ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΟΙΚΤΑ ΔΙΑΣΥΝΔΕΔΕΜΕΝΑ ΔΕΔΟΜΕΝΑ ΥΠΗΡΕΣΙΩΝ ΚΑΙ ΠΡΟΪΟΝΤΩΝ ΑΠΟ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ, ΣΕ ΠΕΡΙΒΑΛΛΟΝΤΑ ΗΛΕΚΤΡΟΝΙΚΟΥ ΕΜΠΟΡΙΟΥ

Διπλωματική Εργασία

της

Καράμπελα Αναστασίας

Θεσσαλονίκη, Νοέμβριος 2021

ΑΝΟΙΚΤΑ ΔΙΑΣΥΝΔΕΔΕΜΕΝΑ ΔΕΔΟΜΕΝΑ ΥΠΗΡΕΣΙΩΝ ΚΑΙ ΠΡΟΪΟΝΤΩΝ ΑΠΟ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ, ΣΕ ΠΕΡΙΒΑΛΛΟΝΤΑ ΗΛΕΚΤΡΟΝΙΚΟΥ ΕΜΠΟΡΙΟΥ

Καράμπελα Αναστασία

Πτυχίο Εφαρμοσμένης Πληροφορικής 2011

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπων Καθηγητής
Γεωργιάδης Χρήστος

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 05/11/2021

Γεωργιάδης Χρήστος	Βλαχοπούλου Μάρω	Χατζηγεωργίου Αλέξανδρος
.....	.....	.....

Καράμπελα Αναστασία

.....

Περίληψη

Η παρούσα έρευνα παρουσιάζει τον κόσμο των ανοιχτών διασυνδεδεμένων δεδομένων και πως αναπαρίστανται σε αυτόν δεδομένα που προέρχονται από τα social media. Σκοπός της εργασίας είναι η εξοικείωση του αναγνώστη με την καινοτόμα τεχνολογία των ανοιχτών διασυνδεδεμένων δεδομένων και η παρουσίαση των δυνατοτήτων που αυτά προσφέρουν μέσα από εφαρμογές τους (π.χ. μηχανές αναζήτησης, recommendation systems κ.α) στον σύγχρονο ψηφιακό κόσμο για την ανάπτυξη και βελτίωση ενός ηλεκτρονικού καταστήματος. Γίνεται αναφορά σε οντολογίες που χρησιμοποιούνται για την αναπαράσταση περιεχομένου των κοινωνικών δικτύων. Γλώσσες με τις οποίες μπορούμε να συντάξουμε ερωτήματα και να εξάγουμε δεδομένα που μας ενδιαφέρουν. Καθώς και εργαλεία οπτικοποίησης των αποτελεσμάτων των εφαρμογών αυτών στον κόσμο του διαδικτύου. Επίσης, αναλύεται η σημασία και η επιρροή του περιεχομένου των social media στην καταναλωτική συμπεριφορά των χρηστών και πως αυτό μπορεί να χρησιμοποιηθεί για τη βελτίωση ενός ηλεκτρονικού καταστήματος. Τέλος, παρουσιάζεται στην πράξη η οντολογία eClassOWL, η οποία χρησιμοποιείται στην αναπαράσταση δεδομένων του κλάδου του ecommerce, μέσα από παραδείγματα και εφαρμογές ερωτημάτων.

Λέξεις Κλειδιά: ανοιχτά διασυνδεδεμένα δεδομένα, σημασιολογικός ιστός, κοινωνικά δίκτυα, ηλεκτρονικό εμπόριο

Abstract

The present research presents the world of link open data and how social media data are represented in it. The aim of this paper is to acquaint the reader with the innovative technology of linked open data, to present the possibilities that they offer through their applications (eg search engines, recommendation systems, etc.) to the modern digital world and how they can help to develop and improve an online store. Reference is made to ontologies used to represent the content of social networks, languages for writing queries and extracting data. As well as tools for visualizing the results of these applications. It also analyzes the importance and influence of social media content on consumer behavior and how it can be used to improve an online store. Finally, the eClassOWL ontology is presented, an ontology used to represent data from the ecommerce industry, through examples and query applications.

Keywords: linked open data, sematic web, social media, ecommerce

Περιεχόμενα

1 Εισαγωγή

1.1 Πρόβλημα - Σημαντικότητα του θέματος

Το Web 2.0 διαφέρει πολύ από τον προκάτοχό του Web 1.0. Προάγει την δημιουργία διαφόρων ειδών περιεχομένου από απλούς χρήστες, την ανάπτυξη διαδικτυακών σχέσεων και τη δημιουργία διαδικτυακών κοινοτήτων στις οποίες οι χρήστες αλληλεπιδρούν, συνεργάζονται και μοιράζονται περιεχόμενο και πληροφορίες. Ένας μεγάλος αριθμός social media platforms έχει δημιουργηθεί βασιζόμενος σε αυτά τα χαρακτηριστικά του. Σύμφωνα με τις βασικές δυνατότητες του Web 2.0 τα social media περιλαμβάνουν:

1. Περιεχόμενο παραγόμενο από τους χρήστες: οι χρήστες καταχωρούν περιεχόμενο το οποίο είναι διαθέσιμο σε άλλους χρήστες για ανάγνωση, σχολιασμό και βαθμολόγηση.
2. Κοινωνική δικτύωση: οι χρήστες των social media συμμετέχουν σε online κοινότητες, οι οποίες τους επιτρέπουν να βλέπουν πληροφορίες των προφίλ των χρηστών με τους οποίους συνδέονται, να μοιράζουν δεδομένα και να αλληλεπιδρούν μεταξύ τους.
3. Συνεργασία: οι χρήστες συμμετέχουν σε συζητήσεις, συν-δημιουργούν περιεχόμενο, συνεργάζονται για επίλυση προβλημάτων και προχωρούν σε ενέργειες συλλογής δεδομένων. [18]

Κάθε μέρα παράγονται 2,5 πεντάκις εκατομμύρια δεδομένα με τον ρυθμό παραγωγής τους να αυξάνεται συνεχώς με την εξέλιξη του Διαδικτύου των Δεδομένων (IoT – Internet of Thinks). Είναι χαρακτηριστικό ότι τα τελευταία 2 χρόνια παράχθηκε το 90% των δεδομένων που έχουν παραχθεί παγκοσμίως καθώς πλέον 3,7 δισεκατομμύρια άνθρωποι χρησιμοποιούν το διαδίκτυο, αριθμός που αυξήθηκε κατά 7,5% από το 2016. [21] Για τον αποδοτική αξιοποίησή των δεδομένων αυτών είναι απαραίτητη η ανάγκη αποθήκευσης και διάθεσης τους με τρόπο εύκολα προσβάσιμο και κατανοητό από ανθρώπους και μηχανές. Ο σημασιολογικός ιστός σημαίνει για τους χρήστες συνδέσμους μεταξύ εννοιών και όχι απαραίτητα ιστοσελίδων. Σε αντίθεση με την παραδοσιακή μορφή του διαδικτύου η οποία απαιτεί γνώση της φυσικής γλώσσας, η κατανόηση του σημασιολογικού ιστού απαιτεί παρακολούθηση και κατανόηση των συνδέσμων και των λογικών κανόνων που τους διέπουν. [5]

Μια ακόμα διάσταση του διαδικτύου με την οποία θα ασχοληθούμε στην παρούσα εργασία είναι τα Social Media. Τα Social Media αποτελούν τη μεγαλύτερη συλλογή δεδομένων της κοινωνίας μας που υπήρξε ποτέ προσφέρωντας μια πλούσια πηγή στοιχείων για την ανθρώπινη συμπεριφορά. Ωστόσο, η κατανόηση και η ουσιαστική χρήση τους αποτελεί μεγάλο πρόβλημα. Η άντληση της σωστής πληροφορίας συχνά γίνεται δύσκολη λόγω των εργαλείων ανάλυσης τα οποία ακόμα βρίσκονται σε πρώιμο στάδιο και παράγουν ανακριβή αποτελέσματα τα οποία είναι δύσκολο να ερμηνεύουν σωστά. Χρειαζόμαστε επομένως περισσότερο εξελιγμένες μορφές

ανάλυσης προκειμένου να κατανοήσουμε το περιεχόμενο που παράγουν οι χρήστες του διαδικτύου. [13]

1.2 Σκοπός Στόχοι

Σκοπός της παρούσας εργασίας είναι να εισάγει τους αναγνώστες σε κάποιες βασικές έννοιες του σημασιολογικού ιστού ώστε να κατανοήσουν την χρησιμότητα του και τις απεριόριστες δυνατότητες τις οποίες μπορεί να προσφέρει στον κόσμο του διαδικτύου και ειδικότερα στον τομέα του ecommerce. Η τεχνολογία των διασυνδεδεμένων δεδομένων παρότι βρίσκεται ακόμη σε αρχικό στάδιο, χρησιμοποιείται από ένα πλήθος επιχειρήσεων με ιδιαίτερα ισχυρή παρουσία στο ηλεκτρονικό εμπόριο. Ταυτόχρονα, τα social media είναι γνωστό σε όλους μας ότι πρωταγωνιστούν στον κλάδο του διαδικτυακού marketing και αποτελούν αναπόσπαστο κομμάτι των εργασιών ενός ηλεκτρονικού καταστήματος. Στόχος μου, μέσα από την παρούσα εργασία είναι να παρουσιάσω πως αυτές οι δύο τεχνολογίες μπορούν να συνδυαστούν και συνεισφέρουν στην εξέλιξη και κερδοφορία ενός ηλεκτρονικού καταστήματος.

1.3 Βασική Ορολογία (Σημασιολογικός Ιστός, Ανοιχτά διασυνδεδεμένα δεδομένα)

Big Data: Ο όρος Big Data χρησιμοποιείται για να περιγράψει σύνολα δεδομένων τόσο μεγάλα ή σύνθετα που ξεφεύγουν από τις δυνατότητες καταγραφής, αποθήκευσης και ανάλυσης των παραδοσιακών τεχνικών επεξεργασίας δεδομένων.
Linked Open Data: Ο όρος Διασυνδεδεμένα δεδομένα περιλαμβάνει ένα σύνολο αρχών σχεδιασμού για την κοινή χρήση διασυνδεδεμένων δεδομένων που διαβάζονται από μηχανές στον Ιστό. Όταν συνδυάζονται με τα Ανοικτά Δεδομένα, ονομάζονται Ανοιχτά Διασυνδεδεμένα Δεδομένα (Linked Open Data - LOD).
Linked Open Datasets (LODs): Τα LODs αποτελούνται από δημοσιευμένες ανοιχτές βάσεις δεδομένων σε μορφή RDF (Resource Description Framework) και συνδέσμους RDF για την σύνδεση διαφορετικών πηγών δεδομένων. [10]
Semantic Annotation: Ο σημασιολογικός σχολιασμός ή η προσθήκη ετικετών είναι η διαδικασία προσάρτησης σε ένα έγγραφο κειμένου ή άλλου μη δομημένου περιεχομένου, μεταδεδομένων σχετικά με διάφορες έννοιες (π.χ. άτομα, μέρη, οργανισμούς, προϊόντα ή θέματα) που σχετίζονται με αυτό.

1.4 Διάρθρωση της εργασίας

Στο δεύτερο κεφάλαιο παρουσιάζονται οι Τεχνολογίες του σημασιολογικού ιστού, οι βασικές έννοιες οι οποίες τις διέπουν (οντολογίες, γλώσσες, LODs) και οι βασικές εφαρμογές τους στο διαδίκτυο τους όπως οι μηχανές αναζήτησης, recommendation systems και data mining. Επίσης, γίνεται αναφορά στα δεδομένα του κλάδου του ecommerce, πως αυτά αναπαρίστανται στον κόσμο των ανοιχτών διασυνδεδεμένων δεδομένων και πως μπορούν να επηρεάσουν την online καταναλωτική συμπεριφορά των χρηστών. Στο τρίτο κεφάλαιο ακολουθεί η παρουσίαση της οντολογίας eClassOWL στον κλάδο του ecommerce μαζί με παραδείγματα και εφαρμογές ερωτημάτων sparql για την εξόρυξη χρήσιμων δεδομένων. Στο τέταρτο κεφάλαιο ακολουθεί ο επίλογος, στο πέμπτο κεφάλαιο περιλαμβάνεται μια σύνοψη της εργασίας μαζί με τα συμπεράσματα στα οποία κατέληξα. Στο έκτο κεφάλαιο ακολουθούν τα όρια και οι περιορισμού της έρευνας. Τέλος, στο έβδομο κεφάλαιο αναφέρονται οι μελλοντικές επεκτάσεις.

2 Βιβλιογραφική επισκόπηση

2.1 Τεχνολογίες του σημασιολογικού ιστού

Οι τεχνολογίες σημασιολογικού ιστού είναι μια προσπάθεια διαχείρισης την πληροφορίας που υπάρχει στο διαδίκτυο χρησιμοποιώντας οντότητες και πρότυπα του σημασιολογικού ιστού. [5]

Ο T.Berners-Lee ήταν ο πρώτος που εισήγαγε τον όρο Διασυνδεδεμένα Δεδομένα και όρισε τις αρχές για τη δημοσίευση και σύνδεση δομημένων δεδομένων στο διαδίκτυο. Από τεχνικής πλευράς, τα διασυνδεδεμένα δεδομένα αξιοποιούν το πρωτόκολλο HTTP και την τεχνολογία RDF για τη δημοσίευση και διασύνδεση δομημένων δεδομένων που προέρχονται από διαφορετικές πηγές στο διαδίκτυο. Αποτελούν ένα κομμάτι του σημασιολογικού ιστού καθώς πολλά από τα χαρακτηριστικά του σημασιολογικού ιστού υπάρχουν και στα διασυνδεδεμένα δεδομένα. Ωστόσο, η αναζήτηση, βελτίωση και η ταξινόμηση τους αποτελούν προκλήσεις για τις μηχανές αναζήτησης ανοικτών διασυνδεδεμένων δεδομένων. [19]

Το σημαντικότερο χαρακτηριστικό των LODs είναι ότι οι συνδέσεις ανάμεσα στα δεδομένα διέπονται από κανόνες του σημασιολογικού ιστού. Μέσω αυτών των κανόνων οι σύνδεσμοι μεταξύ των γράφων RDF αντιπροσωπεύουν τους δεσμούς ανάμεσα στα συνδεδεμένα δεδομένα τα οποία μπορούν να προσδιοριστούν και να ανακτηθούν χρησιμοποιώντας συγκεκριμένα ερωτήματα διατυπωμένα σε κατάλληλες γλώσσες ερωτημάτων, όπως η SPARQL. [12]

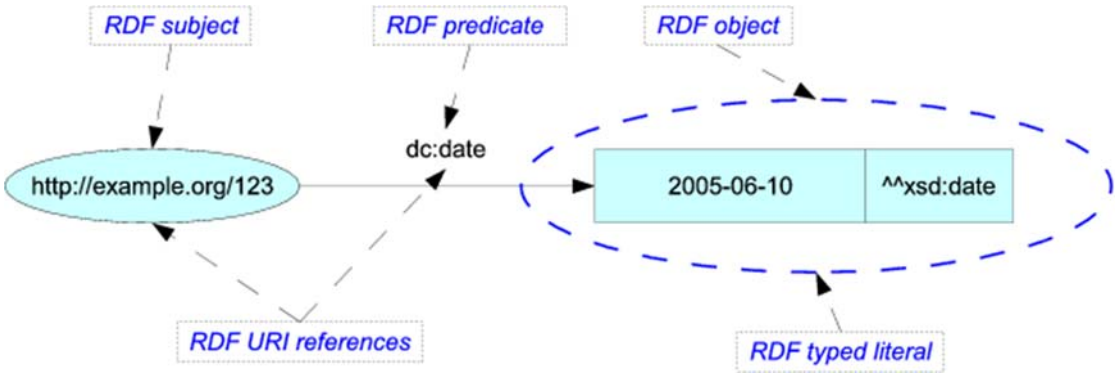
Τα RDF περιέχουν αποθηκευμένες περιγραφές αντικειμένων οι οποίες αφορούν λειτουργικά και μη-λειτουργικά χαρακτηριστικά τους και μπορούν να διεξαχθούν μέσα από απαντήσεις σε ερωτήματα SPARQL. Τα αντικείμενα αυτά σχηματίζουν γράφους RDF μοντελοποιώντας τις μεταξύ τους σχέσεις. Οι απαντήσεις στα ερωτήματα SPARQL μπορούν να είναι σε πολλές μορφές όπως XML, JSON, CSV, TSV. [12]

2.2 Γλώσσες του σημασιολογικού ιστού: RDF, OWL, XML, HTML

Ο σημασιολογικός ιστός εισήγαγε γλώσσες ειδικά σχεδιασμένες για την επεξεργασία των Διασυνδεδεμένων δεδομένων, ακολουθούν οι σημαντικότερες από αυτές:

- RDF (Resource Description Framework): Τα διασυνδεδεμένα δεδομένα εξαρτώνται από έγγραφα που περιέχουν δεδομένα σε μορφή RDF. Το μοντέλο δεδομένων RDF περιλαμβάνει δεδομένα τα οποία βασίζονται σε γράφους και χρησιμοποιούνται για την δημοσίευση δεδομένων στο διαδίκτυο. Οι δηλώσεις των πόρων έχουν την μορφή υποκείμενο - κατηγόρημα - αντικείμενο. Το υποκείμενο υποδηλώνει τον πόρο, το αντικείμενο υποδηλώνει την αξία ενός χαρακτηριστικού του πόρου και το κατηγόρημα υποδηλώνει τη σχέση μεταξύ

υποκειμένου και αντικειμένου. Π.χ. στην φράση “Το αυτοκίνητο έχει χρώμα μαύρο”, το “αυτοκίνητο” είναι το υποκείμενο, το “μαύρο” το αντικείμενο και το “έχει χρώμα” είναι το κατηγόρημα. Το μοντέλο RDF αναπαριστά δεδομένα σε μορφή τριπλετών (υποκείμενο - κατηγόρημα - αντικείμενο). Ο προσδιορισμός των πόρων γίνεται με τη χρήση URIs (Uniform Resource Identifiers). Το υποκείμενο θα πρέπει να είναι ένα αναγνωριστικό URI που αναπαριστά έναν πόρο. Το αντικείμενο μπορεί να είναι ένα URI ή και ένα απλό string (συμβολοσειρά). Το κατηγόρημα επίσης θα πρέπει να είναι URI καθώς αναπαριστά τη σχέση ανάμεσα στο υποκείμενο και το κατηγόρημα.



Εικόνα 1 : Παράδειγμα URI – τριπλέτα (υποκείμενο - κατηγόρημα - αντικείμενο)

- Ontology: Η οντολογία ορίζει την σχέση η οποία επιτρέπει την ενσωμάτωση των δεδομένων. Χρησιμοποιείται για να ορίσει και να περιγράψει μια συγκεκριμένη περιοχή ενδιαφέροντος. Είναι μια συλλογή από URIs με συγκεκριμένη έννοια και περιεχόμενο. Επίσης, αναφέρεται και ως ένα έγγραφο RDFs. Ο κύριος ρόλος μιας οντολογίας είναι ταξινόμηση των αντικειμένων με βάση την σημασία τους. Αυτό πραγματοποιείται με την περιγραφή των αντικειμένων (Individuals), των κλάσεων (Classes), των χαρακτηριστικών τους (Attributes) και των μεταξύ τους σχέσεων (Relations). Οι κλάσεις είναι οι συλλογές των διαφόρων θεμάτων, τα χαρακτηριστικά είναι οι ιδιότητες των αντικειμένων και οι σχέσεις είναι οι τρόποι με τους οποίους σχετίζονται τα αντικείμενα με τις κλάσεις.
- OWL (Web Ontology Language):
- XML (Extensible Markup Language):
- HTML (Hypertext Markup Language):

Οι RDF, OWL και XML περιγράφουν αυθαίρετες έννοιες/ αντικείμενα όπως π.χ. ανθρώπους, meetings ή μέρη του σκάφους ενός αεροπλάνου.

