM     στην G
      πρόσθεσε όλους τους δεσμούς των GL και GM στην G
      για τους δεσμούς με κοινές ομάδες κράτησε αυτόν με τη
      μεγαλύτερη δύναμη
```

```
      αφαίρεσε της GL και GM
      σταμάτησε την επανάληψη
αν συνέβη μια συγχώνευση στην προηγούμενη επανάληψη εκτέλεσε
και πάλι τον αλγόριθμο συγχώνευσης, με GL = G
```

Σχήμα 18: Ο αλγόριθμος συγχώνευσης ομάδων

Στο τέλος, υπάρχει ένα στάδιο μετα-επεξεργασίας με στόχο τη συγχώνευση μικρών ομάδων με μεγαλύτερες. Με αυτό τον τρόπο επιτυγχάνεται η βελτίωση των παραγόμενων περιλήψεων μέσω της αποφυγής περιφερειακών απόψεων, που θα μπορούσαν να διεκδικήσουν δυσανάλογο χώρο στην περίληψη. Για το σκοπό αυτό ορίζουμε δύο είδη ομάδων: 1) Μεγάλη ομάδα, η οποία είναι μια ομάδα με συνολικό πλήθος αναρτήσεων μεγαλύτερο από ένα ορισμένο όριο (t_{lgpt}), 2) μικρή ομάδα, η οποία είναι μια ομάδα που το μέγιστο πλήθος των αναρτήσεών της που απαντούν σε αναρτήσεις από συζητητές μιας άλλης ομάδας, είναι μικρότερο από ένα συγκεκριμένο όριο (t_{SGPT}).

Το στάδιο της μετα-επεξεργασίας αποτελείται από τη συγχώνευση μικρών ομάδων με μεγαλύτερες, ακόμη και αν δεν έχουν σχέση συμφωνίας μεταξύ τους, αρκεί όμως να μην έχουν ούτε σχέση διαφωνίας. Μια μικρή ομάδα συγχωνεύετε με μια μεγάλη ομάδα αν οι σχέσεις της είναι υποσύνολο των σχέσεων της μεγάλης ομάδας. Με αυτό τον τρόπο διατηρούμε τις σχέσεις της μεγάλης ομάδας ανέπαφες, ενώ ταυτόχρονα και η μικρή ομάδα κρατιέται από το να έχει σημαντική επίπτωση στην τελική περίληψη.

6. Εξαγωγή Αναρτήσεων για τη Δημιουργία Περίληψης

Οι μέθοδοι που έχουν παρουσιαστεί στα κεφάλαια 4 και 5 συνδυάζονται για να δημιουργήσουν μια περίληψη των απόψεων που εκφράζονται σε ένα νήμα συζήτησης. Οι δημιουργούμενες περιλήψεις αποτελούνται από ένα μικρό ποσοστό των αναρτήσεων του νήματος συζήτησης, επιλεγμένες με τέτοιο τρόπο ώστε να μεγιστοποιηθεί η κάλυψη των θεμάτων και των επιμέρους υπο-θεμάτων της συζήτησης, και να αναδειχθούν όλες οι διαφορετικές απόψεις των συμμετεχόντων.

Ως στάδιο προεπεξεργασίας, ανακτούμε το σύνολο των αναρτήσεων, χωρισμένες ανά υπο-συζήτηση, όπως περιγράφτηκε στην παράγραφο 4. Στη συνέχεια, αυτές οι αναρτήσεις ταξινομούνται κατά φθίνουσα σειρά του πλήθους των χρήσιμων φράσεων-λέξεων (λέξεις-κλειδιά) που περιέχουν, μιας και αναρτήσεις με περισσότερες λέξεις-κλειδιά είναι πιθανότερο να περιγράφουν το θέμα της συζήτησης πιο αποτελεσματικά από τυχαίες αναρτήσεις.