2.3 Schema.org κοινότητα που προσφέρει πρότυπο λεξικό απόδοσης αναγνωστικών στον σημασιολογικό ιστό και microdata

Η κοινότητα Schema.org προσφέρει ένα πρότυπο λεξικό για την υιοθέτηση των αρχών του σημασιολογικού ιστού. Χρησιμοποιείται από έναν ευρύ αριθμό εταιρειών

και οργανώσεων προκειμένου να καλύψει τις επιχειρηματικές τους ανάγκες. Η κοινότητα ιδρύθηκε από τις εταιρείες Google, Microsoft, Yahoo και Yandex. Η διαμόρφωση του λεξιλογίου αποτελεί διαδικασία η οποία αναπτύσσεται από μια ανοιχτή κοινότητα με τη χρήση της mailing list. [23]
Οι κοινότητα Shema.org έχει σαν αποστολή της να δημιουργεί, να συντηρεί και να προωθεί σχήματα για τα δομημένα δεδομένα του διαδικτύου, όπως αυτά που υπάρχουν σε ιστοσελίδες, σε μηνύματα ηλεκτρονικού ταχυδρομείου κ.α. Το λεξιλόγιο αυτό μπορεί να χρησιμοποιηθεί με διάφορες κωδικοποιήσεις όπως οι RDFa, Microdata και JSON-LD. Αναπαριστά τις οντότητες, τις σχέσεις μεταξύ των οντοτήτων και των τις ενέργειες τους και μπορεί εύκολα να επεκταθεί με τη χρήση ενός καλά τεκμηριωμένου μοντέλου επέκτασης. [23]
Πάνω από 10 εκατομμύρια ιστότοποι χρησιμοποιούν το μοντέλο Schema.org καθώς και πολλές εφαρμογές της Google και της Microsoft καθώς και οι εφαρμογές Pinterest και Yandex ήδη χρησιμοποιούν το λεξιλόγιο αυτό προκειμένου να προσφέρουν στο κοινό τους πλούσιες και επεκτάσιμες εμπειρίες χρήσης. [23]
Οι μηχανισμοί αναζήτησης προϊόντων/υπηρεσιών σε συνδυασμό με τη σημασιολογική βελτιστοποίηση μιας ιστοσελίδας (semantic SEO) με την χρήση του προτύπου Schema.org, αποτελούν εργαλεία για τους προγραμματιστές ώστε να ενσωματώσουν δομημένα δεδομένα στις σελίδες τους και να επωφεληθούν από την βελτίωση της παρουσίασης τους στα αποτελέσματα των μηχανών αναζήτησης. [5]

Συγκεκριμένα, τα microdata, τα οποία αναφέρθηκαν παραπάνω χρησιμοποιούνται ως μέσο για τη βελτίωση της αποτελεσματικότητας των μηχανών αναζήτησης. Πιο συγκεκριμένα, η βάση Schema.org περιλαμβάνει ένα λεξικό microdata τα οποία μπορούν να εμπλουτίσουν το περιεχόμενο και την πληροφορία που περιέχει μια ιστοσελίδα βοηθώντας στην βελτίωση της ευρετηρίασης των ιστοσελίδων στις μηχανές αναζήτησης και κατά επέκταση την παραγωγή καλύτερων αποτελεσμάτων. Με βάση το λεξικό Scgema.org οι ιστοσελίδες "μαρκάρονται" με microdata, όπως τα γνωστά tags της γλώσσας HTML5. Τα microdata βοηθούν:

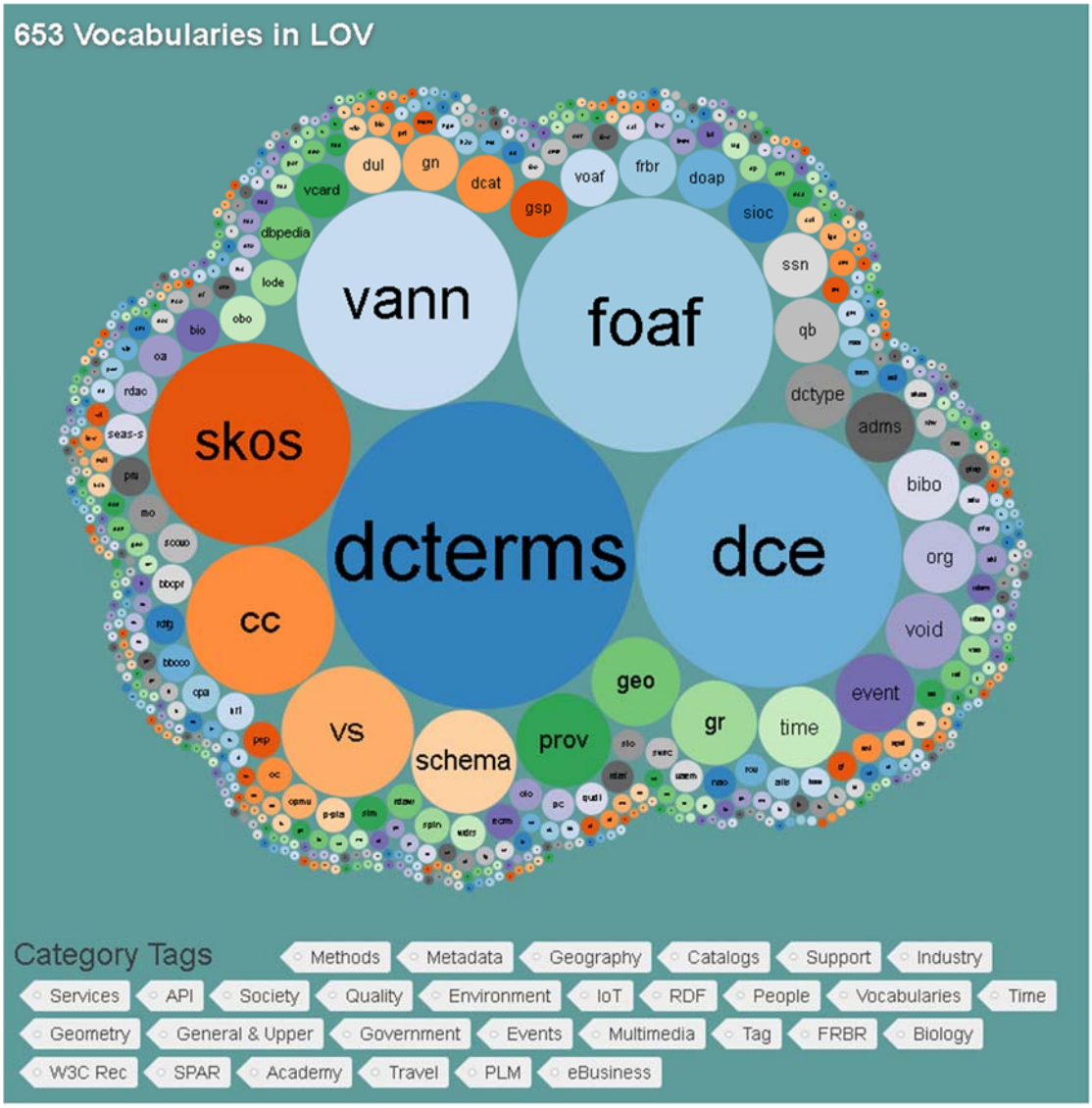
1. Τις μηχανές αναζήτησης και άλλες εφαρμογές να κατανοήσουν καλύτερα τα περιεχόμενα ενός website.
2. Τους χρήστες ώστε να βρίσκουν πιο εύκολα και με μεγαλύτερη ακρίβεια πληροφορίες σχετικά με όσα τους ενδιαφέρουν, εκμεταλλευόμενοι τις επιπλέον πληροφορίες που προστίθενται στον κώδικα HTML. [10]

Η χρήση των microdata έχει καθιερωθεί μεταξύ άλλων μορφών ως αυτή που χρησιμοποιείται για την ενσωμάτωση δομημένων δεδομένων σε μια ιστοσελίδα καθώς αποτελεί τον προτιμώμενο τρόπο σύνταξης σύμφωνα με το πρότυπο Schema.org. [10]

Τα Ανοιχτά Διασυνδεδεμένα Δεδομένα αποτελούν ακόμη έναν τρόπο με τον οποίο το περιεχόμενο ενός ιστοτόπου μπορεί να γίνει κατανοητό από τις μηχανές αναζήτησης. Τα LODs αποτελούνται από δημοσιευμένες ανοιχτές βάσεις δεδομένων σε μορφή RDF (Resource Description Framework) και συνδέσμους RDF για την σύνδεση διαφορετικών πηγών δεδομένων. Η δημοσίευση αυτών των δεδομένων ακολουθεί τις

αρχές που διέπουν τα Διασυνδεδεμένα Δεδομένα. Τα Διασυνδεδεμένα Δεδομένα παρ' ότι αποτελούν έναν τρόπο επέκτασης του περιεχομένου των ιστοσελίδων δεν θα πρέπει να συγχέονται με τα microdata καθώς διέπονται από διαφορετικές αρχές. Η προσπάθεια δημιουργίας συνδέσμων ανάμεσα στις δύο αυτές δομές δίνει τη δυνατότητα στους χρήστες να εξάγουν ανοιχτά διασυνδεδεμένα δεδομένα από ιστοσελίδες οι οποίες έχουν ήδη αντιστοιχιστεί με όρους του λεξιλογίου του Schema.org. Για να πραγματοποιηθεί αυτό χρειάζεται αντιστοίχιση ανάμεσα στις κλάσεις και στις ιδιότητες του Schema.org με το λεξιλόγια που χρησιμοποιούν τα Ανοιχτά Διασυνδεδεμένα Δεδομένα. [10]

Στα Ανοιχτά Διασυνδεδεμένα Δεδομένα δεν υπάρχουν υποχρεωτικά λεξιλόγια ωστόσο έχουν υιοθετηθεί κάποια τα οποία έχουν γίνει δημοφιλή από τις κοινότητες που τα χρησιμοποιούν. Προκειμένου να αποφασίσουμε ποιο λεξιλόγιο θα χρησιμοποιήσουμε θα πρέπει να λάβουμε υπόψη τον βαθμό χρήσης του κάθε λεξιλογίου. Οι ποσοτικές πληροφορίες σχετικά με την χρήση των κλάσεων και των ιδιοτήτων που περιλαμβάνουν τα λεξιλόγια αυτά είναι διαθέσιμες στο Linked Open Vocabulary (LOV) (https://lov.linkeddata.es/dataset/lov). Το LOV είναι συλλογή λεξιλογίων από διαφορετικά πεδία, όπως το ecommerce, και σκοπός του είναι η παροχή πρόσβασης στα λεξιλόγια αυτά, ο προσδιορισμός των μεταξύ τους σχέσεων και εξαρτήσεων καθώς και το πως αυτά συνδέονται στο Linked Data Cloud. Επίσης, συλλέγει αναλυτικές πληροφορίες για τις οντολογίες οι οποίες αντιπροσωπεύουν τα λεξιλόγια αυτά καθώς και στατιστικά στοιχεία που σχετίζονται με τα ανοιχτά διασυνδεδεμένα δεδομένα ή γραφικές σχέσεις ανάμεσα στα λεξιλόγια. [10]



Εικόνα 2 : Η συλλογή λεξιλογίων LOV (Linked Open Vocabulary)

Το 2011 η κοινότητα Schema.org ξεκίνησε να παράγει τις οντολογίες της σε μορφή OWL5 (Web Ontology Language). Το λεξιλόγιο καλύπτει πολλές περιοχές ενδιαφέροντος, ωστόσο μπορούμε να τις διαχωρίσουμε σε δύο κυρίως μέρη. Το πρώτο μέρος περιλαμβάνει ένα μικρό σύνολο στοιχείων τα οποία περιγράφουν πρωτεύοντα είδη δεδομένων, όπως αριθμοί και κείμενο. Στο κομμάτι αυτό μπορούν να βρεθούν κλάσεις όπως οι Boolean, Date ή Number. Οι υπόλοιπες κλάσεις και ιδιότητες χρησιμοποιούνται για την περιγραφή πιο πολυσύνθετων πεδίων όπως οι Οργανισμοί ή οντότητες που σχετίζονται με την Ιατρικη, τα ΜΜΕ κ.α. [10]

2.4 Η οντολογία Goodrelations

Η GoodRelations αποτελεί την ισχυρότερη οντολογία για την αναπαράσταση και δημοσιοποίηση πληροφοριών στο κλάδο του ecommerce. Οι πληροφορίες αυτές αφορούν κυρίως προϊόντα, προσφορές, εκπτώσεις, τιμές ή όρους χρήσης των αγορών

στο διαδίκτυο. Με λίγα λόγια, αποτελεί το ισχυρότερο εννοιολογικό μοντέλο για την απεικόνιση πληροφοριών στο ηλεκτρονικό εμπόριο. Μπορεί να χρησιμοποιηθεί σε όλες της γλώσσες που υποστηρίζουν σύνταξη προτάσεων RDF (όπως RDF/XML, Turtle, RDFa, JSON-LD κ.α.), Microdata και σε οποιαδήποτε σύνταξη υποστηρίζει το μοντέλο υποκείμενο-κατηγόρημα-αντικείμενο, γενικότερα. [24]
Ξεκίνησε ως μια ανεξάρτητη οντολογία το 2007, ενώ από τον Νοέμβριο του 2012 ενσωματώθηκε πλήρως στο λεξικό schema.org και αποτελεί πλέον το επίσημο μοντέλο αναπαράστασης δεδομένων ηλεκτρονικού εμπορίου. Έχει παραμείνει ένα ανεξάρτητο project και έχει διατηρήσει την επίσημη έκδοση του μοντέλου αναπαράστασης. Ενώ οι πλειοψηφία των δεδομένων του μοντέλου GoodRelations που δημοσιεύεται στο διαδίκτυο είναι με τη μορφή του προτύπου schema.org, η πρωτότυπη μορφή αποτελεί τον προκαθορισμένο τρόπο χειρισμού των δεδομένων αυτών στο πλήρες περιβάλλον των γλωσσών RDF/SPARQL/OWL. [24]
Η οντολογία GoodRelations παρέχει ένα βασικό λεξιλόγιο για να εκφράσει πληροφορίες όπως π.χ. η περιγράφη μιας προσφοράς ενός Web site σε κάποια τηλέφωνα μιας συγκεκριμένης εταιρείας ή η διαμόρφωση ενός μοντέλου τιμολόγησης σε ένα ηλεκτρονικό κατάστημα που πουλάει μουσικά όργανα για τα πιάνα που ζυγίζουν λιγότερο από 150 κιλά ή ακόμα και την περιγραφή του μοντέλου μιας εταιρείας ενοικίασης αυτοκινήτων που μισθώνει αυτοκίνητα μιας συγκεκριμένης μάρκας και μοντέλου από ένα συγκεκριμένο σύνολο υποκαταστημάτων σε ολόκληρη τη χώρα. Με λίγα λόγια κάθε λεπτομέρεια εμπορικής ή λειτουργικής φύσεων του κλάδου του ecommerce μπορεί να εκφραστεί με τη χρήστη της οντολογίας αυτής, όπως π.χ. οι επιλογές πληρωμής και παράδοσης, ποσοτικές εκπτώσεις, ώρες λειτουργίας κ.α. Τέλος, αξίζει να αναφέρουμε ότι η χρήση του μοντέλου GoodRelations βελτιώνει την απόδοση των μηχανών αναζήτησης των Google, Yahoo, Bing και Yandex, και αντίστοιχα υποστηρίζεται και χρησιμοποιείται από τις παραπάνω μέσω του schema.org. [24]

2.5 Γλώσσα ανάκτησης πληροφοριών από LODs: SPARQL

Η γλώσσα ερωτημάτων SPARQL χρησιμοποιείται στην ανάκτηση πληροφοριών από δομημένα και ημηδομημένα δεδομένα. Μας δίνει τη δυνατότητα να εξερευνήσουμε τα δεδομένα αυτά μέσα από την εκτέλεση ερωτημάτων σε σχέσεις που μέχρι τώρα μας ήταν άγνωστες. Για να το πετύχει αυτό πραγματοποιεί σύνθετες συνδέσεις μεταξύ διαφορετικών βάσεων δεδομένων με τη χρήση ενός απλού ερωτήματος.

Τα ερωτήματα SPARQL εκτελούνται σε σύνολα δεδομένων RDF αποτελούμενα από RDF γράφους. Ένα SPARQL endpoint δέχεται ερωτήματα και επιστρέφει αποτελέσματα μέσω του πρωτοκόλλου HTTP. Υπάρχουν “γενικά” endpoints που μπορούν να πραγματοποιήσουν αναζητήσεις σε όλα τα RDF δεδομένα που υπάρχουν διαθέσιμα αυτή τη στιγμή στο διαδίκτυο. Όπως επίσης, υπάρχουν “συγκεκριμένα” endpoints τα οποία συνδέονται απευθείας με συγκεκριμένα σύνολα δεδομένων. Τα αποτελέσματα των SPARQL ερωτημάτων μπορούν να επιστραφούν ή να

μετατραπούν σε διάφορες μορφές αρχείων όπως οι XLM, JSON, RDF και HTML. Η γλώσσα SPARQL έχει ορίσει ένα συγκεκριμένο λεξιλόγιο XML που χρησιμοποιείται για την επιστροφή πινάκων αποτελεσμάτων. Τέλος, υπάρχει η δυνατότητα αποθήκευσης του λεξιλογίου αυτού σε μορφή JSON για χρήση σε διαδικτυακές εφαρμογές. Ορισμένα SPARQL αποτελέσματα επιστρέφουν απαντήσεις σε μορφή RDF, οι οποίες με τη σειρά τους μπορούν να κωδικοποιηθούν σε μορφές RDF/XML, N-Triples, Turtle κ.α.). Τέλος, υπάρχει και η μορφή HTML η οποία χρησιμοποιείται σε ερωτήματα SPARQL που ορίζονται μέσω διαδραστικών φορμών. Συχνά υλοποιείται με την εφαρμογή μετασχηματισμού από XSL σε XML format.

Στην παρούσα εργασία θα πραγματοποιήσω μια σύντομη παρουσίαση των βασικών αρχών τις γλώσσας αυτής με χρήση απλών παραδειγμάτων. Η βασική δομή της γλώσσας περιλαμβάνει:

- Δεδομένα RDF σε μορφή τριπλετών "υποκείμενο-κατηγόρημα-αντικείμενο"
- Οι πόροι αυτοί αναπαριστώνται σε μορφή URIs
- Το "αντικείμενο" μιας τριπλέτας εκτός από URI μπορεί να είναι ακόμα και μια σταθερή τιμή (string, integer, boolean κ.α.)
- Prefixes (δηλώσεις προθεμάτων) για συντομογραφία των URIs π.χ. PREFIX foo: <http://example.com/resources/>
- Ορισμός του συνόλου δεδομένων με τη δήλωση του γράφου RDF στον οποίον θέτουμε το ερώτημα
  FROM ...
- Μια ρήτρα αποτελέσματος η οποία προσδιορίζει την πληροφορία που επιθυμούμε να εμφανίσουμε από τα αποτελέσματα του ερωτήματος
  SELECT …
- Το ερώτημα το οποίο θέτουμε στη βάση δεδομένων
  WHERE …
- "Τροποποιητές" των ερωτημάτων για κατάτμηση, ταξινόμηση ή αλλαγή της διάταξης των αποτελεσμάτων
  π.χ. ORDER BY ...

Μετά τη σύντομη περιγραφή των βασικών στοιχείων της γλώσσας SPARQL παραθέτω ορισμένα απλά ερωτήματα ώστε να κατανοήσουμε την χρήση της στην πράξη:

```
PREFIX foaf:  <http://xmlns.com/foaf/0.1/>
SELECT ?name
WHERE {
?person foaf:name ?name .
}
```

Το παραπάνω ερώτημα βρίσκει όλα τα υποκείμενα (ανθρώπους-person) και αντικείμενα (ονόματα-names) τα οποία συνδέονται με την ιδιότητα foaf:name και

επιστρέφει όλες τις τιμές της μεταβλητής ?name. Εφαρμόζοντας το ερώτημα αυτό στον γράφο http://www.w3.org/People/Berners-Lee/card βρίσκουμε όλα τα ονόματα ανθρώπων που υπάρχουν καταχωρημένοι στον γράφο FOAF του Tim Berners-Lee. Όπως θα δούμε και αργότερα η FOAF είναι μια από τις σημαντικότερες οντολογίες η οποία περιγράφει ανθρώπους και τις μεταξύ τους σχέσεις.

Οι μεταβλητές στην γλώσσα SPARQL συμβολίζονται με ένα ? και μπορούν να ταιριάξουν με κάθε κόμβο του γράφου RDF. Η εντολή SELECT επιστρέφει έναν πίνακα μεταβλητών με τις τιμές οι οποίες ικανοποιούν τη συνθήκη του ερωτήματος. Η συνθήκη του ερωτήματος μπορεί να αποτελείται από μια ή και παραπάνω τριπλέτες, όπως το παράδειγμα που ακολουθεί, το κάθε τμήμα τις οποίας μπορεί να αντικατασταθεί από μια μεταβλητή.