Ο αλγόριθμος που χρησιμοποιείται για να δημιουργήσει την περίληψη έχει ως εξής:

```
Για κάθε υποσυζητήση SD
    Για κάθε ζεύγος ομάδων G1 και G2 που συνδέονται με σχέση διαφωνίας
      Για κάθε ανάρτηση P της SD σε φθίνουσα σειρά πλήθους αναρτήσεων
            αν P αναρτήθηκε από συζητητή της G1
                  GG = G1, GS = G2
            αλλιώς αν P αναρτήθηκε από συζητητή της G2
                  GG = G2, GS = G1
            αλλιώς συνέχισε με την επόμενη ανάρτηση P
            για κάθε ανάρτηση P2 της SD
                  αν P2 απαντά σε P και έγινε από συζητητή της GS
                        για κάθε ανάρτηση P3 της SD
                              αν P3 απαντά σε P2 και έγινε από συζητητή της GG
                                    πρόσθεσε (P, P2, P3) στην περίληψη
                                    σταμάτησε και τα 3 επίπεδα επανάληψης και συνέχισε
                                     με το επόμενο ζεύγος ομάδων G1, G2
```

Σχήμα 19: Ο αλγόριθμος δημιουργίας περίληψης

Οι πλειάδες που συλλέγονται μέσω του προηγούμενου αλγορίθμου, παρουσιάζονται με τη μορφή γράφου, έτσι ώστε ο τελικός χρήστης να μπορεί να εντοπίσει εύκολα τις συνδέσεις μεταξύ των εμφανιζόμενων αναρτήσεων. Να σημειωθεί ότι μια πλειάδα που αποτελείται από τρεις αναρτήσεις ουσιαστικά παρουσιάζει τη γνώμη της ομάδας, την απάντηση μιας ομάδας που διαφωνεί με την προηγούμενη γνώμη και, τέλος, την δευτερολογία της αρχικής ομάδας. Αυτό το βάθος είναι συνήθως αρκετό για να παρουσιαστεί αποτελεσματικά το σημείο τριβής μεταξύ των δύο ομάδων και οι απόψεις των δύο πλευρών.

Κάθε περίληψη υποστηρίζεται και από πληροφορίες μεταδεδομένων που μπορούν να χρησιμοποιηθούν από τους παραδοσιακούς μηχανές αναζήτησης για την ανάκτηση συγκεκριμένων περιλήψεων.

7. Πειραματική Αξιολόγηση

Για τους σκοπούς της αξιολόγησης των μεθόδων που παρουσιάσαμε, δημιουργήσαμε μια πλατφόρμα αξιολόγησης όπου άνθρωποι αξιολογητές μπορούσαν να διαβάσουν και να βαθμολογήσουν περιλήψεις που δημιουργήθηκαν με τη χρήση αλγορίθμων μας.

Στην ενότητα που ακολουθεί, παρουσιάζουμε τη λίστα με τις ακριβείς τιμές που χρησιμοποιήθηκαν στην πλατφόρμα αξιολόγησης για τα διάφορα όρια που παρουσιάστηκαν σε προηγούμενα τμήματα της πτυχιακής. Μετά από αυτό, παρουσιάζουμε τη διαδικασία αξιολόγησης λεπτομερώς και αναλύουμε τα αποτελέσματά της.

7.1 Λεπτομέρειες Υλοποίησης

Οι ακόλουθες τιμές επιλέχθηκαν μετά από εκτενή πειραματισμό και αξιολόγηση της ποιότητας των δημιουργούμενων ομάδων και περιλήψεων. Είναι σημαντικό να τονιστεί ότι τα νήματα συζήτησης που χρησιμοποιήθηκαν για το σκοπό αυτό είναι διαφορετικά από εκείνα που χρησιμοποιήθηκαν για την αξιολόγηση.

Πίνακας 2. Οι οριακές τιμές, όπως χρησιμοποιήθηκαν στην πλατφόρμα αξιολόγησης

$th_{bp} = 30$	$th_{cpet} = 3$	$th_{snp} = 3$	$th_{sgpt} = 10$
th_{lgpt} = μέσο πλήθος αναρτήσεων ανά ομάδα			
$th_{conf = 12}$		$t_{ol} = 3$	

Ο πίνακας 3 περιέχει τις ακριβείς αντιστοιχίες μεταξύ των χαρακτηριστικών γνωρισμάτων και των κανονικοποιημένων τιμών σε {1,2,3,4,5}. Αυτές οι αντιστοιχίσεις είναι οι ίδιες για κάθε χαρακτηριστικό. Έτσι, για παράδειγμα, αν η *ένταση της αλληλεπίδρασης* μεταξύ του συζητητή D και της ομάδας G είναι 15%, τότε η κανονικοποιημένη τιμή, σύμφωνα με τον παρακάτω πίνακα, θα είναι 2. Σε αυτό το σημείο αξίζει να τονιστεί ξανά ότι μόνο οι κανονικοποιημένες τιμές χρησιμοποιούνται για τον υπολογισμό του βαθμού βεβαιότητας συμφωνίας και διαφωνίας (ενότητα 5.2).

Πίνακας 3. Αντιστοιχήσεις Χαρακτηριστικών

1	2	3	4	5
0%-9%	10%-20%	21%-32%	33%-50%	51%+

7.2 Αποτελέσματα Αξιολόγησης

Για την αξιολόγηση χρησιμοποιήθηκαν 4 νήματα συζήτησης που είχαν επιλεγεί τυχαία από ένα μεγαλύτερο σύνολο από 30 νήματα συζήτησης που προέρχονται από ένα γνωστό φόρουμ του Διαδικτύου. Κάθε νήμα συζήτησης περιείχε κατά μέσο όρο 175 αναρτήσεις, και συνολικά χρησιμοποιήθηκαν περίπου 700 αναρτήσεις.

Μετά την εφαρμογή αλγορίθμων μας, κάθε νήμα συζήτησης χωρίστηκε, κατά μέσο όρο, σε 2 μικρότερες υπο-συζητήσεις. Επίσης, για κάθε νήμα, δημιουργήθηκε ένας μικρός αριθμός (2 έως 4) μεγάλων ομάδων από συζητητές. Δημιουργήθηκε επίσης, ένα σύνολο από μικρότερες ομάδες, με λιγότερους συζητητές και αναρτήσεις, αλλά οι ομάδες αυτές δεν είχαν μεγάλο αντίκτυπο στις τελικές περιλήψεις.

Μετά τον εντοπισμό των επιμέρους συζητήσεων και ομάδων, δημιουργήθηκαν περιλήψεις για κάθε νήμα, χρησιμοποιώντας τον αλγόριθμο που περιγράφτηκε στο κεφάλαιο 6. Κατά μέσο όρο, παρήχθησαν 4,5 πλειάδες (P, P2, P3) για κάθε νήμα συζήτησης. Δεδομένου ότι κάθε πλειάδα περιέχει 3 αναρτήσεις, οι συνολικές περιλήψεις για κάθε νήμα περιείχαν 13,5 αναρτήσεις. Αυτό είναι περίπου το 8% του μεγέθους ενός νήματος.

Κάθε περίληψη αξιολογήθηκε από 7 αξιολογητές οι οποίοι κλήθηκαν να βαθμολογήσουν την ποιότητα των περιλήψεων και την ακρίβεια των ομάδων που δημιουργήθηκαν και των σχέσεων μεταξύ τους. Το σύστημα αξιολόγησης περιελάμβανε 5 επίπεδα βαθμολογίας, από 0 έως και 5, με το 0 είναι ο χειρότερος βαθμός. Δεδομένου ότι δεν δόθηκε επιλογή για ουδέτερη βαθμολόγηση, οι τιμές 2,1,0 είναι αυξανόμενα αρνητικές και από οι τιμές 3,4,5 είναι αυξανόμενα θετικές. Η αξιολόγηση οργανώθηκε ως εξής:

Αρχικά, ζητήθηκε από τους αξιολογητές να περιηγηθούν γρήγορα στις αναρτήσεις των επιμέρους υπο-συζητήσεων που επιλέχθηκαν για τη δημιουργία περιλήψεων. Ο χρόνος που δόθηκε για το έργο αυτό ήταν μικρός, έτσι ώστε να υπήρχε μια ευκαιρία να εξοικειωθούν με το θέμα, αλλά όχι τόσο πολύ ώστε να είναι προ-κατειλημμένοι και να συμπληρώσουν κενά στις περιλήψεις με ό, τι είχαν διαβάσει. Μετά από αυτό, παρουσιάστηκαν οι περιλήψεις που παρήχθησαν για την τρέχουσα υπο-συζήτηση και οι αξιολογητές κλήθηκαν να τις μελετήσουν προσεκτικά και στη συνέχεια να τις βαθμολογήσουν στις ακόλουθες κατηγορίες:

1. Ποιότητα των περιλήψεων. Αν ήταν σε θέση να κατανοήσουν τις απόψεις και τα επιχειρήματα των ομάδων συζητητών και ένιωθαν ικανοποιημένοι από την κάλυψη του θέματος, τους ζητήθηκε να δώσουν ένα θετικό βαθμό. Αν είχαν την αίσθηση ότι οι περιλήψεις ήταν δύσκολο να κατανοηθούν και περιείχαν πληροφορίες που ήταν είτε άσχετες με το θέμα είτε ελλιπείς, τους ζητήθηκε να δώσουν έναν αρνητικό βαθμό.