- PREFIX foaf: <http://xmlns.com/foaf/0.1/>
  SELECT *
  WHERE {
  ?person foaf:name ?name .
  ?person foaf:mbox ?email .
  }

Το παραπάνω ερώτημα χρησιμοποιεί πολλαπλές τριπλέτες για να επιστρέψει πολλαπλές τιμές μεταβλητών από την συγκεκριμένη βάση. Η εντολή SELECT * επιστρέφει τις τιμές όλων των μεταβλητών οι οποίες αναφέρονται στο ερώτημα και όχι μόνο της ?name όπως στο πρώτο παράδειγμα. Η τιμές των υποκειμένων και αντικειμένων των τριπλετών μπορούν να είναι strings, urls ή ακόμα και ιδιότητες μιας κλάσης:

- ?craft foaf:name "Apollo 7" .
- <http://nasa.dataincubator.org/spacecraft/1968-089A> space:discipline ?disc .
- ?craft a space:Spacecraft

Επιπλέον μπορούμε να περιορίσουμε τα επιστρεφόμενα αποτελέσματα με τη χρήση διαφόρων περιορισμών όπως οι παρακάτω:

- FILTER (?population > 15000000) . (φιλτράρει τις τιμές μιας μεταβλητής)
- LIMIT 50 (επιστρέφει μόνο τα 50 πρώτα αποτελέσματα)
- SELECT DISTINCT ?concept (επιστρέφει μόνο της διακριτές τιμές μιας μεταβλητής, δεν συμπεριλαμβάνει διπλά αποτελέσματα)
- ORDER BY DESC(?population) (ταξινόμει τα επιστρεφόμενα αποτελέσματα από το μικρότερο στο μεγαλύτερο) [29]

2.6 Οι πιο γνωστές βάσεις ανοικτών διασυνδεδεμένων δεδομένων: DBpedia, Yago

Η DBpedia είναι μια από τις πιο διαδεδομένες ανοιχτές βάσεις διασυνδεδεμένων δεδομένων η οποία βασίζεται στην τεχνολογία cloud και περιλαμβάνει δεδομένα από ένα ευρύ φάσμα άρθρων της Wikipedia τα οποία δημοσιεύει ως ανοιχτά διασυνδεδεμένα δεδομένα, τα μετατρέπει δηλαδή σε μορφή RDF. [8] Έχει διαμορφωθεί σταδιακά με βάση τις πιο συχνά χρησιμοποιούμενες εγγραφές της Wikipedia. Η οντολογία Dbpedia αυτή τη στιγμή περιλαμβάνει 685 κλάσεις οι οποίες έχουν διαμορφωθεί με μια συγκεκριμένη ιεραρχία και περιγράφονται με 2.795 διαφορετικές ιδιότητες. [27] Είναι διαθέσιμη σε πολλές γλώσσες (σε κάθε γλώσσα διαφέρει ο όγκος της πληροφορίας), ανάμεσα σε αυτές είναι και τα Ελληνικά (http://wiki.el.dbpedia.org/el). [8]

Η έκδοση 3.2 της DBpedia εισήγαγε μια νέα μέθοδο αντιστοίχισης της πληροφορίας από τη Wikipedia στην οντολογία DBpedia η οποία βασίζεται στην χαρτογράφηση που έχει γίνει με ανθρώπινη παρέμβαση στα infoboxes της Wikipedia. Η χαρτογράφηση αυτή ορίζει κανόνες για τον υπολογισμό των τιμών που θα δοθούν στα infoboxes αυτά καθώς και επισημαίνει αδυναμίες του συστήματος της Wikipedia, όπως διαφορετικά infoboxes που αντιστοιχούν στην ίδια κλάση, χρήση διαφορετικών ονομάτων για την ίδια ιδιότητα και όχι καλά ορισμένους τύπους δεδομένων για τιμές ιδιοτήτων. Επομένως, στην έκδοση αυτή τα δεδομένα της οντολογίας είναι "καθαρότερα" και σαφώς καλύτερα δομημένα. Με την έκδοση αυτή έχουμε πλέον διαθέσιμη μια δημόσια βιβλιοθήκη για την χαρτογράφηση της πληροφορίας που προέρχεται από την Wikipedia, την επεξεργασία της υφιστάμενης πληροφορίας καθώς και την επεξεργασία της Dbpedia οντολογίας. Αυτό επιτρέπει ακόμα και σε εξωτερικούς συνεισφέροντες να χαρτογραφούν πληροφορία η οποία τους ενδιαφέρει και να επεκτείνουν την υπάρχουσα οντολογία με επιπλέον κλάσεις και ιδιότητες. [27]

Από την έκδοση 3.7 η Dbpedia είναι πλέον ένα κατευθυνόμενο άκυκλο γράφημα και όχι ένα δέντρο. Οι κλάσεις μπορούν να έχουν πολλές υπερκλάσεις οι οποίες είναι χρήσιμες στην χαρτογράφηση του schema.org. Η τρέχουσα έκδοση της οντολογίας είναι διαθέσιμη στον παρακάτω σύνδεσμο: http://mappings.dbpedia.org/server/ontology/classes/. Αυτή τη στιγμή αποτελείται από 4.233.000 αναφορές. Ο παρακάτω πίνακας δείχνει τον αριθμό των αναφορών που υπάρχουν για τις διάφορες κλάσεις της οντολογίας: [27]

Κλάση	**Αναφορές**
Resource (συνολικά)	4,233,000
Place	735,000

Person	1,450,000
Work	411,000
Species	251,000
Organisation	241,000

Πίνακας 1: Οι κλάσεις της οντολογίας Dbpedia

Μια ακόμα ιδιαίτερα διαδεδομένη βάση ανοιχτών διασυνδεδεμένων με ελεύθερη πρόσβαση είναι η YAGO(Yet Another Great Ontology) η οποία αναπτύχθηκε από το Ινστιτούτο για την Επιστήμη των Υπολογιστών Max Planck με έδρα το Saarbrücken της Γερμανίας. Μέχρι και το 2020, η βάση YAGO3, περιλαμβάνει περισσότερες από 17 εκατομμύρια οντότητες και περισσότερα από 150 εκατομμύρια στοιχεία για τις οντότητες αυτές. Τα περιεχόμενα της βάση αυτής εξάχθηκαν αυτόματα από τις βάσεις Wikipedia (π.χ. κατηγορίες, ανακατευθύνσεις, infoboxes), WordNet (π.χ. synsets, hyponymy) και NeoNames. Η ακρίβεια της βάσης YAGO έχει εκτιμηθεί πάνω από 95% με βάση ένα τυχαίο δείγμα γεγονότων. Προκειμένου να ενσωματωθεί στο cloud, η YAGO έχει συνδεθεί με την οντολογία Dbpedia καθώς και με την SUMO(Suggested Upper Merged Ontology).[28]

Η YAGO3 παρέχεται σε μορφή γλώσσας Turtle και tsv(Tab Separated Values). Στο διαδίκτυο υπάρχουν αποθέματα ολόκληρης της βάσης καθώς και πιο εξειδικευμένα κομμάτια της με συγκεκριμένη θεματολογία. Αναζήτηση στη βάση YAGO3 μπορεί να πραγματοποιηθεί μέσω διαφόρων φυλλομετρητών καθώς και μέσω ενός SPARQL endpoint που φιλοξενείται από την OpenLink Software. Ο πηγαίος κώδικας της βάσης είναι διαθέσιμος στο GidHub. Τέλος, αξίζει να αναφέρουμε ότι η YAGO χρησιμοποιείται στο σύστημα τεχνητής νοημοσύνης Watson.[28]

2.7 Αλγόριθμοι Αναζήτησης σε Linked Open Data

Δεδομένα που αφορούν τις προτιμήσεις των χρηστών καθώς και προτάσεις τους σχετικά με προϊόντα/υπηρεσίες αποτελούν σημαντικό κομμάτι του ecommerce. Ο όγκος των δεδομένων έχει αυξηθεί σε τόσο μεγάλο βαθμό που η επεξεργασία τους από άνθρωπο είναι πλέον αδύνατη. Οι κλασικές μηχανές αναζήτησης οι οποίες διαθέτουν αλγορίθμους που βασίζονται στην ιστορικότητα των αναζητήσεων ενός χρήστη πάσχουν σε τρία βασικά σημεία:

1. Για τους νέους χρήστες μιας ιστοσελίδας οι αλγόριθμοι που παράγουν τις προτάσεις είναι ανίκανοι να φέρουν σωστά αποτελέσματα καθώς λόγω έλλειψης ιστορικότητας υπάρχει επαρκής ανατροφοδότηση πληροφοριών.
2. Όταν τεθεί σε λειτουργία μια μηχανή αναζήτησης, Το πρώτο χρονικό διάστημα, δεν μπορεί να παράξει σωστές προβλέψεις καθώς έχει ανεπαρκή και ελάχιστα δεδομένα.
3. Όταν ένα νέο προϊόν εισάγεται για πρώτη φορά σε ένα eshop συνήθως υπάρχει περιορισμένη ποσότητα προϊόντων που να ανήκουν στην ίδια κατηγορία με αυτό και επομένως και πάλι τα αποτελέσματα των αλγορίθμων μπορεί να είναι αβέβαια λόγω έλλειψης πληροφοριών. [11]

Ο πρώτος αλγόριθμος αναζήτησης της Google που δημιουργήθηκε για να κατατάξει τις ιστοσελίδες σε βαθμίδες ποιότητας ήταν ο PageRank. Η πρώτη αναφορά σε αυτόν έγινε σε ένα άρθρο του 1998. Στον Pagerank λαμβάνονται θετικά υπόψη όλες οι διασυνδέσεις που έχει μια ιστοσελίδα σε άλλες ιστοσελίδες (εσωτερικές και εξωτερικές) καθώς όσο περισσότερα είναι τα links τόσο πιο πολύ αυξάνεται η σημαντικότητα της, επομένως ο αλγόριθμος χαρακτηρίζει την ιστοσελίδα με υψηλή σπουδαιότητα. Στην αρχή του υπολογισμού η πιθανότητες να κάνει κλικ ένας χρήστης σε έναν λινκ κατανέμονται εξίσου μεταξύ όλων των λινκς που ανήκουν στο ίδιο γκρουπ. Ωστόσο, η πιθανότητα αυτή στα ανοιχτά διασυνδεδεμένα δεδομένα δεν είναι ομοιόμορφα κατανεμημένη. Επομένως δε μπορεί να χρησιμοποιηθεί ο ίδιος αλγόριθμος αναζήτησης. Έχουν γίνει προσπάθειες ανάπτυξης μηχανών αναζήτησης που ειδικεύονται στα ανοιχτά διασυνδεδεμένα δεδομένα οι οποίες βασίζονται σε 3 βήματα: 1) τα αποτελέσματα διαμορφώνονται συλλέγοντας δεδομένα από διαφορετικούς τομείς διασυνδεδεμένων δεδομένων χρησιμοποιώντας απλή αναζήτηση κειμένου. 2) Τα αποτελέσματα αυτά βελτιστοποιούνται με σκοπό την διαγραφή διπλότυπων δεδομένων. 3) Εφαρμόζεται μια μέθοδος ταξινόμησης/ βαθμολόγησης (ranking). [19]

Ο μεγάλος αριθμός διασυνδεδεμένων δεδομένων που παράγεται καθημερινά έχει κεντρίσει το ενδιαφέρον των ερευνητών οι οποίοι προσπαθούν να αναπτύξουν πλατφόρμες για την αξιοποίηση των δυνατοτήτων τους. Παραδείγματα μηχανών αναζήτησης Open Data οι οποίες ανιχνεύουν διασυνδεδεμένα δεδομένα από τον σημασιολογικό ιστό είναι η Swoogle Semantic Web Search Engine (SWSE), Falcons, Sindica και Watson.[19]

Η Google ανιχνεύει τριπλέτες διασυνδεδεμένων δεδομένων για την διαμόρφωση του Public Chart API προκειμένου να πετύχει την ελαχιστοποίηση του κατακερματισμού των αποτελεσμάτων αναζήτησης που μπορούν να αφορούν προϊόντα, reviews ή ακόμα και ανθρώπους. Η Yahoo μέσω της BOSS API πραγματοποιεί ανίχνευση διασυνδεδεμένων δεδομένων σε συνδυασμό με το εργαλείο searchMonkey για τη βελτίωση των αποτελεσμάτων αναζήτησης. Οι Pagerank και HIT μετρούν τη χρησιμότητα των σελίδων αναλύοντας τη δομή τους. Σε αυτές τις προσεγγίσεις η πιθανότητα να κάνει ο χρήστης κλικ σε κάποιο link κατανέμεται ομοιόμορφα σε όλες

τις σελίδες του γκρουπ. Ωστόσο, όπως ήδη αναφέρθηκε, όταν η πιθανότητα του κλικ σε έναν σύνδεσμο δεν είναι ομοιόμορφη όπως συμβαίνει στον σημασιολογικό ιστό οι παραπάνω προσεγγίσεις αποτυγχάνουν να βγάλουν τα σωστά αποτελέσματα. [19]

Ο Pop-Rank είναι ένας αλγόριθμος ανάλυσης συνδέσμων σε επίπεδο αντικειμένων που προσδίδει αυτόματα έναν παράγοντα διάδοσης σημαντικότητας σε κάθε είδους σχέση μεταξύ των αντικειμένων. Το σύστημα ObjectRank από την άλλη πραγματοποιεί αναζήτηση βάση keywords για να βαθμολογήσει τα αντικείμενα μιας βάσης τα οποία σχετίζονται σημασιολογικά μεταξύ τους. Ο αλγόριθμος ξεκινάει από έναν αρχικό κόμβο (αντικείμενο) ο οποίος περιλαμβάνει την λέξη κλειδί και στη συνέχεια η αξία του μεταβιβάζεται σε άλλα αντικείμενα σύμφωνα με τις σημασιολογικές σχέσεις που υπάρχουν. [19]

Η μηχανή αναζήτησης Swoogle χρησιμοποιεί τον αλγόριθμο OntoRank, μια εκδοχή του PageRank για τον σημασιολογικό ιστό. Ο OntoRank βασίζεται στη μέθοδο ανάλυσης συνδέσμων. Η σημαντικότητα μιας οντολογίας αξιολογείται με στατικό τρόπο ενώ το ερώτημα του χρήστη δεν θεωρείται σημαντικός παράγοντας στην ταξινόμηση των αποτελεσμάτων. Η μηχανή Swoogle χρησιμοποιεί ένα τυχαίο μοντέλο περιήγησης το οποίο περιγράφει τα διάφορα είδη συνδέσμων που μπορούν να προκύψουν ανάμεσα στα έγγραφα του σημασιολογικού ιστού. Υπολογίζουν την σημαντικότητα των πόρων πραγματοποιώντας ανάλυση των συνδέσμων την ώρα που ο χρήστης πληκτρολογεί το ερώτημά του. [19]

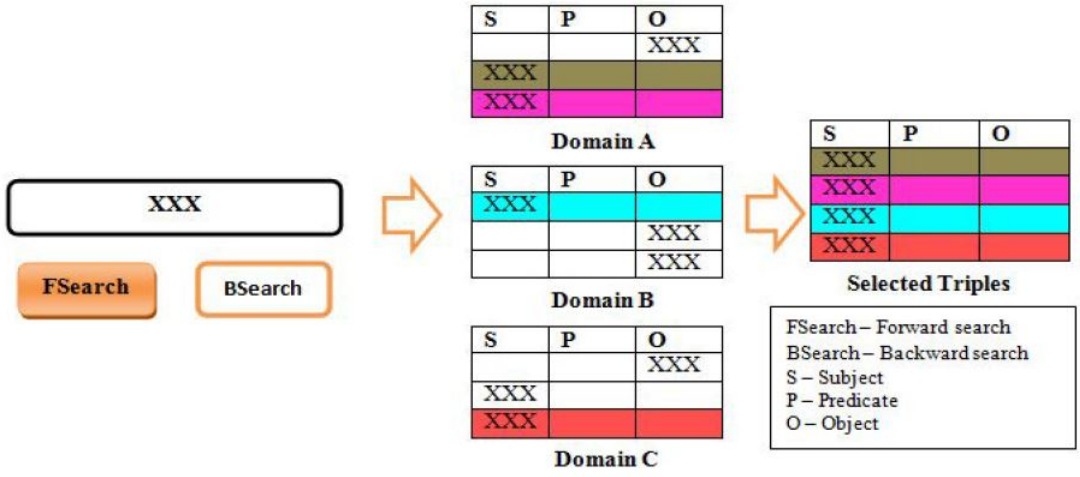
Η μηχανή αναζήτησης Falcons αναζητά οντότητες σε δεδομένα RDF. Χαρτογραφεί φράσεις κλειδιά ως σχέσεις μεταξύ των οντοτήτων και περιορίζει άμεσα τα αρχικά αποτελέσματα χρησιμοποιώντας ιεραρχίες κλάσεων. Η μηχανή αναζήτησης Dink προσφέρει σημασιολογική ταξινόμηση των δεδομένων RDF και βασίζεται στο λεξιλόγιο VoID (Vocabulary of Interlinked Datasets) για την περιγραφή των datasets. Αναλύει τους συνδέσμους ανάμεσα στα datasets χρησιμοποιώντας τις πληροφορίες που προέρχονται από τις εξηγήσεις που υπάρχουν στο VoID. Λαμβάνει υπόψη του τους τύπους των σχέσεων και τον αριθμό των συνδέσμων. Ωστόσο, δεν υπάρχουν πολλές περιγραφές στο VoID επομένως η προσέγγιση αυτή είναι λιγότερο επεκτάσιμη. [19]

2.7.1 Αναζήτηση RDF Τριπλετών

Όταν γίνεται αναζήτηση ενός όρου, οι πληροφορίες που εντοπίζονται είναι είτε χαρακτηριστικά του όρου αναζήτησης είτε πληροφορίες σχετικά με τους όρους οι οποίοι έχουν αναφερθεί ως χαρακτηριστικό του όρου αναζήτησης. Στις βάσεις Ανοικτών Διασυνδεδεμένων Δεδομένων η αναζήτηση πραγματοποιείται σε βάσεις αποθήκευσης τριπλετών όλων των τομέων οι οποίοι περιέχουν open data και τα αποτελέσματα παρουσιάζονται σε μορφή RDF τριπλετών. Η κάθε μια από αυτές τις βάσεις μπορεί να ανήκει σε διαφορετικό server. [19]

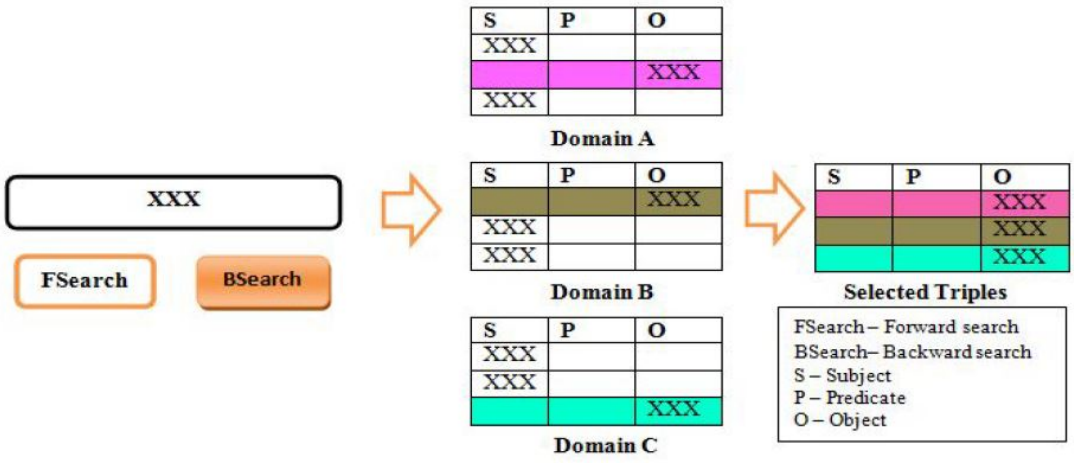
Όπως φαίνεται στο παράδειγμα στις παρακάτω εικόνες έχουμε 3 διαφορετικούς τομείς οι οποίοι αντιπροσωπεύουν την δομή μιας τριπλέτας RFD. Ο κάθε τομέας χωρίζεται σε στήλες, η 1η αφορά το υποκείμενο, η 2η το κατηγόρημα και η 3η το αντικείμενο. Με βάση την επιλογή του χρήστη η αναζήτηση μπορεί να πραγματοποιηθεί προς τα εμπρός ή προς τα πίσω:

- Αναζήτηση προς τα εμπρός: όταν ο χρήστης που πραγματοποιεί αναζήτηση στα χαρακτηριστικά ενός όρου διαθέτει το υποκείμενο τις τριλέτας και αναζητά το αντικείμενο η αναζήτηση χαρακτηρίζεται ως αναζήτηση προς τα εμπρός (forward search). Σε πρώτη φάση επιλέγουμε τους τομείς που περιλαμβάνουν τον όρο της αναζήτησης και στη συνέχεια ξεχωρίζουμε τις τριπλέτες που περιέχουν τον όρο ως υποκείμενο. Π.χ. αν θέλουμε να αναζητήσουμε τα χαρακτηριστικά ενός συγκεκριμένου κινητού τηλεφώνου ο χρήστης πραγματοποιεί αναζήτηση προς τα εμπρός καθώς θέλουμε να βρούμε τριπλέτες με το προϊόν που μας ενδιαφέρει ως υποκείμενο. [19]



Εικόνα 3 : Αναζήτηση προς τα εμπρός

Αναζήτηση προς τα πίσω: όταν ο χρήστης θέλει να αναζητήσει έναν όρο ο οποίος εμφανίζεται ως χαρακτηριστικό σε άλλους όρους, χρησιμοποιεί αναζήτηση προς τα πίσω. Στην περίπτωση αυτή ο χρήστης γνωρίζει το αντικείμενο των τριπλετών και ψάχνει για το υποκείμενο. Αρχικά επιλέγουμε τους τομείς που περιλαμβάνουν τον όρο της αναζήτησης και συλλέγουμε τις τριπλέτες που περιέχουν τον όρο ως αντικείμενο. Π.χ. όταν ο χρήστης αναζητά για κινητά τηλέφωνα τα οποία διαθέτουν ένα συγκεκριμένο χαρακτηριστικό χρησιμοποιεί αναζήτηση προς τα πίσω καθώς επιλέγει τις τριπλέτες που περιέχουν το χαρακτηριστικό ως αντικείμενο. [19]



Εικόνα 4 : Αναζήτηση προς τα πίσω

2.7.2 Εξαγωγή των δεδομένων

Το πρώτο βήμα είναι η ανάκτηση των περισσότερο σχετικών δεδομένων από τα αποτελέσματα του ερωτήματος αναζήτησης. Τα εξαγόμενα δεδομένα από τις διαφορετικές πηγές αποθηκεύονται σε ένα σετ από τριπλέτες ως αποτέλεσμα της αναζήτησης. Υπάρχουν διάφορες μέθοδοι για την εξαγωγή των δεδομένων:

- Προσθήκη/προσάρτηση datasets: Τα επιθυμητά σύνολα δεδομένων ανακτώνται από τις αποθήκες τριπλετών των διαφόρων τομέων. Οι εξαγόμενες τριπλέτες τοποθετούνται η μία μετά την άλλη. Αυτός είναι και ο απλοϊκότερος τρόπος συλλογής δεδομένων από διαφορετικές αποθήκες δεδομένων.
- Εξαγωγή τριπλετών με χρήση ερωτημάτων SPARQL: η γλώσσα SPARQL χρησιμοποιείται για την εξαγωγή δεδομένων από μεγάλες βάσεις δεδομένων RDF. Τα ανοιχτά διασυνδεδεμένα δεδομένα μπορούν να εξαχθούν με τρεις τρόπους: 1) Εξάγουμε τα δεδομένα από τις διάφορες πηγές δεδομένων εφαρμόζοντας το ερώτημα SPARQL στην κάθε μία ξεχωριστά και στη συνέχεια συνδέουμε τα εξαγόμενα δεδομένα. 2) Εφαρμόζουμε το ερώτημα σταδιακά σε κάθε τομέα δεδομένων. Βασιζόμενοι στο αποτέλεσμα του τελευταίου κάθε φορά ερωτήματος αντικαθιστούμε το ερώτημα με το αποτέλεσμα αυτό, έτσι το ερώτημα βήμα βήμα εφαρμόζεται σε όλες τις πηγές δεδομένων. 3) Ο τελευταίος τρόπος βασίζεται στην μαζική εφαρμογή του ερωτήματος SPARQL που μας δίνει πρόσβαση σε διαφορετικούς τύπους δεδομένων σχετικής προέλευσης.
- Μετακίνηση ιστοσελίδων HTML: σε αυτή τη μέθοδο ο επιθυμητός τομέας δεδομένων είναι προκαθορισμένος και βασίζεται σε πρότυπα καταλληλότητας δεδομένων(HTML). Τα επιλεγμένα δεδομένα εξάγονται αυτόματα από τα διαθέσιμα σύνολα δεδομένων (ιστοσελίδες) χρησιμοποιώντας κατάλληλα εργαλεία εξόρυξης. Τέλος, τα εξαγόμενα δεδομένα που παράγονται από τα διάφορα σύνολα ενώνονται. Σε μια διαφορετική προσέγγιση, ο μηχανισμός