2. Η ακρίβεια του αλγορίθμου δημιουργίας ομάδων. Για κάθε πλειάδα (P, P2, P3) ζητήθηκε να προσδιοριστεί εάν η P2 έδειχνε πράγματι διαφωνία και εξέφραζε

διαφορετική άποψη από αυτό που εκφραζόταν στην P, και αν η P3 διαφωνούσε με την P2 και εξέφρασε μία παρόμοια άποψη με την P.

Τα αποτελέσματα της αξιολόγησης (Πίνακας 4), υπολογίστηκαν με την αφαίρεση τόσο της υψηλότερης όσο και της χαμηλότερης βαθμολογίας από το κάθε σύνολο των αξιολογήσεων. Παρουσιάζονται επίσης, τα αποτελέσματα χωρίς αυτήν την τροποποίηση, και φαίνονται στον ίδιο πίνακα σε παρένθεση. Είναι σημαντικό να παρατηρηθεί ότι αυτές οι τιμές είναι πολύ κοντά μεταξύ τους, κάτι που σημαίνει ότι η αξιολόγηση ήταν συνεπής και δεν υπήρχαν μεγάλες αποκλίσεις στις βαθμολογίες. Όπως μπορούμε να δούμε, οι βαθμοί και για τις δύο κατηγορίες αξιολόγησης είναι στη θετική κλίμακα.

Πίνακας 4. Τα αποτελέσματα των αξιολογήσεων

Ποιότητα Περίληψης	Ακρίβεια Δημιουργούμενων Ομάδων
3,34 (3,4)	**3,77** (3,8)

Εκτός από αυτές τις βαθμολογίες, ζητήσαμε επίσης από τους αξιολογητές να μας προσφέρουν σχόλια υπό τη μορφή ελεύθερου κειμένου. Μια κοινή παρατήρηση όλων των αξιολογητών ήταν ότι βρήκαν το κείμενο των επιμέρους αναρτήσεων που περιέχονταν στις περιλήψεις να είναι κάπως μεγάλο, αν και οι περιλήψεις οι ίδιες ήταν ένα πολύ μικρό ποσοστό του αρχικού μεγέθους του νήματος. Αυτή είναι μια καλή ένδειξη ότι πρέπει να εργαστούμε στην ανάπτυξη μεθόδων που δεν κάνουν εξόρυξη ολόκληρων αναρτήσεων, αλλά μόνο συγκεκριμένων προτάσεων μέσα τους που περιέχουν τις πιο πλήρεις πληροφορίες για το θέμα.

Παρ 'όλα αυτά, οι περισσότεροι αξιολογητές επίσης ανέφεραν ότι οι επιλεχθείσες αναρτήσεις ήταν πολύ διαφωτιστικές για το θέμα της συζήτησης και ότι οι διαφορετικές απόψεις των συζητητών ήταν ιδιαίτερα σαφείς.

Συνολικά, είμαστε πολύ ευχαριστημένοι με τα αποτελέσματα της αξιολόγησης και είμαστε αισιόδοξοι για την επιτυχία των προτεινόμενων μεθόδων μας.

8. Μελλοντικές Κατευθύνσεις

Στην τρέχουσα εργασία μας επικεντρωθήκαμε στην εξόρυξη του ενός μικρού ποσοστού αναρτήσεων ώστε να επιτευχθεί μια αποτελεσματική περίληψη των διαφορετικών απόψεων που παρουσιάζονται μέσα σε ένα νήμα συζήτησης. Ο επόμενος στόχος μας είναι να προσδιορίσουμε, αξιοποιώντας τις αναρτήσεις που ήδη έχουμε εξορύξει ως μέρος των περιλήψεων, τα διάφορα είδη των συναισθημάτων που εκφράζονται από τους χρήστες σχετικά με τα διάφορα θέματα της συζήτησης. Σκοπεύουμε να δημιουργήσουμε μια εκτεταμένη οντολογία των συναισθημάτων αυτών και να την υποστηρίξουμε με κανόνες συμπερασμού. Οι κανόνες αυτοί μπορούν έπειτα να χρησιμοποιηθούν για τη δημιουργία νέων δεσμών συναισθημάτων μεταξύ των ομάδων, των χρηστών και των θεμάτων, όπως αυτά εξάγονται από την τρέχουσα δουλειά μας.