εξόρυξης αρχικά επιλέγει ένα σύνολο δεδομένων για μια πρώτη συλλογή των σχετικών πληροφοριών και στη συνέχεια διασχίζει τα υπόλοιπα σχετικά σύνολα με βάση τις πληροφορίες που έχει συλλέξει από τα δεδομένα που έχει εξάγει προηγουμένως. [19]

Όποια μέθοδο και να επιλέξουμε, το εργαλείο εξόρυξης των δεδομένων θα πρέπει να είναι συμβατό με την συλλογή διαφορετικών δεδομένων που ανήκουν σε διαφορετικές οντολογίες με διαφορετική αρχιτεκτονική. [19]

Η βελτίωση των αποτελεσμάτων αναζήτησης μπορεί να πραγματοποιηθεί με τις παρακάτω τεχνικές:

- Χαρτογράφηση των οντολογιών: Με την ραγδαία αύξηση των διαθέσιμων οντολογιών στο διαδίκτυο υπάρχει ανάγκη για το κατάλληλο software προκειμένου να πετύχουμε τη μέγιστη αξιοποίησή τους. Επίσης, διάφορες οντολογίες θα πρέπει να είναι προσπελάσιμες από διάφορες συσκευές. Η χαρτογράφηση των οντολογιών αυτών προσφέρει μια κοινή πλατφόρμα για την ανταλλαγή πληροφοριών και τον προσδιορισμό του σημασιολογικού νοήματός τους. Τα περισσότερα σύνολα δεδομένων χρησιμοποιούν παραπάνω από μία οντολογίες για τον προσδιορισμό των δεδομένων τους και είναι σύνηθες κάθε dataset να περιέχει διαφορετικές οντολογίες. Επομένως, στα αποτελέσματα μπορεί να προκύψουν τα ίδια δεδομένα με διαφορετικά ονόματα ή ακόμα και διαφορετικά δεδομένα με ίδιο όνομα. Μια λύση στο πρόβλημα αυτό θα μπορούσε να αποτελέσει η χαρτογράφηση των οντολογιών. [19]
- Διπλότυπα δεδομένα: δημιουργούνται για 2 λόγους: 1) όταν υπάρχουν δύο πανομοιότυπα υποκείμενα τα οποία εμφανίζονται σε δύο διαφορετικές βάσεις δεδομένων με διαφορετική τιμή αντικειμένου και 2) όταν υπάρχουν διαφορετικές τιμές αντικειμένου για το ίδιο κατηγόρημα ενός υποκειμένου. Επομένως, από τα αποτελέσματα θα πρέπει να επιλεχθεί μόνο μια από τις τιμές αυτές. Η λύση σε αυτό το πρόβλημα είναι η επιλογή των δεδομένων με την μεγαλύτερη αυθεντικότητα και ο αποκλεισμός των υπολοίπων. Η αυθεντικότητα των δεδομένων μπορεί να κριθεί από τον βαθμό αυθεντικότητας του τομέα στον οποίο ανήκουν. [19]

2.8 Recommendation Systems και Ανοιχτά διασυνδεδεμένα δεδομένα

Ακόμα μια τεχνολογία η οποία αντιμετωπίζει παρόμοια προβλήματα και περιορισμούς όπως με τις μηχανές αναζήτησης είναι τα Recommendation Systems. Οι παρακάτω προτάσεις παρουσιάζουν τα προβλήματα των σημερινών recommendation systems στο κομμάτι του ecommerce και πως αυτά μπορούν να λυθούν με τη βοήθεια των ανοικτών διασυνδεδεμένων δεδομένων:

1. Περιεκτικότητα: όπως προαναφέρθηκε, οι ecommerce πλατφόρμες έχουν να αντιμετωπίσουν το πρόβλημα της έλλειψης επαρκών δεδομένων. Κυρίως, μικρά sites δεν έχουν επαρκώς μεγάλη πελατειακή βάση ώστε να τους παρέχει αρκετά ratings. Στο σημείο αυτό βοηθούν οι βάσεις Ανοικτών Διασυνδεδεμένων Δεδομένων οι οποίες ξεπερνούν το πρόβλημα των περιορισμών των δεδομένων καθώς το νέφος LOD παρέχει δισεκατομμύρια τριπλέτες οι οποίες προέρχονται από διάφορους τομείς για να τις αξιοποιήσουν. Τέτοιου είδους δεδομένα μπορούν να χρησιμοποιηθούν π.χ. στην αγορά των multimedia ή σε ταξιδιωτικά πρακτορεία.
2. Προσαρμοστικότητα: τα recommendation systems των περισσότερων ecommerce sites δε δίνουν την δυνατότητα στους χρήστες να περιορίσουν τα επιστρεφόμενα αποτελέσματα με βάση συγκεκριμένα κριτήρια. Για να πετύχουμε πιο εξατομικευμένα αποτελέσματα θα ήταν επιθυμητό να εφαρμόσουμε φίλτρα πριν ή μετά τα αποτελέσματα των προτεινόμενων προϊόντων. Π.χ. ας πάρουμε έναν χρήστη ενός streaming site ο οποίος παρά το ιστορικό των αγορών του θέλει να παρέχει την πληροφορία ότι ενδιαφέρεται για τις ευρωπαϊκές ταινίες. Επίσης, σε τομείς όπως ο τουρισμός, οι προτιμήσεις των χρηστών εξαρτώνται από πολλούς παράγοντες όπως είναι οι σύντροφοι του ταξιδιού ή οι προτιμήσεις του χρήστη σε ταξιδιωτικούς προορισμούς. Τέλος, η παραμετροποίηση μπορεί να πραγματοποιηθεί εκτός από τους ίδιους τους χρήστες και από επαγγελματίες του marketing. Όπως π.χ οι προωθητικές ενέργειες που θα μπορούσαν να πραγματοποιούνται σε ειδικές εορταστικές περιόδους του χρόνου για να προωθήσουν προϊόντα μακράς διάρκειας. Αυτές οι προσεγγίσεις απαιτούν εφαρμογές πλούσιες σε δεδομένα τα οποία προσεγγίζονται με εκφραστικά ερωτήματα, όπως η γλώσσα SPARQL.
3. Προτάσεις σε πραγματικό χρόνο: να συνηθισμένα συστήματα προτάσεων βασίζονται σε προϋπολογισμένα αποτελέσματα. Ωστόσο, η αποτελεσματικότητα στα recommendation systems είναι άρρηκτα συνυφασμένη με την παραγωγή προτάσεων σε πραγματικό χρόνο. Οι απαιτήσεις των πελατών καθώς και των επιχειρήσεων δεν μπορούν να προβλεφθούν και επομένως τα συστήματα προτάσεων θα πρέπει να διαμορφώνουν τα αποτελέσματα τους σε πραγματικό χρόνο ώστε ο χρήστης να μπορεί να επιλέξει τα δεδομένα που του χρειάζονται. [11]

Τα παραπάνω προβλήματα λύνονται με την χρήση των Linked Open Data τα οποία προσφέρουν δωρεάν και ανοικτή πρόσβαση σε βάσεις διασυνδεδεμένων δεδομένων. Οι βάσεις αυτές είναι διαθέσιμες στο διαδίκτυο και περιλαμβάνουν πληροφορίες δοσμένες σε μορφή κατανοητή και επεξεργάσιμη από υπολογιστές (τριπλέτες RDF) οι οποίες συνδέονται μεταξύ τους με μοναδικούς συνδέσμους (URIs). [8]

Υπάρχουν δύο προσεγγίσεις ως προς τον τύπο του μηχανισμού φιλτραρίσματος της πληροφορίας που χρησιμοποιούν τα recommendation systems για την παραγωγή προτάσεων προϊόντων και υπηρεσιών σε έναν χρήστη:

1. Το φιλτράρισμα με βάση το περιεχόμενο: Η μέθοδος αυτή προτείνει αντικείμενα παρόμοια με αυτά στα οποία έχει δείξει προτίμηση ο χρήστης στο παρελθόν. Ο μηχανισμός αυτός συνήθως χρησιμοποιεί χαρακτηριστικά και πληροφορίες που αντλεί από το περιεχόμενο τους για να προσδιορίσει τα προφίλ του χρήστη και του αντικειμένου που αναζητά.
2. Το συνεργατικό φιλτράρισμα: Η τεχνική αυτή προτείνει αντικείμενα που προτιμούν άτομα τα οποία έχουν παρόμοιες προτιμήσεις με αυτές του χρήστη. Ο μηχανισμός αυτός δουλεύει με ομοιότητες ανάμεσα στα αντικείμενα οι οποίες βασίζονται σε αξιολογήσεις/βαθμολογίες χρηστών και επομένως δεν βασίζεται σε περιεχόμενο το οποίο μπορεί εύκολα να αναλυθεί και να επεξεργαστεί από μια μηχανή. [9]

Εκτός από τις παραπάνω κατηγορίες έχουν αναπτυχθεί υβριδικές προσεγγίσεις οι οποίες προσπαθούν να ενσωματώσουν χαρακτηριστικά από διάφορες μεθοδολογίες προκειμένου να αυξήσουν την αποτελεσματικότητα και την απόδοση των παραγόμενων προτάσεων. [5]

Τα περισσότερα συνεργατικά συστήματα έχουν σχεδιαστεί για να χειρίζονται αριθμητικές βαθμολογίες όπως πχ. τα αστέρια στην Amazon και στο Netflix. Ωστόσο, σε πολλές ecommerce πλατφόρμες και σε κοινωνικά δίκτυα οι προτιμήσεις των χρηστών εκφράζονται μέσα από θετικές/αρνητικές ή θετικές μόνο ψήφους, όπως τα likes του Facebook ή το thumbs up του YouTube. Επίσης, χαρακτηριστικά μπορούν να εξαχθούν από περιγραφές των αντικειμένων καθώς και από ετικέτες (tags) τα οποία έχουν δώσει οι ίδιοι οι χρήστες στα αντικείμενα. [9]

Με την χρήση του Facebook Graph API, τα αντικείμενα αυτά λαμβάνουν αναγνωριστικά ονόματα τα οποία αποτελούνται από απλό κείμενο που έχουν παράγει οι χρήστες. Με τη χρήση του API τα likes των χρηστών μετατρέπονται σε στοιχεία τεσσάρων διαστάσεων: αναγνωριστικό, όνομα, κατηγορία και χρονική στιγμή δημιουργίας του like (π.χ. {id: "35481394342", name: "The Godfather", category: "Movie", created_time: "2015-05-14T12:35:08+0000"}). Το όνομα αναφέρεται στον τίτλο που έχει δώσει ο χρήστης στη σελίδα που δημιούργησε, έτσι σελίδες με διαφορετικό όνομα μπορούν να αναφέρονται στο ίδιο αντικείμενο και επομένως τα likes των χρηστών μπορούν να εμφανίζονται σε διαφορετικές σελίδες αλλά να αναφέρονται στο ίδιο αντικείμενο.[9]

Πολλές έρευνες προτείνουν την παρακάτω μεθοδολογία για την παραγωγή μεταδεδομένων των αντικειμένων που μας αφορούν. Πρώτα συνδέουμε τα αντικείμενα με τις αντίστοιχες οντότητες μιας εξωτερικής πηγής δεδομένων, όπως η αντιστοίχιση των ονομάτων τους με τα URIs σημασιολογικών βάσεων δεδομένων όπως η DBpedia. Τα ανοιχτά διασυνδεδεμένα δεδομένα δεν επιτρέπουν απλώς την περιγραφή των αντικειμένων με βάση τα μεμονωμένα χαρακτηριστικά τους αλλά και τη δημιουργία σημασιολογικών δικτύων που αποτελούνται από σχετικά αντικείμενα και χαρακτηριστικά αντικειμένων. Δομώντας όλα τα διαθέσιμα δεδομένα σε έναν

γράφο αποκτούμε πολλά πλεονεκτήματα. Ανάμεσα σε αυτά είναι και η χρήση γνωστών αλγορίθμων οι οποίοι εφαρμόζονται σε σημασιολογικούς γράφους για άντληση διαφόρων ειδών πληροφορίας. Γνωστοί αλγόριθμοι αναζήτησης σε αυτού του είδους γράφους είναι ο PathRank, επέκταση του γνωστού PageRank που ανακαλύπτει διαφορετικά μονοπάτια σε ένα ετερογενές γράφο, ο HeteRec και ο SPrank.[9]

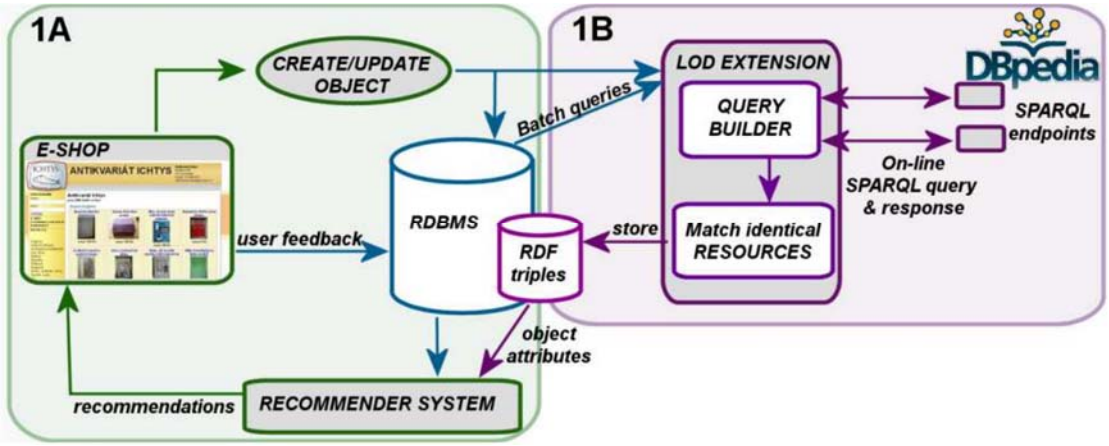
Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν σε ένα υβριδικό σύστημα προτάσεων το οποίο μπορεί να εκμεταλλευτεί τα διάφορα είδη πληροφορίας και των μεταξύ τους σχέσεων που απεικονίζονται σε έναν ομοιόμορφο γράφο. Τα γραφίματα αυτά μπορούν να επεκταθούν με πληροφορίες αντλούμενες από βάσεις Linked Open Data. [9]

Επίσης, οι ερευνητές έχουν ξεκινήσει να χρησιμοποιούν πηγές πληροφοριών Συνδεδεμένων Δεδομένων για να αντιμετωπίσουν το πρόβλημα της ανεπαρκής πληροφόρησης για τα προϊόντα. Το νέφος LOD περιλαμβάνει δεδομένα για κάθε είδος πληροφορίας γενικού ή ειδικού σκοπού αλλά και δεδομένα από ειδικούς τομείς. Η μέχρι τώρα πορεία τους έχει δείξει ότι τα συστήματα προτάσεων που βασίζονται σε Διασυνδεδεμένα Δεδομένα (LDRS - Linked Data Recommendation Systems) είναι περισσότερο ανταγωνιστικά και παράγουν πιο ακριβή αποτελέσματα σε σύγκριση με τις κλασικές μεθόδους παραγωγής προτάσεων. Ωστόσο, δεν εκμεταλλεύονται πλήρως το νεφος LOD καθώς οι προσεγγίσεις που έχουν γίνει απαιτούν μεγάλη προσοχή για την σωστή επιλογή και εξαγωγή των χαρακτηριστικών των αντικειμένων που μας ενδιαφέρουν πριν προχωρήσουμε στην εφαρμογή των μοντέλων. Όταν επιλεχθούν τα χαρακτηριστικά αυτά, το μοντέλο προτάσεων εντάσσεται στην ιστοσελίδα που θα το χρησιμοποιήσει και δεν μπορεί να προσαρμοστεί στις αλλαγές και τις ανάγκες της επιχείρησης. [10]

Το πρόβλημα της έλλειψης πληροφοριών μπορεί να αντιμετωπιστεί με προσεγγίσεις που βασίζονται στο περιεχόμενο των LODs για να βελτιώσουν τις διαθέσιμες πληροφορίες που υπάρχουν για τα χαρακτηριστικά ενός προϊόντος. Υπάρχουν ελάχιστα συστήματα που εξετάζουν τις προτιμήσεις των χρηστών αξιοποιώντας τεχνολογίες ανοικτών διασυνδεδεμένων δεδομένων, όπως η γλώσσα SPARQL, τα οποία μπορούν να διατυπώσουν τα ακριβή ερωτήματα των χρηστών. Τα συστήματα που βασίζονται σε ερωτήματα σε βάσεις ανοικτών διασυνδεδεμένων δεδομένων συχνά ακολουθούν εργασίες για ανίχνευση κοινών στοιχείων μεταξύ των δεδομένων των βάσεων ή εντοπίζουν κοινά πρότυπα SPARQL και αντιμετωπίζουν μεγάλους χρόνους εκτέλεσης όταν έχουν να επεξεργαστούν μεγάλο όγκο τριπλετών. [10]

Το παρακάτω σύστημα αποτελεί παράδειγμα σύνδεσης μιας μηχανής αναζήτησης με τη Βάση Διασυνδεδεμένων Δεδομένων DBpedia. Το σύστημα διατηρεί συνδέσμους ενός ή περισσοτέρων SPARQL Endpoinds σε διάφορα σύνολα ανοικτών διασυνδεδεμένων δεδομένων. Οι σύνδεσμοι αυτοί είναι συνήθως REST APIs ή απλά HTTP Services. Όταν ένα αντικείμενο δημιουργείται ή μεταβάλλεται το σύστημα

αυτομάτως αντιστοιχίζει μια σύνδεση SPARQL με ένα μοναδικό αναγνωριστικό του αντικειμένου, σε περίπτωση που το ίδιο αναγνωριστικό δεν υπάρχει το αντιστοιχίζει με αυτό που ταιριάζει καλύτερα.



Εικόνα 5 : Recommendation System ενός eshop που συνδέεται με βάση ανοικτών διασυνδεδεμένων δεδομένων

Το τμήμα 1Α της παραπάνω εικόνας παρουσιάζει ένα eshop που διαθέτει μια μηχανή αναζήτησης η οποία βασίζεται τόσο στην ανατροφοδότηση από το περιεχόμενο που παράγουν οι χρήστες του site όσο και σε recommendation system το οποίο συνδέεται με μια βάση ανοικτών διασυνδεδεμένων δεδομένων. Το τμήμα 1Β παρουσιάζει την επέκταση του συστήματος αυτού για την δημιουργία ερωτημάτων και την αποθήκευση δεδομένων σε μια βάση ανοικτών διασυνδεδεμένων δεδομένων. [8]

2.9 Πως οι τεχνολογίες σημασιολογικού ιστού μπορούν να βελτιώσουν το ecommerce

Οι υπηρεσίες του διαδικτύου, όπως είναι και ο κλάδος του ecommerce, μπορούν να θεωρηθούν αυτόνομα κομμάτια μιας επιχείρησης και ταυτόχρονα μέρη ενός συνόλου τα οποία προσφέρουν προστιθέμενη αξία στην επιχείρηση με τις υπηρεσίες που προσφέρουν. Με τη χρήση των υπηρεσιών διαδικτύου μια επιχείρηση μπορεί να μειώσει το κόστος των διαδικασιών της και να καλύψει επιχειρηματικές ανάγκες που προκύπτουν κατά τη διάρκεια των online συναλλαγών με τους πελάτες και τους συνεργάτες της. Με τη βοήθεια των τεχνολογιών του σημασιολογικού ιστού, τα μεταδεδομένα τα οποία περιγράφουν τις ιδιότητες και τις χρήσεις των υπηρεσιών διαδικτύου χρησιμοποιούνται για την ανάπτυξη κανόνων για την αυτοματοποίηση και ενορχήστρωση των διαφόρων μερών μιας εταιρείας. [12]

Μια πρόσφατη προσέγγιση η οποία χρησιμοποιείται για την αναγνώριση των υπηρεσιών διαδικτύου οι οποίες πληρούν τις απαιτήσεις και τις προϋποθέσεις που επιθυμούμε είναι και η χρήση των Ανοικτών Διασυνδεδεμένων Δεδομένων (Linked Open Data - LOD). Όπως ήδη αναφέρθηκε η μέθοδος αυτή βασίζεται στη χρήση

δομημένων δεδομένων τα οποία μπορούν να διαβαστούν από μηχανές/υπολογιστές. [12]

Το ενδιαφέρον είναι ότι οι χρήστες θεωρούν ότι οι παραγωγή και χρήση γράφων πληροφοριών είναι σημαντικός παράγοντας στην βελτίωση ενός διαδικτυακού τόπου ηλεκτρονικού εμπορίου. Στη κατεύθυνση αυτή είναι σημαντικό και βοηθητικό για έναν χρήστη, είτε προχωρήσει στην αγορά ενός προϊόντος είτε όχι, να έχει τη δυνατότητα εύρεσης των προϊόντων και υπηρεσιών που τον ενδιαφέρουν μέσα από κατηγορίες και υποκατηγορίες των χαρακτηριστικών τους. Οι τακτικές αυτές είναι χρήσιμες στην ανάπτυξη αποτελεσματικών μηχανών αναζήτησης στο ηλεκτρονικό εμπόριο καθώς αυξάνουν την αξία χειρισμού των δεδομένων. [5]