Η επιτυχία σε αυτή την προσπάθεια μπορεί να μας επιτρέψει να δημιουργούμε περιλήψεις χωρίς τη χρήση λέξεων-κλειδιών, αλλά με τη χρήση σημασιολογικών κανόνων για την αναγνώριση της σημασίας των ερωτημάτων που τίθενται στη βάση δεδομένων μας. Η τρέχουσα έρευνα στις τεχνολογίες σημασιολογικού ιστού μπορεί να λειτουργήσει ως ένας πολύτιμος οδηγός στις προσπάθειές μας.

9. ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Web Forum / Web Discussion Board	Φόρουμ / Διαδικτυακός Τόπος Συζήτησης
Post	Ανάρτηση
Thread	Νήμα
Thread Extraction	Ανάλυση Ροής Συζήτησης
Topic Extraction	Εξαγωγή Θέματος
Topic Keyword	Θεματική Λέξη-Κλειδί
Post Thematic Context	Θεματικό Πλαίσιο Ανάρτησης
Information Extraction (IR)	Εξαγωγή Πληροφορίας

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

SIGMOD	Special Interest Group on Management Of Data

ΑΝΑΦΟΡΕΣ

[1] Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. 2004. Identifying agreement and disagreement in conversational speech: use of Bayesian networks to model pragmatic dependencies. In Proceedings of the 42nd Annual Meeting on Association For Computational Linguistics.

[2] Germesin, S. and Wilson, T. 2009. Agreement detection in multiparty conversation. In Proceedings of the 2009 international Conference on Multimodal interfaces.

[3] Hahn, S., Ladner, R., and Ostendorf, M. 2006. Agreement/disagreement classification: exploiting unlabeled data using contrast classifiers. In Proceedings of the Human Language Technology Conference of the Naacl, Companion Volume: Short Papers.

[4] Hillard, D., Ostendorf, M., and Shriberg, E. 2003. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In Proceedings of the 2003 Conference of the North American Chapter of the Association For Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003--Short Papers - Volume 2.

[5] Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining.

[6] Ku, L.-W., Liang, Y.-T. and Chen, H.-H. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW).

[7] Zhuang, L., Jing, F., and Zhu, X. 2006. Movie review mining and summarization. In Proceedings of the 15th ACM international Conference on information and Knowledge Management.

[8] Seki, Y. 2008. Summarization Focusing on Polarity or Opinion Fragments in Blogs. In Proceedings of the First Text Analysis Conference (TAC 2008).

[9] Balahur, A., Lloret, E., Boldrini, E., Montoyo, A., Palomar, M., and Martínez-Barco, P. 2009. Summarizing threads in blogs using opinion polarity. In Proceedings of the Workshop on Events in Emerging Text Types.

[10] Kim, H. and Zhai, C. 2009. Generating comparative summaries of contradictory opinions in text. In Proceeding of the 18th ACM Conference on information and Knowledge Management.

[11] Kim, J., Candan, S. and Donderler, M. 2005. Topic Segmentation of Message Hierarchies for Indexing and Navigation Support. In Proceedings of the14th International Conference on World Wide Web (WWW2005).

[12] Adams, P. and Martell, C. 2008. Topic Detection and Extraction in Chat. In Proceedings of the Second International Conference on Semantic Computing (ICSC08).

[13] Huang, J., Zhou, M. and Yang, D. 2007. Extracting Chatbot Knowledge from Online Discussion Forums. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI07).

[14] Wang, Y. C., Joshi, M., Cohen, W. W. and Rosé, C. P. 2008. Recovering Implicit Thread Structure in Newsgroup Style Conversations. In Proceedings of the 2nd International Conference on Weblogs and Social Media (ICWSM II).

[15] Labadié, A. and Prince, V. 2008. Intended Boundaries Detection in Topic Change Tracking for Text Segmentation. In Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS08).

[16] Alchemy API, http://www.alchemyapi.com