Οι καταναλωτές επιθυμούν να έχουν πρόσβαση στις πληροφορίες που σχετίζονται με τα χαρακτηριστικά των προϊόντων που πιθανόν τους ενδιαφέρουν. Μια πρόκληση που καλούμαστε να αντιμετωπίσουμε είναι η βελτίωση των αναλύσεων των αποτελεσμάτων αναζήτησης ώστε να δίνουμε απαντήσεις σε ερωτήματα όπως: Ποιο είναι το επόμενο προϊόν που θα θελήσει να αγοράσει ο καταναλωτής; Ποια θα είναι η επόμενη προωθητική ενέργεια; Ποια θα είναι η επόμενη καλή περίοδος για αγορά αποθεμάτων; Κ.α. Όλα αυτά είναι ερωτήματα τα οποία απαντώνται μέσα από επιπλέον αναλύσεις στο περιεχόμενο της φυσικής γλώσσας επεξεργασίας. [5]

Όταν αναφερόμαστε στις τεχνολογίες σημασιολογικού ιστού γνωρίζουμε ότι ένα μεγάλο και σημαντικό κομμάτι τους αποτελούν τα Big Data. Οι τεχνολογίες του σημασιολογικού ιστού μπορούν να χρησιμοποιηθούν για την διαχείριση μεγάλου όγκου δεδομένων. Ουσιαστικά τα Big Data αποτελούν κομμάτι των Open Data, καθώς η αξία των big data αυξάνεται όταν αυτά μετατρέπονται σε Open Data και είναι πλέον προσβάσιμα και αξιοποιήσιμα από τον οποιονδήποτε (μηχανές και ανθρώπους). [1] Οι τεχνολογίες σημασιολογικού ιστού προσφέρουν ένα κοινό πλαίσιο το οποίο επιτρέπει τον διαμοιρασμό και επαναχρησιμοποίηση των δεδομένων από πολλές εφαρμογές, οργανισμούς και κοινότητες. Επίσης, χρησιμοποιείται ως σύνδεσμος ανάμεσα σε πληροφοριακά συστήματα και εφαρμογές με διαφορετικό περιεχόμενο. Οι επιχειρηματικές εφαρμογές που βασίζονται στις τεχνολογίες σημασιολογικού ιστού βελτιώνουν την διαδικασία ανάκτησης των πληροφοριών, οι οποίες είναι απαλλαγμένες από περιττές πληροφορίες, και αυξάνουν την ακρίβεια των δεδομένων που ανακτώνται. [5]

Σύμφωνα με πρόσφατη έρευνα που πραγματοποιήθηκε στις Ηνωμένες Πολιτείες, οι εταιρείες που δραστηριοποιούνται στον κλάδο του ecommerce και έχουν εισάγει τα Big Data Analytics στην αλυσίδα αξίας τους το 2016 σημείωσαν 5-6% αύξηση στην παραγωγικότητά τους σε σύγκριση με τους ανταγωνιστές τους. Μια λίγο παλαιότερη έρευνα του 2014 της BSA Software Alliance και πάλι στις Ηνωμένες Πολιτείες, έδειξε ότι στο 56% των εταιρειών που συμμετείχαν η υιοθέτηση των Big Data Analytics συνείσφερε τουλάχιστον στο 10% στην συνολική ανάπτυξή τους. [1]

Τα Big Data, και περισσότερο τα Big Open Data, βοηθούν τις εταιρείες που δραστηριοποιούνται τον τομέα του ecommerce να:

1. Ανιχνεύσουν τις κινήσεις των ανταγωνιστών τους και να τις αξιοποιήσουν ώστε να μετατρέψουν τους πελάτες που θα πραγματοποιήσουν μια αγορά σε σταθερούς πελάτες.
2. Πετύχουν μεγαλύτερο Conversion Rate.
3. Βελτιώσουν την διαδικασία λήψης αποφάσεων.
4. Δώσουν επιπλέον δύναμη και αξία στους πελάτες τους.
5. Βελτιώσουν της αποδοτικότητα του κόστους διαχείρισης της συναλλαγής (π.χ. Recommendation Algorithms που θα εξετάσουμε αργότερα) και τον χρόνο (π.χ. χρόνος αναζήτησης).
6. Εντοπίσουν πιστούς και κερδοφόρους πελάτες.
7. Επιτύχουν στοχευμένη διαφήμιση μέσα από τον εντοπισμό δυνητικών πελατών.
8. Καθορίσουν την βέλτιστη τιμολογιακή πολιτική για τα προϊόντα ή τις υπηρεσίες που προσφέρουν.
9. Βελτιώσουν τις υπηρεσίες υποστήριξης προς τους πελάτες τους.
10. Εντοπίσουν προβλήματα στην ποιότητα των προϊόντων ή υπηρεσιών μέσα από τις εμπειρίες και τις απόψεις των πελατών τους.
11. Αναπτύξουν καινοτόμα μοντέλα υπηρεσιών όπως αυτά που έχουν υιοθετήσει κατά καιρούς μεγάλες εταιρίες όπως η Rolls Royce, Amazon, Google και Netflix. [1]

Η χρήση των Business Analytics για την λήψη αποφάσεων θα πρέπει να συμβαδίζει με την στρατηγική της επιχείρησης και να συμβάλει στην αναγνώριση νέων ευκαιριών, νέων προϊόντων και υπηρεσιών που φέρνουν επιπλέον αξία στην επιχείρηση. Τα σύγχρονα εργαλεία του σημασιολογικού ιστού παρατηρούν το εξωτερικό περιβάλλον και αναγνωρίζουν πιθανά εξελισσόμενα γεγονότα και ευκαιρίες μέσα από την αναγνώριση συμπεριφορών χρηστών. [1]

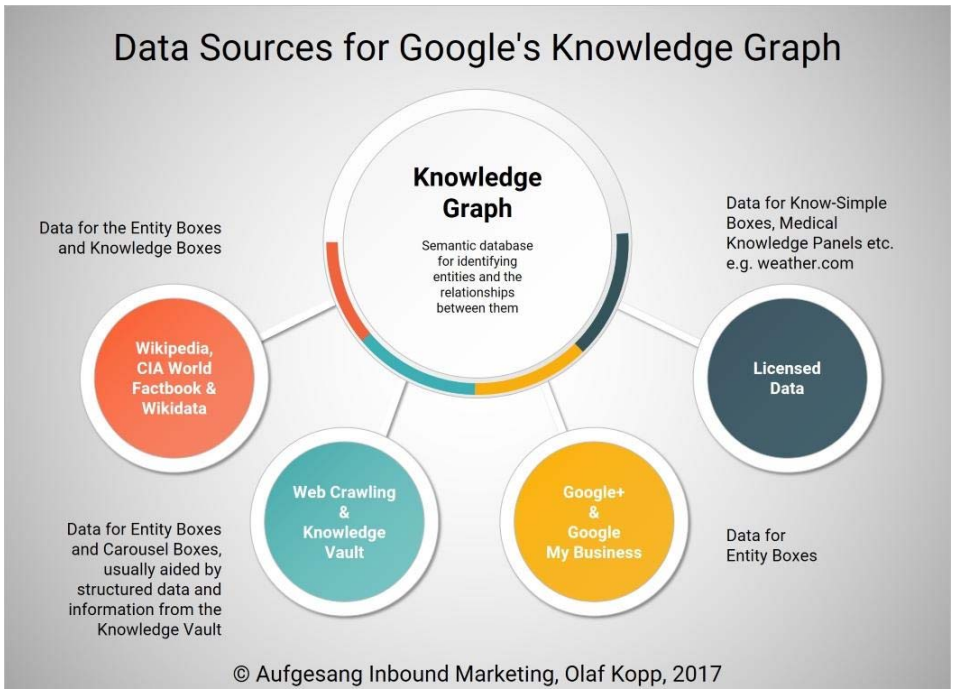
Τα δεδομένα του σημασιολογικού ιστού αποτελούν πληροφορίες οι οποίες μπορούν να διαβαστούν από μηχανές (υπολογιστές) και έτσι βοηθούν τους managers να περιγράψουν τη δομή της πληροφορίας που περιέχουν και να δώσουν νόημα στο περιεχόμενό τους. Με τον τρόπο αυτόν ένας υπολογιστής μπορεί να επεξεργαστεί την πληροφορία αυτούσια χρησιμοποιώντας διαδικασίες και τεχνικές παρόμοιες με την κριτική ανθρώπινη σκέψη. Επίσης, εξάγει συμπεράσματα παράγοντας σημαντικά αποτελέσματα που βοηθούν τους υπολογιστές στην αυτοματοποιημένη έρευνα και συλλογή χρήσιμων πληροφοριών. [5]

Τα Open Data προχωρούν την όλη διαδικασία ένα βήμα πιο πέρα αξιοποιώντας τις τεχνολογίες σημασιολογικού ιστού οι οποίες έχουν υιοθετηθεί από μεγάλες εταιρείες. Είναι χαρακτηριστικό το γεγονός ότι η μηχανή αναζήτησης της Google αξιοποιεί τα big data προκειμένου να βελτιώσει τα αποτελέσματα των αναζητήσεων και τους αλγορίθμους που χρησιμοποιεί στον υπολογισμό των στοχευμένων διαφημίσεων που θα προβάλει στους χρήστες του διαδικτύου. [1]

Ακόμη ένα παράδειγμα αποτελεί ο αλγόριθμος που χρησιμοποιεί η Amazon για τα προϊόντα που προτείνει στους πελάτες της. Κάνει χρήση τεχνικών μοντελοποίησης οι οποίες φιλτράρουν τα δεδομένα αυτά και τροφοδοτούν τον μηχανισμό προτεινόμενων προϊόντων που εμφανίζονται στους πελάτες τους ("You may also like"), τεχνική που ακολουθούν πληθώρα εταιρειών σήμερα. Μέσα από την ανάλυση μεγάλου όγκου δεδομένων τόσο ηλεκτρονικών συναλλαγών όσο και κριτικών προϊόντων που υπάρχουν στο διαδίκτυο προσπαθεί να αναγνωρίσει τις πιθανές ανάγκες των πελατών της. Είναι αξιοσημείωτο το γεγονός ότι το 30% των πωλήσεων της Amazon προέρχονται από αυτό το εργαλείο. [1]

Τέλος, το Netflix κατηγοριοποίησε τους πελάτες τους ανάλογα με τις προτιμήσεις τους (π.χ. περιπέτειες, θρίλερ κ.α.) αναλύοντας παραπάνω από ένα δισεκατομμύριο reviews τα οποία χώρισε σε κατηγορίες (π.χ. liked, loved, hated κ.α.). Με τα δεδομένα αυτά και χρησιμοποιώντας προηγμένα εργαλεία προβλέψεων κατά απαίτηση δημιούργησε την επιτυχημένη σειρά "House of Cards" προσαρμόζοντας την στις επιθυμίες και τις προτιμήσεις των θεατών. [1]
Ακολουθούν κάποιες τεχνικές βελτίωσης ενός διαδικτυακού τόπου που δραστηριοποιείται στον κλάδο του ecommerce:

1. <u>Προσθήκη συνωνύμων λέξεων στην μηχανή αναζήτησης του ιστοτόπου:</u> Όταν ο χρήστης πραγματοποιεί μια αναζήτηση σε μια ιστοσελίδα ο αλγόριθμος που υπολογίζει τα αποτελέσματα λαμβάνοντας υπόψιν του και συνώνυμα της λέξης αναζήτησης αυτομάτως βελτιώνει τα επιστρεφόμενα αποτελέσματα. Οι ερευνητές έχουν αναπτύξει διάφορες τεχνικές για την αυτόματη παραγωγή «θησαυρών» λέξεων οι οποίες χρησιμοποιούνται στην ενίσχυση των ερωτημάτων αναζήτησης. Οι περισσότερες από αυτές τις μεθόδους βασίζονται στην ανάλυση των επαναλήψεων εμφάνισης λέξεων μέσα σε έγγραφα από την οποία παράγονται ισχυρές λίστες σχετικών μεταξύ τους λέξεων.
2. <u>Διαμόρφωση κατηγοριών και υποκατηγοριών χαρακτηριστικών προϊόντων/υπηρεσιών:</u> Βέλτιστη πρακτική αποτελεί το γεγονός οι χρήστες να μην περιορίζουν τις προτιμήσεις τους ώστε να μην μειώνουν αλλά να διευρύνουν το εύρος της αναζήτησης τους και των αποτελεσμάτων που λαμβάνουν αντίστοιχα. Μια τεχνική θα αποτελούσε το User Interface να ζητάει τη συναίνεση του χρήστη για την πιθανή αντιστοίχιση ανάμεσα σε χαρακτηριστικά των προϊόντων. Π.χ. σε ένα περιβάλλον RDF αντίστοιχα αξιώματα μπορούν εύκολα να αντιστοιχιστούν σε υπάρχοντα δεδομένα, όπως συγκεκριμένα γραφήματα RDF τα οποία μπορούν να διαμορφωθούν ακόμα και αποκλειστικά για τον κάθε χρήστη που πραγματοποιεί μια αναζήτηση. [5]
3. <u>Διαμόρφωση λεξικών όρων (σε μορφή γραφημάτων):</u> Με τη μαζική εμφάνιση πηγών Ανοιχτών Διασυνδεδεμένων Δεδομένων, όπως η DBpedia, έχει παρουσιαστεί το ενδιαφέρον διαμόρφωσης γράφων που αναπαριστούν κομμάτια της γενικευμένης παγκόσμιας γνώσης όπως ο γράφος Google Knowledge τον οποίον εισήγαγε το 2012 η Google. Η τεχνική αυτή έχει αποδειχθεί ότι είναι ο καλύτερος τρόπος παραγωγής δομημένων δεδομένων για χρήση σε μηχανές αναζήτησης, ακόμη και σε προηγμένα εργαλεία προσωπικών βοηθών (intelligent personal assistant).



Εικόνα 6 : Αναπαράσταση των πηγών δεδομένων του Google Knowledge Graph [25]

4. Ανάλυση χαρακτηριστικών των προϊόντων/υπηρεσιών ανάλογα με τη σημαντικότητά τους: Η χρήση πληροφοριών που προέρχονται από ετικέτες (tags) χαρακτηριστικών μπορούν να ενισχύσουν την αποτελεσματικότητα της ταξινόμησης και του τρόπου εμφάνισης των προϊόντων.
5. Βαθμολογίες και σχόλια: Αντλούνται από την ροή της πληροφορίας με χρήση ερωτημάτων SPARQL σε δεδομένα RDF. Η τακτική αυτή μετατρέπει τους ικανοποιημένους πελάτες σε "influencers" οι οποίοι διαδίδουν πληροφορίες για τα προϊόντα/ υπηρεσίες. Αποτελεί τη βάση για παραγωγή προσωποποιημένων αποτελεσμάτων αναζήτησης και προτάσεων τα οποία βασίζονται στα μέλη ενός κοινωνικού δικτύου που είναι περισσότερο δημοφιλή και είναι πιθανότερο να εμπιστευτεί κάποιος τη γνώμη τους με βάση ένα συγκεκριμένο σενάριο. Αντίθετα, οι συμβατικές υπηρεσίες αξιολόγησης που προσφέρει μια ιστοσελίδα έχουν ορισμένους περιορισμούς. Προέρχονται από ένα πιο περιορισμένο κοινό, εστιάζουν σε συγκεκριμένα προϊόντα ή προϊόντα συγκεκριμένης κατηγορίας τα οποία διαθέτει μια συγκεκριμένη εταιρεία και τα οποία είναι καταγεγραμμένα σε μια συγκεκριμένη πλατφόρμα (ιστοσελίδα) καταχώρησης βαθμολογιών και σχολίων. [5]

Στο σημείο αυτό αξίζει να αναφέρουμε ορισμένα χαρακτηριστικά ενός ηλεκτρονικού καταστήματος στα οποία η πλειοψηφία των online καταναλωτών δείχνουν ενδιαφέρον:

1. Περιγραφή των προϊόντων/υπηρεσιών χρησιμοποιώντας συνώνυμα και οικογένειες λέξεων οι οποίες ανήκουν στην ίδια οντότητα.

2. Εμφάνιση των αποτελεσμάτων αναζήτησης ενός διαδικτυακού καταστήματος ομαδοποιημένα με βάση της περιγραφές τους σε επιμέρους κατηγορίες και υποκατηγορίες.
3. Οι περιγραφές των προϊόντων να συνοδεύονται από επιπλέον πληροφορίες οι οποίες αναφέρονται σε όρους διαδικτυακών λεξικών, όπως αυτές που χρησιμοποιούνται στις μηχανές αναζήτησης που αναφέρθηκαν προηγουμένως, μέσω των οποίων ο χρήστης μπορεί να αντλήσει επιπλέον πληροφορίες για το προϊόν/υπηρεσία που τον ενδιαφέρει.
4. Επιθυμούν τα χαρακτηριστικά το προϊόντων να κατηγοριοποιούνται και να ταξινομούνται με βάση την σημαντικότητά τους (στοιχεία από βαθμολογίες και σχόλια) όταν εμφανίζονται στους χρήστες.

2.9.1 Είδη δεδομένων στο ecommerce

Στον κλάδο του ecommerce τα δεδομένα αποτελούν το κλειδί για την ανάλυση της αγοραστικής συμπεριφοράς των καταναλωτών ώστε μια εταιρεία που δραστηριοποιείται στον κλάδο του ecommerce να προσφέρει στους πελάτες της προσωποποιημένες προσφορές μέσα από την ανάλυση δεδομένων που συλλέγονται και αναλύονται σε πραγματικό χρόνο. Τα δεδομένα που χειρίζονται οι εταιρείες του κλάδου του ecommerce είναι δύο ειδών:

- Δομημένα: δημογραφικά στοιχεία, ονομ/νυμο, ηλικία, φύλο, ημ/νια γέννησης, διεύθυνση, προτιμήσεις κ.α.
- Αδόμητα: κλικς, tweets, ήχος, video κ.α. [1]

Πρόκληση αποτελεί ο συνδυασμός τους με σκοπό τη διεξαγωγή χρήσιμων συμπερασμάτων ώστε να βελτιώσουμε την απόδοση και την κερδοφορία ενός ecommerce site μέσα από την αύξηση του conversion rate. [1]
Μια ακόμη κατηγορία δεδομένων αποτελούν τα δεδομένα ροής που προέρχονται από τα κλικ των χρηστών στο διαδίκτυο. Συνήθως προέρχονται από τα social media (tweets, Facebook posts, Instagram, blog posts κ.α.) ή διαδικτυακές διαφημίσεις. [1]

Μια διαφορετική προσέγγιση διαχωρίζει τα δεδομένα στις παρακάτω κατηγορίες:

- Τα ίδια τα δεδομένα: posts, likes, shares, hashtags, users κ.α. Υπάρχει η δυνατότητα υπολογισμού της εμβέλειας τους και των ποσοτικών προσδιοριστικών χαρακτηριστικών τους καθώς και ανάλυσης της δύναμής τους.
- Τα μεταδεδομένα τους (metadata): Πολλοί υποστηρίζουν ότι είναι πιο σημαντικά από τα ίδια τα δεδομένα. Περιέχουν πληροφορίες όπως π.χ. ποιος είναι ο χρήστης που το πραγματοποίησε κάποιο post ή share; Ποια είναι η τοποθεσία του; Με ποιους χρήστες είναι συνδεδεμένος; Ποια είναι η επιρροή τους και πόσο ενεργοί είναι; Ποια είναι τα πρότυπα συμπεριφοράς τους και ποιες είναι οι προτιμήσεις τους; [2]

Γίνονται αξιόλογες προσπάθειες για τον χειρισμό της γνώσης η οποία μπορεί να συμβάλει στο συντονισμό των διαδικασιών ηλεκτρονικού επιχειρείν μέσα από την αξιοποίηση των τεχνολογιών σημασιολογικού ιστού. Οι τεχνολογίες αυτές

θεωρούνται οι σημαντικότερες και περισσότερο χρησιμοποιούμενες μέσα από την χρήση οντοτήτων οι οποίες βοηθούν στην ανταλλαγή γνώσεων, την ολοκλήρωση και διαλειτουργικότητα των βάσεων δεδομένων. [5]

Τη σημαντικότερη οντότητα του σημασιολογικού ιστού αποτελούν τα Ανοιχτά Διασυνδεδεμένα Δεδομένα η οποία επιτρέπει τη δημιουργία συνδέσμων μεταξύ εννοιών που ανήκουν σε διαφορετικά σύνολα δεδομένων. [5] Σήμερα η παραγωγή των Linked Open Data έχει αυξηθεί λόγω της ραγδαίας εξάπλωσης των μέσων κοινωνικής δικτύωσης, της αύξησης της χρήσης του διαδικτύου, των smartphones και όλων των άλλων τεχνολογιών μαζικής αποθήκευσης και παραγωγής δεδομένων. Ένας ακόμη παράγοντας που συνέβαλε σε αυτό είναι η μείωση του κόστους του αποθηκευτικού χώρου, της υπολογιστικής ισχύς και των νέων εργαλείων ανάλυσης των δεδομένων που βοηθούν της εταιρείες ecommerce να εντάξουν τα Open Data στην διαδικασία λήψης αποφάσεων. [1]

Τα χαρακτηριστικά των δεδομένων στο ecommerce είναι:

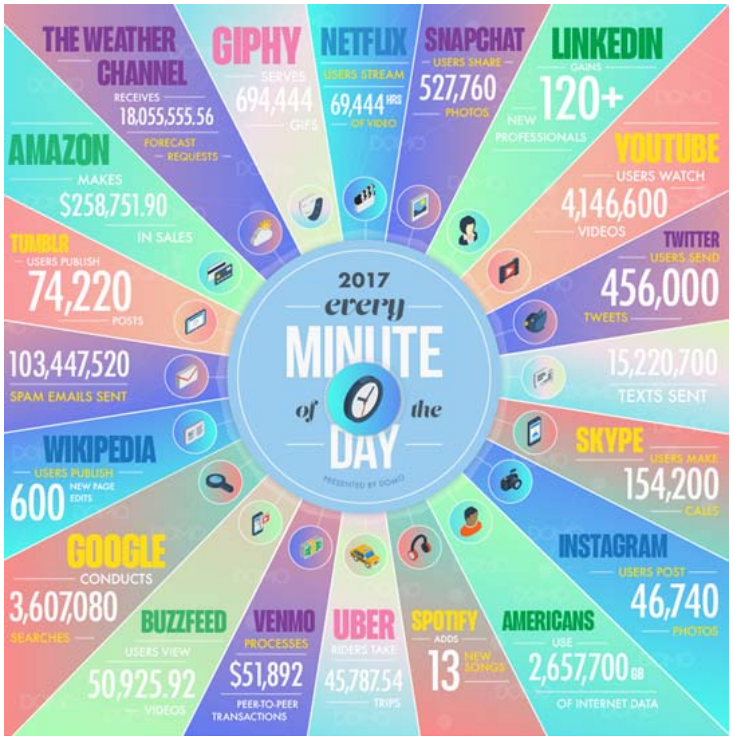
1. Μεγάλος όγκος δεδομένων: Οι εταιρείες προσπαθούν να συλλέξουν μεγάλο όγκο δεδομένων προκειμένου να βελτιώσουν τη διαδικασία λήψης αποφάσεων. Γνωστές εταιρείες, όπως η Amazon, eBay, Expedia, Travelocity κ.α. χρησιμοποιούν μαζικά δεδομένα κοινωνικών δικτύων (φωτογραφίες, video, posts, weblinks, newsfeeds κ.α.) τα οποία αναλύουν σε πραγματικό χρόνο προκειμένου να εμφανίσουν διαφημίσεις και προωθητικές ενέργειες στους επισκέπτες των διαδικτυακών τους τόπων. Ωστόσο, ο μεγάλος όγκος των δεδομένων εκτός από πλεονεκτήματα έχει φέρει και μειονεκτήματα καθώς η ενοποίηση διαφορετικών μορφών δεδομένων που προέρχονται από διαφορετικές πηγές απαιτεί την χρήση κοστοβόρων και χρονοβόρων προηγμένων μεθόδων ανάλυσης και τεχνικών machine learning.
2. Ποικιλία δεδομένων: Όπως ήδη αναφέρθηκε, τα δεδομένα παράγονται από πολλαπλές πηγές και σε διαφορετικές μορφές οι οποίες περιέχουν πολυδιάστατα πεδία δεδομένων. Τα μοντέλα ανάλυσης που χρησιμοποιούνται στο ecommerce περιλαμβάνουν μεγάλη ποικιλία δομημένων δεδομένων, όπως τα προφίλ των πελατών, την ιστορικότητα των συναλλαγών τους, την εποχικότητα των αγορών αλλά και αδόμητων, όπως στοιχεία από social media για την επίβλεψη της δημοφιλίας των προϊόντων ή τα αποτελέσματα προωθητικών ενεργειών. Για την αποδοτική χρήση των δεδομένων αυτών απαιτείται δέσμευση από την πλευρά του management της εταιρείας ώστε να βελτιωθούν οι επιχειρηματικές διαδικασίες για την παραγωγή σωστών αποτελεσμάτων και αναφορών.
3. Ταχύτητα παραγωγής δεδομένων: Πολλές εταιρείες του κλάδου του ecommerce υιοθέτησαν προηγμένα συστήματα εξόρυξης δεδομένων προκειμένου να εντοπίσουν, να αποθηκεύσουν και να αναλύσουν τα δεδομένα σε πραγματικό χρόνο ώστε να λάβουν τις κατάλληλες αποφάσεις που θα τους παρέχουν ανταγωνιστικό πλεονέκτημα.
4. Εγκυρότητα των δεδομένων: Πολύ σημαντικό παράγοντα αποτελεί η ικανότητα εντοπισμού έγκυρων, πιστοποιημένων και επεξεργάσιμων δεδομένων, κάτι το

οποίο λύνεται με την χρήση των Linked Open Data, απαλλαγμένων από "προβληματική" πληροφορία. [1]

2.9.2 Ο ρόλος των social media στο ecommerce

Όπως βλέπουμε και στο παρακάτω γραφικό που παρουσίασε η εταιρεία Domo στο Data Never Sleeps 5.0 το 2017 ο αριθμός των δεδομένων που παράγονται κάθε λεπτό στα Social Media ξεπερνά πλέον κάθε προσδοκία. Το Facebook πλέον μετράει 2 δισεκατομμύρια χρήστες, από αυτούς το 1.5 δισεκατομμύριο είναι ενεργοί καθημερινά. 5 νέα προφίλ δημιουργούνται κάθε δευτερόλεπτο και 300 εκατομμύρια εικόνες "ανεβαίνουν" καθημερινά. Κάθε λεπτό δημοσιεύονται 510.000 posts και 293.000 status updates. Παρότι το Facebook αποτελεί το μεγαλύτερο μέσο κοινωνικής δικτύωσης, το Instagram είναι αυτό που αναπτύσσεται με μεγαλύτερους ρυθμούς. Υπάρχουν 600 εκ. προφίλ στο Instagram με τα 400 εκ. από αυτά να είναι καθημερινά ενεργά. 56 εκ. εικόνες και video να ανεβαίνουν καθημερινά και 100 εκ. να χρησιμοποιούν το εργαλείο "insta stories". [22]

Τα δεδομένα αυτά παίζουν σημαντικό ρόλο στην διαμόρφωση της στρατηγικής προώθησης μιας επιχείρησης που δραστηριοποιείται στον κλάδο του ecommerce καθώς μπορούν να χρησιμοποιηθούν για την πρόβλεψη των προτιμήσεων των πελατών καθώς και στην λήψη στρατηγικών αποφάσεων.



Εικόνα 7 : Ο αριθμός των δεδομένων που παράγονταν κάθε λεπτό στα Social Media το 2017

Δεν υπάρχει κάποιο αυτόνομο εργαλείο/σύστημα το οποίο μπορεί να εφαρμοστεί για τον σκοπό αυτό, ωστόσο υπάρχει ένα σύνολο open source εργαλείων που βασίζονται

σε ανοικτές υπηρεσίες διαδικτύου και μεθοδολογίες οι οποίες συνδυάζονται για την κάλυψη των αναγκών της εκάστοτε εφαρμογής. [13]

2.10 Ο ρόλος των Social Media στον Σημασιολογικό Ιστό

Η συμβολή των τεχνολογιών σημασιολογικού ιστού στην διαδικασία ανάλυσης των δεδομένων από τα Social Media είναι καθοριστικής σημασίας καθώς αυτόματα αναγνωρισμένες οντότητες και θέματα που προέρχονται από τα Social Media αποσαφηνίζονται και συνδέονται με πόρους των Open Data με την χρήση των URIs (π.χ. DBpedia, GeoNames). Επίσης, χρησιμοποιούνται γνώσεις του σημασιολογικού ιστού για την πραγματοποίηση σε βάθος αναζήτησης στα κείμενα της ροής των social media και αναζητούνται οι επαναλαμβανόμενες εμφανίσεις μιας λέξης σε κείμενα καθώς και οι αντίστοιχες οντότητες που προκύπτουν από την αναζήτηση αυτή. Τα βήματα που ακολουθούνται για την ανάπτυξη μιας τέτοιας εφαρμογής είναι η συλλογή των δεδομένων, η ανάλυση, η συνένωση, η αναζήτηση στον σημασιολογικό ιστό και η χρήση εργαλείων οπτικοποίησης των αποτελεσμάτων. Η μέθοδος αυτή επιτρέπει τη σε βάθος ανάλυση των δεδομένων και μπορεί να εφαρμοστεί για την απάντηση περίπλοκων ερωτημάτων που βασίζονται σε εκατομμύρια posts σε πραγματικό χρόνο καθώς τα εργαλεία οπτικοποίησης ανακαλύπτουν νέους συσχετισμούς ανάμεσα στα δεδομένα. [13]

Με σκοπό την περαιτέρω κατανόηση του τρόπου λειτουργίας των Social Media και πως αυτός αξιοποιείται από τον σημασιολογικό ιστό ακολουθεί μια κατηγοριοποίηση των Social Media με βάση τους συνδέσμους που αναπτύσσονται ανάμεσα στους χρήστες, τους τρόπους που μοιράζεται η πληροφορία καθώς και το πως οι χρήστες αντιδρούν στη ροή των media:

1. Μέσα που αποτελούνται από γράφους με βάση τα ενδιαφέροντα των χρηστών: πχ Twitter, οι σύνδεσμοι μεταξύ των χρηστών σχηματίζονται με βάση τα ενδιαφέροντα τους, ανεξαρτήτως με το αν γνωρίζονται στην προσωπική τους ζωή. Οι συνδέσεις δεν είναι αμοιβαίες και οι πληροφορία που μοιράζεται εμφανίζεται με μορφή μηνυμάτων σε χρονολογική σειρά.
2. Σελίδες κοινωνικής δικτύωσης: π.χ. Facebook, ενθαρρύνουν τους χρήστες να συνδέονται με άτομα που γνωρίζουν και στην προσωπική τους ζωή παρέχοντας έναν τρόπο δημοσίευσης πληροφοριών αλλά και δυνατότητα σχολιασμού posts άλλων χρηστών. Οι χρήστες μοιράζονται γεγονότα που συμβαίνουν στην προσωπική τους ζωή ή συνδέονται με πληροφορίες τις οποίες πιστεύουν ότι οι φίλοι τους θα βρουν ενδιαφέρουσες. Αυτές οι ενημερώσεις των status των χρηστών συνδυάζονται και εμφανίζονται στους χρήστες σε μια ροή με βάση μια χρονολογική σειρά.
3. Σελίδες επαγγελματικής δικτύωσης: πχ LinkedIn, χρησιμοποιούνται από κάποιον χρήστη ώστε να συστηθεί επαγγελματικά σε άτομα που τον ενδιαφέρουν. Όταν κάποιος συνδέεται με ένα άτομο σημαίνει ότι το γνωρίζει ως ένα βαθμό και ότι θα πρότεινε σε άλλους να επικοινωνήσουν μαζί του για επαγγελματικό σκοπό.

4. Υπηρεσίες για μοίρασμα περιεχομένου και πληροφορίας: όπως blogs ή ιστοσελίδες video sharing (π.χ. YouTube, Vimeo) και forums με συζητήσεις/αξιολογήσεις χρηστών (πχ. CNET). Συνήθως αναφερόμαστε σε blogs στα οποία παράγονται ροές άρθρων για να διαβάσουν και να σχολιάσουν οι επισκέπτες τους. Συνηθίζεται, πολλοί χρήστες και ιστοσελίδες που δημοσιεύουν τέτοια άρθρα να διαφημίζουν τις αναρτήσεις τους μέσα από posts στο Facebook ή στο Twitter. [16]

   Το περιεχόμενο των ροών των περισσότερων από τα social media που αναφέρθηκαν παραπάνω είναι σύντομες και ιδιαίτερα θορυβώδεις οδηγώντας συχνά σε αποτελέσματα χαμηλότερης ποιότητας. Τα παρακάτω χαρακτηριστικά των Social Media αποτελούν κίνητρο για την ανάπτυξη νέων τεχνολογιών σημασιολογικού ιστού οι οποίες θα ταιριάζουν καλύτερα στις ροές των μέσων αυτών και θα παράγουν χρήσιμα αποτελέσματα: [16]

- Σύντομα κείμενα (microtexts): Τα περισσότερα μηνύματα που δημοσιεύονται στο Facebook και στο Twitter (140 χαρακτήρες) είναι πολύ μικρά σε έκταση. Επομένως υπάρχει ανάγκη για ανάπτυξη μεθόδων σημασιολογικού ιστού οι οποίες εμπλουτίζουν το περιεχόμενο τους με επιπλέον πληροφορίες οι οποίες προέρχονται από ενσωματωμένα URLs και hashtags.
- Θορυβώδες περιεχόμενο: Αναφερόμαστε σε ασυνήθιστη ορθογραφία (όλα μικρά ή κεφαλαία γράμματα, emoticons κα.). Έχουν αναπτυχθεί μέθοδοι κανονικοποίησης/βελτιστοποίησης του κειμένου σε συνδυασμό με μελέτες που γίνονται στις παραλλαγές των γλωσσικών στιλ ανά τοποθεσία. Τέλος, τα emoticons συχνά χρησιμοποιούνται ως ένδειξη συναισθημάτων σε αλγορίθμους εξόρυξης της γνώμης του κοινού.
- Παροδικότητα: Το περιεχόμενο των social media δεν ενδείκνυται για χρονική ανάλυση. Για την αντιμετώπιση της προσωρινότητας απαιτούνται νέα μοντέλα τα οποία παρακολουθούν τις αλλαγές στα ενδιαφέροντα του χρήστη με την πάροδο του χρόνου.
- Το Κοινωνικό πλαίσιο είναι απαραίτητο για την σωστή ερμηνεία του περιεχομένου των social media: Οι μέθοδοι που βασίζονται στον σημασιολογικό ιστό θα πρέπει να λαμβάνουν υπόψη τους το εκάστοτε κοινωνικό πλαίσιο προκειμένου να παράγουν αυτόματα σημασιολογικά μοντέλα των κοινωνικών δικτύων τα οποία θα μετρούν την σημαντικότητα του χρήστη, θα ομαδοποιούν παρόμοιους χρήστες κ.α.
- Περιεχόμενο παραγόμενο από τους χρήστες: Οι χρήστες παράγουν και καταναλώνουν περιεχόμενο. Υπάρχει μια πλούσια πηγή πληροφοριών για κάθε χρήστη όπως τα δημογραφικά χαρακτηριστικά, το φύλο, η τοποθεσία, η ηλικία κ.α. αλλά και τα ενδιαφέροντα και τις απόψεις του. Πρόκληση αποτελεί το γεγονός ότι σε ορισμένες περιπτώσεις το περιεχόμενο που δημιουργείται από έναν χρήστη δεν επαρκεί και έτσι οι αλγόριθμοι που βασίζονται σε αυτό δε μπορούν να εφαρμοστούν σωστά.
- Πολυγλωσσικότητα: Σύμφωνα με έρευνες λιγότερο από το 50% των tweets είναι στα Αγγλικά με τα Ιαπωνικά, Ισπανικά, Πορτογαλικά και Γερμανικά να ακολουθούν. Δυστυχώς οι περισσότερες σημασιολογικές μέθοδοι βασίζονται στα Αγγλικά και έτσι η αυτόματη αναγνώριση της γλώσσας αποτελεί το πρώτο βήμα ώστε οι εφαρμογές να

ομαδοποιήσουν τα κοινωνικά δίκτυα σε γλωσσικές ομάδες οι οποίες μπορούν να επεξεργαστούν με χρήση διαφορετικών αλγορίθμων. [16]

- Ασάφεια: Αποτελεί το μεγαλύτερο πρόβλημα καθώς δεν μπορούμε εύκολα να συσχετίσουμε τις πληροφορίες που περιλαμβάνονται στα Social Media με ένα συγκεκριμένο θέμα. Σε αντίθεση με τα blog posts τα tweets συνήθως δεν ακολουθούν τη ροή μιας συζήτησης αλλά εμφανίζονται ανεξάρτητα. Τέλος, περιέχουν στοιχεία σαρκασμού και ειρωνείας τα οποία είναι δύσκολα ανιχνεύσιμα από υπολογιστές. Είναι συχνό φαινόμενο ένα tweet να σχετίζεται με παραπάνω του ενός θέματα. Δίνοντας έμφαση σε θέματα τα οποία μπορεί να είναι παρεμφερή με αυτά που περιέχει το tweet ή σε θέματα που συνδέονται μεταξύ τους μπορούμε να βοηθήσουμε στην αποσαφήνιση του νοήματος του tweet. [16]
  Λόγω της σύντομης φύσης τους τα μηνύματα σε Facebook και twitter συχνά δεν μπορούν να γίνουν κατανοητά χωρίς της χρήση εξωτερικών αναφορών στο περιεχόμενο τους. Από το σύνολο των posts ένα ελάχιστο ποσοστό περιλαμβάνουν έτοιμα ULRs. Γι αυτό το λόγο γίνεται χρήση τεχνικών εμπλουτισμού του περιεχομένου των μηνυμάτων ακόμη και με χρήση πολυμέσων. Π.χ στα tweets που δεν υπάρχουν URLs χρησιμοποιούνται αλγόριθμοι TF-IDF (term frequency–inverse document frequency) για τον εντοπισμό ομοιοτήτων ανάμεσα στο περιεχόμενο των tweets και νέα σε posts, hashtags κ.α. Στόχος είναι ο εντοπισμός ομοιοτήτων που βασίζονται σε οντότητες του σημασιολογικού ιστού, όπως π.χ. η χρήση του εργαλείου OpenCalais για την αναγνώριση σημασιολογικών οντοτήτων και θεμάτων. [16]

Συμπερασματικά, ο καλύτερος τρόπος βελτίωσης του σημασιολογικού σχολιασμού των social media είναι να κάνουμε καλύτερη χρήση της απεριόριστης γνώσης που υπάρχει στο Διαδίκτυο των Δεδομένων. Αυτή τη στιγμή η γνώση περιορίζεται στην Wikipedia και από πόρους που αντλούνται από αυτή (DBpedia, YAGO). [16]

2.10.1 Οι γνωστότερες οντολογίες για την αναπαράσταση περιεχομένου Social Media

Ακολουθούν οι σημαντικότερες οντολογίες οι οποίες αναπαριστούν τη σημασιολογία των social media:

- FOAF (Friend-of-a-Friend): Περιγράφει ανθρώπους και Κοινωνικά Δίκτυα. Περιλαμβάνει λεξιλόγιο για προσδιορισμό προσώπων, όπως ονόματα, πληροφορίες επικοινωνίας, δραστηριότητες και τις σχέσεις ενός ατόμου με άλλους ανθρώπους ή αντικείμενα. Επιτρέπει σε ένα γκρουπ ανθρώπων να αναπαραστήσουν ένα κοινωνικό δίκτυο χωρίς να χρειάζεται κεντρική βάση δεδομένων.
- SIOC (Semantically Interlinked Online Communities): Μοντελοποιεί σελίδες Κοινωνικών Δικτύων όπως blogs, wikis, online forums κ.α. Βασικές έννοιες αποτελούν τα forums, sites, posts, user accounts, user groups και tags. Επίσης, μπορεί να χρησιμοποιηθεί για μοντελοποίηση των ενδιαφερόντων των

χρηστών με την χρήση της ιδιότητας sioc:topic η οποία λαμβάνει ένα URI ως τιμή.

- Μοντελοποίηση Microblogs: στην οντολογία SIOC έχουν προστεθεί έννοιες όπως το MicroblogPost, οι ιδιότητες sioc:follows, sioc:addressed_to για posts που περιέχουν έναν συγκεκριμένο όνομα. Η οντολογία Bottari σχεδιάστηκε αποκλειστικά για να περιγράψει σχέσεις στο Twitter, ειδικά για την σύνδεση μεταξύ tweets, τις τοποθεσίες και τα συναισθήματα των χρηστών. Τέλος, έχει δημιουργηθεί μια νέα κλάση, η TwitterUser που περιλαμβάνει τις ιδιότητες follower και following.
- DLPO (Live PostOntology): Χρησιμοποιείται για τη διασύνδεση Κοινωνικών Μέσων, Κοινωνικών Δικτύων και πρακτικών Online Sharing. Η οντολογία περιλαμβάνει ένα περιεκτικό μοντέλο για τα social media posts. Πλαισιώνεται από κύριες οντολογίες όπως οι FOAF, SOIC και Simple Knowledge Organisation System (SKOS). Μοντελοποιεί πληροφορίες για τους χρήστες οι οποίες προκύπτουν από το περιεχόμενο των social media όπως επίσης συνδέει posts ανάμεσα σε προσωπικά social networks. Η οντολογία καταγράφει 6 βασικους τύπους πληροφορίας: online posts, διαφορετικά είδη posts (π.χ. retweets), microposts, online παρουσία, φυσική παρουσία και πρακτικές online sharing. Η συμπεριφορά του κάθε χρήστη καθώς και τα μεμονωμένα χαρακτηριστικά τους μοντελοποιούνται με τη χρήση της οντολογίας SWUM στην οποία θα αναφερθούμε παρακάτω. [16]

Ειδική κατηγορία αποτελούν οι οντολογίες για τη μοντελοποίηση χρηστών οι οποίες αποτελούν το κλειδί για την συγκέντρωση, αναπαράσταση και μοίρασμα των πληροφοριών που αφορούν τους χρήστες και τις αλληλεπιδράσεις τους στα social media. Μια τέτοια οντολογία είναι η General User Modelling Ontology (GUMO) η οποία καλύπτει ένα ευρύ φάσμα πληροφοριών σχετικών με τους χρήστες. Ωστόσο, δεν αναπαριστά τα ενδιαφέροντα των χρηστών καθιστώντας την ακατάλληλη για τα social media. [16]

Έπειτα από ανάλυση 17 κοινωνικών διαδικτυακών εφαρμογών οι Plumbaum et al κατέληξαν σε ένα σύνολο διαστάσεων που χρειάζεται να έχει μια οντολογία μοντελοποίησης χρηστών διαδικτύου. Οι διαστάσεις αυτές περιλαμβάνουν δημογραφικά χαρακτηριστικά, ανάγκες, στόχους, ψυχική και σωματική κατάσταση, γνώσεις και υπόβαθρο, συμπεριφορά και άλλα μοναδικά χαρακτηριστικά. Βασιζόμενοι σε αυτά δημιούργησαν την οντολογία SWUM (Social Web User Model). Μειονέκτημα της συγκεκριμένης οντολογίας είναι η αδυναμία σύνδεσης με άλλες οντολογίες. [16]

Τέλος, η Οντολογία Συμπεριφοράς Χρήστη μοντελοποιεί τις αλληλεπιδράσεις μεταξύ των χρηστών στις διαδικτυακές κοινότητες, όπως τα online forums και τις συζητήσεις στο Twitter. Περιλαμβάνει κλάσεις που μοντελοποιούν το αντίκτυπο των posts, της συμπεριφοράς των χρηστών, τους ρόλους τους, τα χρονικά πλαίσια και άλλες πληροφορίες αλληλεπίδρασης. Η αναπαράσταση της χρονικής διάστασης στα social

media είναι πολύ βασική κυρίως στην μοντελοποίηση των αλλαγών στη διάρκεια του χρόνου. [16]

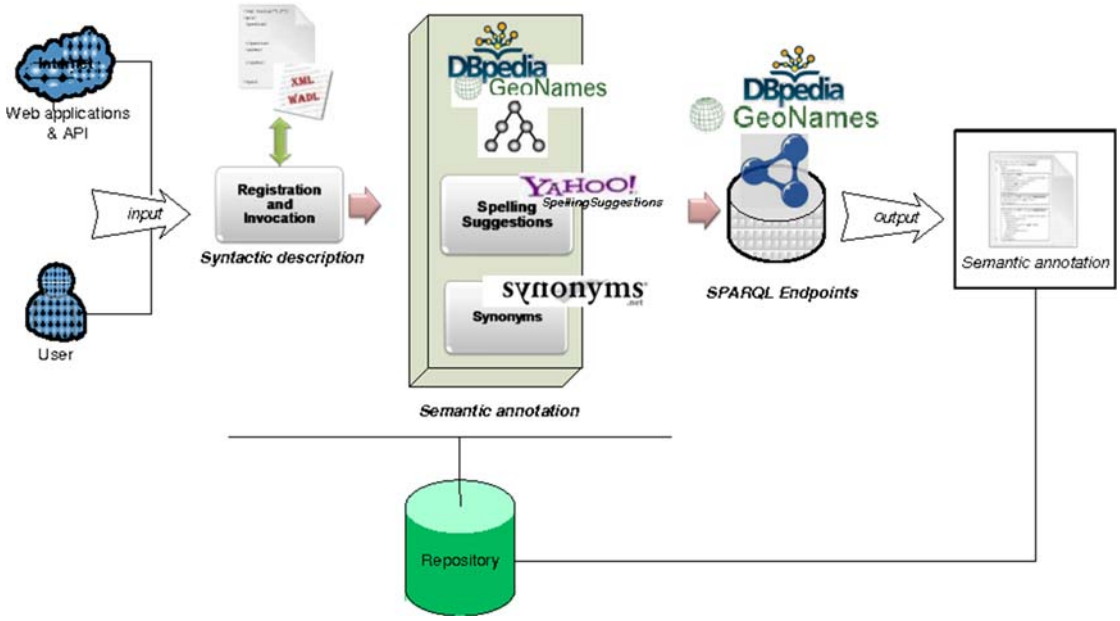
Η μέθοδος εξόρυξης keywords από το Twitter είναι αναποτελεσματική καθώς η θεματολογία του περιέχει πολλές περιττές πληροφορίες. Π.χ. Όσον αφορά την εξόρυξη των δημοφιλών θεμάτων στο Twitter μπορούμε να εντοπίσουμε λέξεις κλειδιά εκμεταλλευόμενοι τον πλεονασμό της πληροφορίας και τον εντοπισμό των συχνών εμφανίσεων ακολουθιών λέξεων. Ένα ακόμη χαρακτηριστικό το οποίο διευκολύνει σε λιγότερο βαθμό τη διαδικασία εξόρυξης είναι η μεγάλη ποικιλία θεμάτων που αναπτύσσονται. [16]

2.10.2 Αντιστοίχιση οντολογιών με περιεχόμενο από τα Social Media (Semantic Annotation)

Η διαδικασία αντιστοίχισης κλάσεων οντολογιών με οντότητες του περιεχομένου των social media με βάση τις γνωστές οντολογίες που αναφέρθηκαν παραπάνω περιλαμβάνει δύο φάσεις: 1)τον εντοπισμό των οντολογιών
και 2)τη σύνδεση τους με την οντότητα. Η διαδικασία αναφέρεται στον εντοπισμό όλων των αναφορών στο κείμενο των κλάσεων από τα Social Media και των αναφορών στην οντολογία (π.χ. DBpedia). Ακολουθεί η σύνδεση της οντότητας στην οποία χρησιμοποιούνται πληροφορίες από το κείμενο καθώς και γνώση από την οντολογία ώστε να επιλεχθεί το σωστό URI. [16]

Οι πρόσφατες έρευνες που σχετίζονται με την αναγνώριση και τη σύνδεση οντοτήτων από τα Social Media με γνωστές οντολογίες χρησιμοποιούν την Wikipedia ως μια μεγάλη και ελεύθερα προσβάσιμη βάση που έχει δημιουργηθεί από τους ίδιους τους χρήστες. Πιο συγκεκριμένα οι βάσεις που χρησιμοποιούνται είναι οι DBpedia και YAGO, οι οποίες προέρχονται από τη Wikipedia και προσφέρουν απευθείας συνδέσμους ανάμεσα σε URIs και στις αντίστοιχες σελίδες της Wikipedia. Για τη σύνδεση των οντοτήτων με τη βάση της Wikipedia επωφελούμαστε από τον μεγάλο όγκο πληροφορίας κειμένου που περιλαμβάνεται στα άρθρα και τις σελίδες της. [16]

Οι μέθοδοι αποσαφήνισης των οντοτήτων ουσιαστικά συλλέγουν ένα λεξικό ετικετών για κάθε οντότητα URI, χρησιμοποιώντας τις πληροφορίες των σελίδων που αφορούν την οντότητα στην Wikipedia. Στη συνέχεια ανακατευθύνουν τις σελίδες της οντότητας στο κείμενο αναφοράς πραγματοποιώντας μια σύνδεση με την αντίστοιχη σελίδα της Wikipedia. Το λεξικό αυτό των URIs χρησιμοποιείται στην αναγνώριση όλων των URIs για ένα συγκεκριμένο κείμενο αναφοράς. Στο στάδιο της αποσαφήνισης, ταξινομούνται όλα τα υποψήφια URIs και βαθμολογούνται με βάση τον βαθμό εγκυρότητας τους. Αν δεν υπάρχει οντότητα που να ταιριάζει με κάποιο URI της βάσης επιστρέφεται μια τιμή NIL. [16]



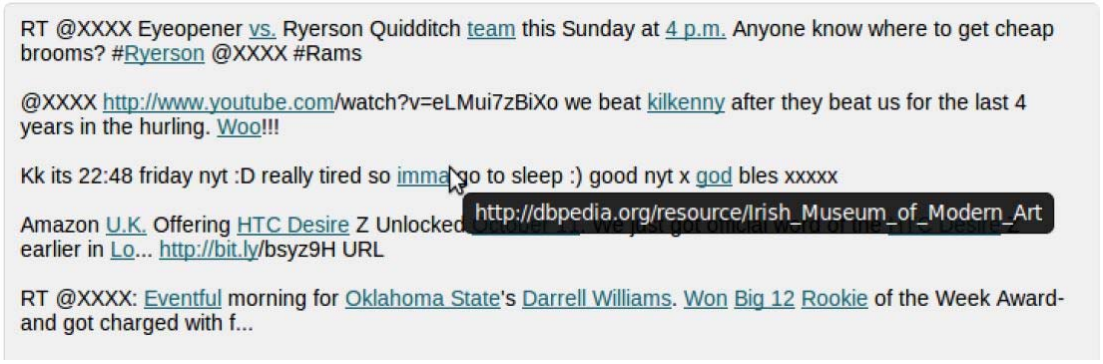
*Εικόνα 8 : Σύστημα Semantic Annotation με χρήση RESTFul API *[26]*

Οι αναφορές κειμένου μπορούν να αποσαφηνιστούν είτε ανεξάρτητα είτε ως μέρος ενός συνολικού κειμένου. Ουσιαστικά, οι μέθοδοι χρησιμοποιούν τα στατιστικά της Wikipedia σε συνδυασμό με τεχνικές που συνδυάζουν το περιεχόμενο των αναφορών κειμένου με τις σελίδες της Wikipedia για κάθε υποψήφια οντότητα. Οι τεχνικές αυτές μπορούν να χρησιμοποιηθούν για να εξαχθούν π.χ. από τα Tweets ενός χρήστη πληροφορίες για το προφίλ του με βάση τις κατηγορίες της Wikipedia. Η ακρίβεια των αλγορίθμων αυτών έχει αξιολογηθεί μέχρι στιγμής σε άρθρα της Wikipedia και σε νέες βάσεις δεδομένων, οι οποίες διαφέρουν στην φύση τους από τα σύντομα μηνύματα των social media streams. [16]

Μια συχνά χρησιμοποιούμενη σημασιολογική βάση είναι η DBpedia Spotlight. Πρόκειται για ένα ελεύθερα προσβάσιμο και παραμετροποιήσιμο σύστημα το οποίο αντιστοιχίζει έγγραφα κειμένου με URIs της DBpedia. Από το σύνολο των κλάσεων της DBpedia μπορούμε να περιορίσουμε ποιες κλάσεις θα χρησιμοποιηθούν στην αναγνώριση της οντότητας με την βοήθεια ερωτημάτων SPARQL. Ο αλγόριθμος πρώτα επιλέγει τις υποψήφιες οντότητες μέσω αναζήτησης σε ένα λεξικό URIs που προέρχεται από την Wikipedia, ακολουθεί το στάδιο ταξινόμησης τους με βάση ένα μοντέλο διανύσματος θέσης. Κάθε πόρος της DBpedia συσχετίζεται με ένα έγγραφο που κατασκευάζεται από όλες τις παραγράφους της Wikipedia στις οποίες αναφέρεται η συγκεκριμένη οντότητα. [16]

Στην παρακάτω εικόνα απεικονίζεται ένα απόσπασμα από Tweets το οποίο έχει αντιστοιχιστεί με URIs με τη χρήση του DBpedia Spotlight. Τα αποτελέσματα δείχνουν την ανάγκη για βελτιστοποίηση του κειμένου και της ορθογραφίας των tweets καθώς επίσης και τη δυσκολία αναγνώρισης των URLs. Ο αλγόριθμος έχει σχεδιαστεί για να αντιστοιχίζει όσο το δυνατό περισσότερες οντότητες χρησιμοποιώντας τα εκατομμύρια των πόρων της DBpedia. Το γεγονός αυτό

δεδομένης της σύντομης ισχύς και της "θορυβώδους" φύσης των tweets μπορεί να οδηγήσει σε ανακριβή αποτελέσματα. Επομένως, η περαιτέρω αξιολόγηση στο περιεχόμενο των μηνυμάτων των social media είναι απαραίτητη για τον καλύτερο καθορισμό των διαφόρων παραμέτρων του DBpedia Spotlight. [16]



Εικόνα 9 : Αντιστοίχιση Tweets με URIs με τη χρήση του DBpedia Spotlight

Ο αλγόριθμος LINDEN συνδέει οντότητες με έννοιες των βάσεων του Σημασιολογικού Ιστού και χρησιμοποιεί την πλουσιότερη σημασιολογική πληροφορία η οποία υπάρχει στην οντολογία YAGO (επιπλέον της πληροφορίας που βασίζεται στην Wikipedia). Η μέθοδος αυτή βασίζεται στο εργαλείο Wikipedia-Miner8, το οποίο χρησιμοποιείται για την ανάλυση αμφιλεγόμενων οντοτήτων και τον εντοπισμό εννοιών τις Wikipedia οι οποίες εμφανίζονται σε αυτό.
[16]

Ενώ κάθε μεμονωμένο post παρέχει ένα επαρκές σημασιολογικό πλαίσιο στο οποίο μπορούμε να βασιστούμε, επιπλέον πληροφορίες μπορούν να αντληθούν από τα προφίλ των χρηστών, τα κοινωνικά δίκτυα και τα αλληλένδετα posts (π.χ. απαντήσεις σε tweets).

Η διαδικασία του Semantic Annotation εμπεριέχει και κάποιους γνωστούς περιορισμούς όπως το πρόβλημα της ταξινόμησης των διαθέσιμων οντοτήτων. Το πρόβλημα αυτό μπορεί να λυθεί με τη χρήση της βάσης Freebase η οποία περιλαμβάνει ένα μεγάλο αριθμό γνωστών και ευρέως χρησιμοποιούμενων οντοτήτων. Το ποσοστό επιτυχίας της τακτικής αυτής είναι μόλις 38% καθώς το 35% των οντοτήτων είναι διφορούμενες, έχουν παραπάνω του ενός τύπους, ενώ το 30% των οντοτήτων των tweets δεν εμφανίζονται καν στη βάση Freebase. [16]

Ένα ακόμα πρόβλημα με το οποίο έχουν ασχοληθεί αρκετές έρευνες αποτελεί ο προσδιορισμός της τοποθεσίας των tags, όπως αυτά που περιέχουν οι πλατφόρμες Flickr και Instagram. Η προσέγγιση της λύσης βασίζεται στην σημασιολογική βάση Yahoo! GeoPlanet η οποία παρέχει ένα μοναδικό URI για κάθε αναφερόμενη τοποθεσία καθώς και μια ταξινόμηση των μεταξύ τους σχετικών ή γειτονικών τοποθεσιών. Η αποσαφήνιση της ετικέτας τοποθεσίας κάνει χρήση όλων των

ετικετών που υπάρχουν σε μια φωτογραφία, όλα τα tags των υπολοίπων φωτογραφιών του χρήστη καθώς και εκτεταμένο περιεχόμενο του χρήστη το οποίο λαμβάνει υπ' όψιν ακόμη και τα tags των επαφών του χρήστη. Η χρήση αυτού του γενικότερου κοινωνικού δικτύου βελτιώνει σημαντικά την συνολική ακρίβεια του προσδιορισμού της τοποθεσίας ενός tag. [16]

Μια ακόμα πηγή πρόσθετης πληροφορίας είναι τα Hashtags στα διάφορα κοινωνικά δίκτυα τα οποία βοηθούν τους χρήστες να παρακολουθήσουν τα posts που αναφέρονται σε ένα συγκεκριμένο θέμα. Οι Laniado & Mika ερεύνησαν τη σημασιολογία των hashtags σε παραπάνω από 369 εκατομμύρια μηνύματα χρησιμοποιώντας 4 μετρικές: συχνότητα χρήσης, ακρίβεια (χρήση του hashtag σε αντίθεση με την ίδια τη λέξη), συνέπεια της χρήσης και σταθερότητα με το πέρασμα του χρόνου. Αυτές οι μετρικές χρησιμοποιούνται για τον προσδιορισμό των hashtags τα οποία μπορούν να αποτελέσουν τη βάση για τη συνδεση με τα URIs της βάσης Freebase που αναφέρθηκε παραπάνω. Τα hashtags επίσης έχουν χρησιμοποιηθεί ως επιπλέον πηγή σημασιολογικής πληροφορίας για τα tweets προσθέτωντας σε αυτά επιπλέον ορισμούς hashtags οι οποίοι υπάρχουν σε πολυπληθή διαδικτυακά λεξικά. [16]

Τα χαρακτηριστικά που σχετίζονται με τους χρήστες και τους κοινωνικούς δεσμούς που αναπτύσσονται μεταξύ τους κωδικοποιούνται με την χρήση της οντολογίας FOAF, ενώ οι σημασιολογική αντιστοίχιση tags με τη χρήση της MOAT. [16]

Ontology	People	Online posts	Social networks	Micro blogs	User Interests	Tags	Geo-location	User Behaviour
FOAF	Yes		knows		Partial			
SIOC(T)	Yes	Yes		Partial	Yes			
MOAT						Yes		
Bottari	Yes	Yes	Yes	Yes		Yes	Yes	
DLPO	Yes	Yes	Yes	Yes	Yes	Yes		
SWUM	Yes				Yes		Yes	Yes
UBO	Yes		Yes		Yes			Yes

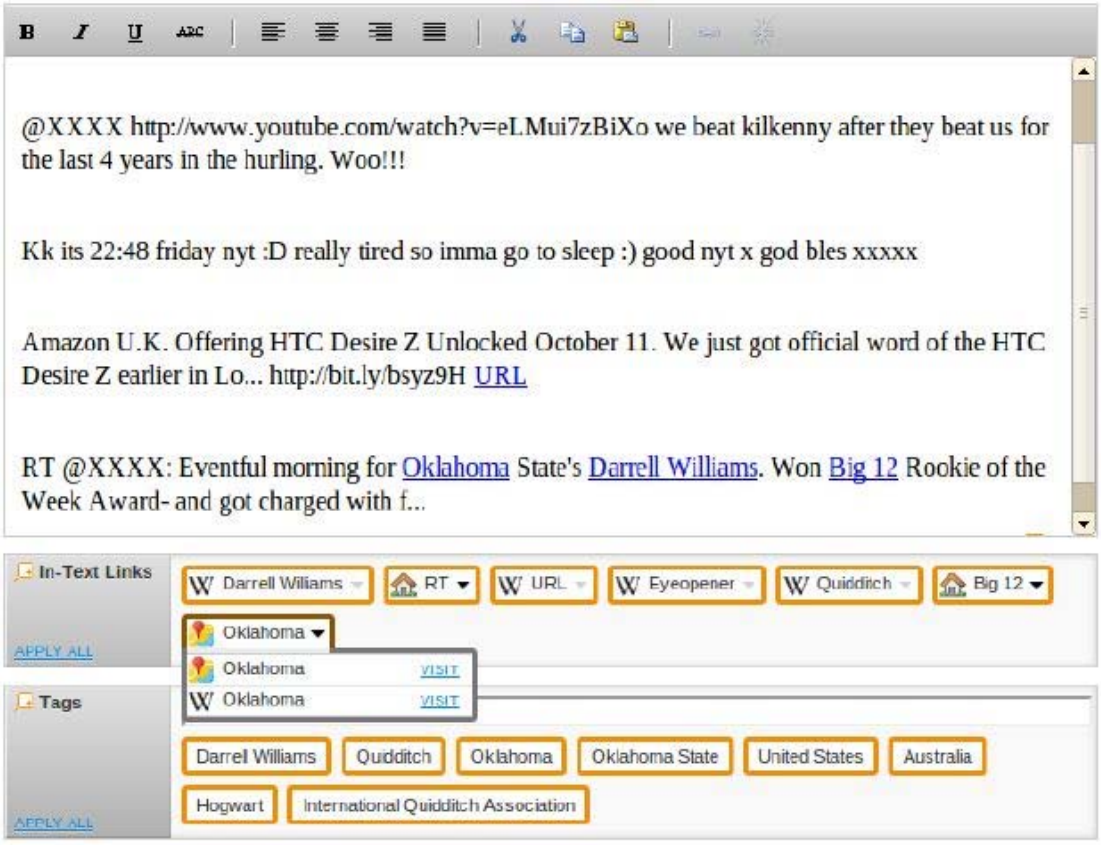
Πίνακας 2: Οι οντολογίες που χρησιμοποιούνται στα social media και τι μοντελοποιούν [16]

2.10.3 Εμπορικές διαδικτυακές υπηρεσίες semantic Annotation

Ακολουθούν μερικές από τις πιο γνωστές εμπορικές εφαρμογές που παρέχουν υπηρεσίες Αναγνώρισης Οντοτήτων. Οι εφαρμογές αυτές συνδέουν έγγραφα με

οντότητες και αντιστοιχίζουν URIs από τα ανοιχτά διασυνδεδεμένα δεδομένα σε αυτά.

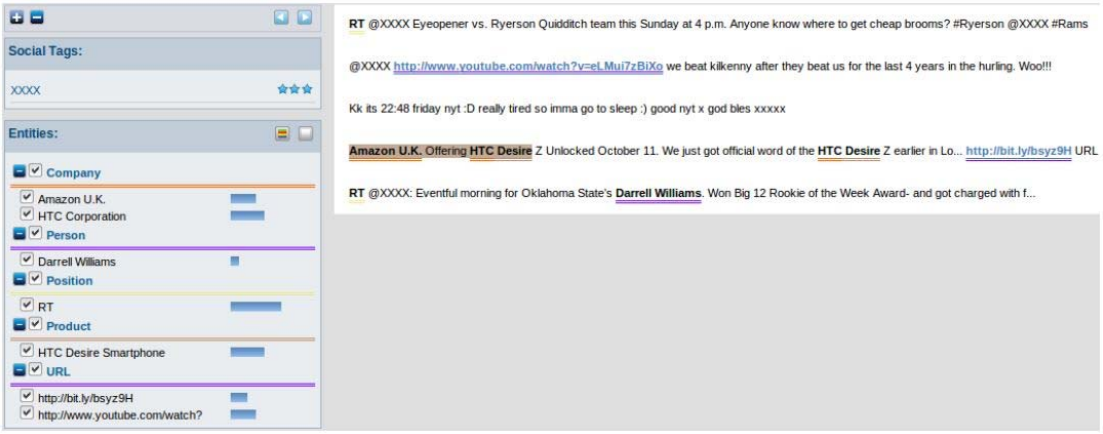
- Zemanta (www.zemanta.com): είναι ένα online εργαλείο που έχει δημιουργηθεί με σκοπό να προτείνει στους χρήστες tags και links για να εισάγουν στο περιεχόμενο των blogs και emails τους με σκοπό να το εμπλουτίσουν με αναφορές στην Wikipedia. Η παρακάτω εικόνα δείχνει ένα παράδειγμα κειμένου και προτεινόμενων tags με πιθανούς συνδέσμους στην Wikipedia και σε άλλα σχετικά άρθρα. [16]



Εικόνα 10 : Προτεινόμενα tags με πιθανούς συνδέσμους στην Wikipedia και σε άλλα σχετικά άρθρα σε κείμενο με το εργαλείο Zemanta

- Open Calais: αποτελεί υπηρεσία διαδικτύου για σημασιολογική αναγνώριση οντοτήτων η οποία έχει χρησιμοποιηθεί από ερευνητές των social media. Π.χ οι Abel et al εισήγαγαν το OpenCalais για την ονομαστική αναγνώριση οντοτήτων σε tweets που σχετίζονται με ειδήσεις. Η παρακάτω εικόνα δείχνει ένα παράδειγμα κειμένου το οποίο έχει αντιστοιχηθεί με ορισμένες οντότητες. Οι οντότητες περιλαμβάνουν URIs τα οποία επιτρέπουν την πρόσβαση μέσω HTTP σε επιπλέον πληροφορίες για την οντότητα σε βάσεις ανοικτών διασυνδεδεμένων δεδομένων. Συνδέεται με 8 Linked Data sets, συμπεριλαμβανομένων τις δικής του βάσης πληροφοριών, την DBpedia, Wikipedia, IMDB και Shopping.com. Ο μεγαλύτερος περιορισμός του εργαλείου αυτού είναι ότι οι χρήστες στέλνουν έγγραφα για να αντιστοιχιστούν από την υπηρεσία η οποία επιστρέφει τα

αποτελέσματα, ωστόσο δεν έχουν τα μέσα για να εισάγουν στο εργαλείο κάποια άλλη οντολογία ή να παραμετροποιήσουν το τρόπο με τον οποίο λειτουργεί η εξαγωγή των οντοτήτων. [16]



Εικόνα 11 : : Αντιστοίχιση κειμένου URIs με το εργαλείο Open Calais

2.10.4 Διαμόρφωση μοντέλων χρήστη από σημασιολογικές πληροφορίες

Ένα μοντέλο χρήστη (User Model - UM) είναι ένας πόρος γνώσης που περιέχει σημασιολογικές πληροφορίες για διάφορες πτυχές του χρήστη τις οποίες αντλεί είτε απευθείας από την πλατφόρμα των social media (π.χ. Facebook profile) είτε προκύπτουν από την γενικότερη συμπεριφορά του χρήστη. Μοντέλα χρηστών τα οποία βασίζονται σε οντολογίες έχουν χρησιμοποιηθεί εκτενώς σε πολλούς άλλους τομείς, εκτός από τα social media, ειδικά στην διαχείριση Προσωπικών Πληροφοριών (Personal Information Management - PIM). Στην περίπτωση των social media, το παραγόμενο από τους χρήστες περιεχόμενο δίνει τη δυνατότητα παραγωγής πλούσιων σημασιολογικών μοντέλων χρήστη. [16]

Με βάση τα είδη των σημασιολογικών πληροφοριών που χρησιμοποιούν οι μέθοδοι διαμόρφωσης μοντέλων χρηστών μπορούν να βασιστούν είτε σε οντότητες σημασιολογικά αποσαφηνισμένες, όπως αυτές που αναφέρουν οι ίδιοι οι χρήστες, είτε σε συνδέσεις που υπάρχουν ανάμεσα στο προφίλ του χρήστη και σε μεγαλύτερα αρχεία του Διαδικτύου καθώς και σε θεματολογία βασισμένη στις κατηγορίες της Wikipedia. [16]

Ακολουθεί μια σύντομη περιγραφή της διαδικασίας εξαγωγής των ενδιαφερόντων των χρηστών από πληροφορίες που υπάρχουν στο Twitter. Ένα μοντέλο χρηστή βασισμένο σε οντότητες διαμορφώνεται ως ένα σύνολο σταθμισμένων οντοτήτων. Το βάρος της κάθε οντότητας υπολογίζεται με βάση τον αριθμό των tweets στα οποία εμφανίζεται και τη συχνότητα εμφάνισης της σε συνδυασμό με τα άρθρα με τα οποία σχετίζεται. Τα προφίλ των χρηστών, οι διαμόρφωση των οποίων βασίζεται στη θεματολογία τους διαμορφώνονται με παρόμοιο τρόπο αλλά αντιπροσωπεύουν

μεγαλύτερο μέρος των κατηγοριών της Wikipedia. Και στους δύο παραπάνω τρόπους διαμόρφωσης προφίλ χρήστη γίνεται χρήση του εργαλείου Open Calais. Έρευνες έχουν δείξει ότι τα hashtags δεν αποτελούν χρήσιμο δείκτη των ενδιαφερόντων των χρηστών. [16]

Δύο ακόμη δείκτες τους οποίους λαμβάνουμε υπ' όψιν στη διαμόρφωση ενός μοντέλου χρήστη είναι οι αλληλεπιδράσεις όπως τα retweets και οι αλλαγές στα ενδιαφέροντα του χρήστη με την πάροδο του χρόνου. Έρευνες έχουν δείξει ότι ένα θέμα το οποίο εξελίσσεται με την πάροδο του χρόνου παράγει μοντέλα ενδιαφερόντων χρηστών τα οποία είναι ιδιαίτερα χρήσιμα για την παραγωγή προτάσεων μέσω του twitter. Επίσης, μπορούμε να αναγνωρίσουμε διαφορετικές ομάδες χρηστών βασισμένες στη διάρκεια των ενδιαφερόντων τους με βάση ένα συγκεκριμένο θέμα. [16]

Πληροφορίες από πολλά και διαφορετικά social media μπορούν να συνδυαστούν για τη διαμόρφωση των ενδιαφερόντων των χρηστών. Π.χ. τα likes του χρήστη στο Facebook ή ακόμα και τα ενδιαφέροντα τα οποία αναφέρει ο ίδιος στην σελίδα του στο LinkedIn ή Facebook μπορούν να συνδυαστούν με πληροφορίες που εξάγονται από τα tweets του. [16]

Το μοντέλο Open Provenance Model11 χρησιμοποιείται για την παρακολούθηση της προέλευσης των ενδιαφερόντων των χρηστών. Επιλεγμένες λέξεις κειμένου των tweets με βάση ορισμένα κριτήρια θεωρούνται υποψήφιες οντότητες και αναζητούνται στα άρθρα της Wikipedia. Σε πρώτο στάδιο αποσαφηνίζονται οι οντότητες που ταιριάζουν καλύτερα. Για κάθε οντότητα λαμβάνεται υπ' όψιν όλο το δέντρο των υποκατηγοριών της Wikipedia που υπάρχει κάτω από αυτή. Στη συνέχεια καταχωρείται ένα θέμα σε κάθε υποψήφια οντότητα, όλες οι κατηγορίες που υποδέντρου αναλύονται ώστε να ανακαλυφθούν οι πιο συχνά εμφανιζόμενες υποκατηγορίες, οι οποίες ορίζουν τα ενδιαφέροντα των χρηστών στο προφίλ τους. [16]

Πολλοί ερευνητές έχουν χρησιμοποιήσει την Wikipedia για την μοντελοποίηση των ενδιαφερόντων των χρηστών. Προτείνουν επίσης μεθόδους για την ενοποίηση των διαφόρων προφίλ ενός χρήστη στα μέσα κοινωνικής δικτύωσης. Ετικέτες από διαφορετικούς ισοτόπους φιλτράρονται με βάση τα συνώνυμα που υπάρχουν στο λεξικό WordNet (λεξιλογική βάση δεδομένων για την Αγγλική γλώσσα) και σχετίζονται με τις αντίστοιχες σελίδες της Wikipedia. Στη συνέχεια, οι κατηγορίες της Wikipedia που θα προκύψουν από τις επιλεγμένες σελίδες χρησιμοποιούνται ώστε να εντοπιστούν αντιπροσωπευτικά θέματα τα οποία θα έχουν υψηλό ενδιαφέρον για τον χρήστη. [16]

Τέλος, ο σημασιολογικός ιστός μπορεί να μας βοηθήσει στην αναγνώριση της συμπεριφοράς του χρήστη. Μπορούμε να κατατάξουμε τους χρήστες των social media σε κατηγορίες ανάλογα με τα χαρακτηριστικά της συμπεριφορά τους. Π.χ. οι

χρήστες του twitter χωρίζονται σε meformers (80%-τα posts αφορούν τον εαυτό τους) και informers (20%). Οι ερευνητές προκειμένου να ορίσουν αυτόματα ρόλους συμπεριφοράς δημιουργούν βασικούς κανόνες με ερωτήματα SPARQL οι οποίοι χαρτογραφούν τα σημασιολογικά χαρακτηριστικά των χρηστών μέσω των αλληλεπιδράσεων τους σε διάφορα επίπεδα συμπεριφοράς. Τα επίπεδα αυτά απορρέουν δυναμικά από τις αλληλεπιδράσεις του χρήστη και μπορούν να μεταβληθούν με την πάροδο του χρόνου καθώς οι διαδικτυακές κοινότητες αναπτύσσονται. Οι ρόλοι των χρηστών και οι αλληλεπιδράσεις τους μοντελοποιούνται σημασιολογικά μέσω της οντολογίας Συμπεριφοράς Χρήστη (User Behaviour Ontology).

2.10.5 Πρόσβαση σε πληροφορίες σημασιολογικού ιστού μέσω των ροών των social media (Σύστημα Twarql)

Ο σημασιολογικός ιστός επιτρέπει στους χρήστες να αναζητούν έγγραφα τα οποία περιέχουν έννοιες προερχόμενες από τις οντολογίες σημασιολογικού ιστού με τις οποίες σχετίζεται. Ανάλογα με τη μέθοδο που χρησιμοποιείται, τα ερωτήματα της αναζήτησης μπορούν να αναμειχθούν με λέξεις κλειδιά ή να γίνει εφαρμογή φίλτρων σε αυτά για καλύτερα αποτελέσματα(GATE Mimir). Τα εργαλεία αυτά συχνά παρέχουν λειτουργίες περιήγησης και αναζήτησης με δυνατότητες βελτίωσης. Η αναζήτηση και περιήγηση του περιεχομένου του σημασιολογικού ιστού είναι πολύ σημαντική λόγω της μεγάλης έντασης και των γρήγορων αλλαγών στο περιεχόμενο των ροών των social media. [16]

Οι αναζητήσεις που βασίζονται σε σημασιολογικά χαρακτηριστικά παραγόμενα από τις οντολογίες της DBpedia Spotlight παρέχουν σημαντικά καλύτερα αποτελέσματα. Το σύστημα Twarql δημιουργεί τριπλέτες RDF από tweets, βασισμένες σε μεταδεδομένα των ίδιων των tweets, σε αναφορές οντοτήτων, hashtags και URLs. Οι τριπλέτες αυτές κωδικοποιούνται με τη χρήση λεξιλογίων Open Data (FOAF, SIOC) και μπορούν να εντοπιστούν με την χρήση ερωτημάτων SPARQL. Είναι επίσης πιθανό μια ροή από tweets να ταιριάζει με ένα περίπλοκο σημασιολογικό ερώτημα π.χ. ποιοι ανταγωνιστές αναφέρονται στο προϊόν μου. [16]

Οι Abel et al χρησιμοποίησαν οντότητες του σημασιολογικού ιστού από το εργαλείο OpenCalais σε συνδυασμό με μοντέλα χρηστών για να δημιουργήσουν σημασιολογικές όψεις. Οι δύο βασικές όψεις βασίζονται στην αναζήτηση λέξεων κλειδιών και hashtags. Τα καλύτερα αποτελέσματα εξάγονται όταν οι όψεις είναι εξατομικευμένες, βασίζονται δηλαδή στις προτιμήσεις του συγκεκριμένου χρήστη για τον οποίο γίνεται η αναζήτηση. Τέλος, είναι σημαντικό οι όψεις αυτές να λαμβάνουν υπόψη τους και το χρονικό πλαίσιο εφαρμογής τους. [16]

2.10.6 Οπτικοποιήσεις

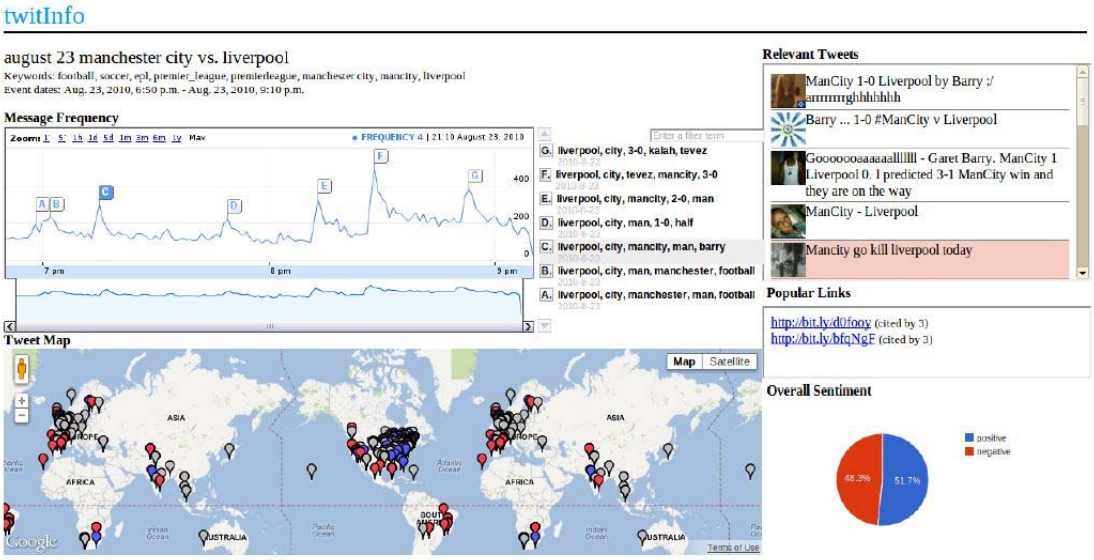
Υπάρχουν εργαλεία οπτικοποίησης τα οποία προσπαθούν να αποτυπώσουν τις σημασιολογικές σχέσεις που υπάρχουν ανάμεσα στα διάφορα θέματα στη ροή των social media. Π.χ. το BlogScope υπολογίζει συσχετίσεις μεταξύ keywords με τον εντοπισμό αμοιβαίων πληροφοριών που υπάρχουν για ζευγάρια keywords χρησιμοποιώντας ένα τυχαίο δείγμα εγγράφων. Ένα ακόμη είδος οπτικοποίησης είναι η απεικόνιση με βάση την τοποθεσία, η οποία απεικονίζει ομοιότητες στη θεματολογία με βάση την εγγύτητα των τοποθεσιών βάση χάρτη. Επίσης, οι αλληλεξαρτήσεις μεταξύ των θεμάτων ενός εγγράφου μπορούν να εντοπιστούν μέσα από οπτικοποιήσεις που βασίζονται σε γράφους που σχηματίζονται ανάλογα με τη δύναμη των θεμάτων. Τέλος, σημαντική παράμετρος στο παραγόμενο από τους χρήστες περιεχόμενο είναι ο τόπος αναφοράς. Π.χ. κάποια tweets είναι γεωγραφικά τοποθετημένα με πληροφορίες ακόμη και για τη συντεταγμένες του σημείου, ενώ πολλά facebook, twitter και blog profiles περιέχουν πληροφορίες σχετικά με την τοποθεσία του χρήστη. Π.χ. στο εργαλείο Twitris ο χρήστης επιλέγει μία πολιτεία των ΗΠΑ από το Google Maps και του εμφανίζονται τα θέματα που συζητήθηκαν περισσότερο στα social media στην περιοχή αυτή. [16]

Εικόνα 12 : Χάρτης από τις προεδρικές εκλογές του 2012.

Οι ερευνητές έχουν επίσης διερευνήσει το πρόβλημα της οπτικοποίησης των συζητήσεων στα social media σχετικά με θέματα του πραγματικού κόσμου, όπως ποδοσφαιρικούς αγώνες, συνέδρια και άλλες ειδήσεις. Βασικό στοιχείο αποτελεί η ικανότητα του εντοπισμού δευτερευόντων γεγονότων και ο συνδυασμός τους με χρονοδιαγράμματα, χάρτες και οπτικοποιήσεις συγκεκριμένης θεματολογίας. [16]

Οι εφαρμογές που χρησιμοποιούν τεχνικές πλοήγησης και οπτικοποίησης μπορούν εύκολα να βελτιωθούν αξιοποιώντας την επιπλέον σημασιολογική γνώση των οντοτήτων που αναφέρονται στις ροές των social media. Όταν οντότητες και θέματα σχετίζονται μέσω URIs με πόρους βάσεων ανοικτών διασυνδεδεμένων δεδομένων, όπως η DBpedia, μπορούν να υποστηρίξουν οπτικοποιήσεις που βασίζονται στις μεταξύ τους σημασιολογικές συσχετίσεις. Τέλος, η εξερεύνηση των social media με

βάση τη θεματολογία τους καθώς και οι οπτικοποιήσεις τους στη ροή του χρόνου μπορούν να εμπλουτιστούν με πολύπλευρες αναζητήσεις με βάση την οντολογίες και με διεπαφές με σημασιολογικά ερωτήματα.



Εικόνα 13 : Οπτικοποίηση με χρήση εργαλείου twitinfo των αντιδράσεων στην ροή του twitter με βάση ένα συγκεκριμένο γεγονός

Οι περισσότερες εφαρμογές τείνουν να υιοθετούν περισσότερες από μια οντολογίες καθώς μοντελοποιούν τις διαφορετικές οπτικές της. Όσον αφορά τους πόρους του Web of Data, οι σημερινές μέθοδοι χρησιμοποιούν κυρίως πόρους της Wikipedia (DBpedia & YAGO). [16]

2.11 Πως τα social media επηρεάζουν την online καταναλωτική συμπεριφορά

Η μοναδική φύση των δεδομένων των social media είναι αυτή που τα κάνει τόσο ενδιαφέροντα καθώς το περιεχόμενο τους είναι δυναμικό και αντανακλά τις κοινωνικές και συναισθηματικές διακυμάνσεις των χρηστών. Οι δραστηριότητά τους συχνά επηρεάζεται από δημοφιλή συμβάντα και θέματα. Αναπτύσσονται με μεγάλη ταχύτητα, δυναμικά και σε μεγάλο όγκο αντανακλώντας τις αλλαγές στις κοινωνικές απόψεις.

Τα Social Media βασίζονται στις τεχνολογίες του Web 2,00. Το Web 2,00 προσφέρει την κατάλληλη υποδομή που υποστηρίζει τις online αλληλεπιδράσεις των χρηστών οι οποίες επηρεάζουν την αγορά προϊόντων και υπηρεσιών και κατά επέκταση όλο τον κλάδο του ecommerce. Οι πλατφόρμες κοινωνικής δικτύωσης, με σημαντικότερες το Facebook, Instagram, Twitter και LinkedIn, εισήγαγαν νέα επιχειρηματικά μοντέλα στον τομέα του ηλεκτρονικού εμπορίου. Παρέχουν πρόσφορο έδαφος για καινοτομία και νέα ερευνητικά πεδία με την ανάπτυξη νέων εννοιών όπως το Co-Creation, όπου

μια εταιρεία συνεργάζεται με τους πελάτες της προκειμένου να παράγει καλύτερες ιδέες, προϊόντα και υπηρεσίες, και το Word of Mouth. Η δύναμή τους πηγάζει από τον μεγάλο αριθμό ατόμων που συμμετέχουν σε αυτά οι δραστηριότητες των οποίων επηρεάζουν τις απόψεις και τη συμπεριφορά ενός χρήστη. [3]

Αξίζει να αναφερθούμε σε πλατφόρμες οι οποίες ασχολούνται με την αναγνώριση συναισθημάτων και απόψεων των χρηστών. Υπάρχει πληθώρα websites τα οποία φιλοξενούν αποκλειστικά απόψεις και αξιολογήσεις για προϊόντα και υπηρεσίες για τα οποία οι χρήστες αισθάνονται την ανάγκη να εκφράσουν διαδικτυακά τις απόψεις τους. Στα social media ωστόσο οι χρήστες τις περισσότερες φορές με τα μηνύματα τους εκφράζουν την διάθεση τους και την προσωπική τους γνώμη παρά τις απόψεις τους για γενικότερα θέματα. Έρευνες έχουν δείξει ότι το 19% των μηνυμάτων σε microblogs αφορά επωνυμίες προϊόντων και τις δραστηριότητές τους, ενώ το 20% περιέχει συναισθήματα για τα προϊόντα αυτά. Επομένως, μπορούμε να καταλήξουμε στο συμπέρασμα ότι οι πληροφορίες που απορρέουν από τα social media μπορούν να χρησιμοποιηθούν για την αύξηση της κερδοφορίας στον κλάδο του ecommerce. [16]

Τα Social Media παίζουν καθοριστικό ρόλο στην φήμη και στην αποδοχή ενός νέου προϊόντος/υπηρεσίας, καθώς και στην πρόθεση αγοράς τους από τους χρήστες. Έχει αποδειχθεί ωστόσο ότι δεν επηρεάζουν την ικανοποίηση των πελατών. Τα Social Media αποτελούν το καλύτερο εργαλείο που έχει στη διάθεση της μια επιχείρηση για να χτίσει τις σχέσεις με τους πελάτες της δημιουργώντας λογαριασμούς στα διάφορα κοινωνικά δίκτυα. Με τον τρόπο αυτόν αυξάνεται η δημοφιλία των προϊόντων της επιχείρησης (brand awareness), καθώς οι χρήστες μέσω των Social Media έχουν πρόσβαση σε μια πληθώρα διαφορετικών απόψεων για το προϊόν που τους ενδιαφέρει. [6]

Ένα ακόμα κομμάτι μιας εταιρείας στο οποίο παίζουν σημαντικό ρόλο τα Social Media είναι ο τομέας του marketing. Καθοριστικοί παράγοντες στην ανάπτυξη του ηλεκτρονικού εμπορίου είναι η τιμή των προϊόντων/υπηρεσιών, οι εκπτώσεις που προσφέρει μια επιχείρηση, η ασφάλεια των συστημάτων παράδοσης των παραγγελιών της καθώς και η εμπιστοσύνη των πελατών προς την διαδικτυακή επιχείρηση. Έχει αποδειχθεί ότι τα πολλαπλά κανάλια marketing συμβάλλουν αποτελεσματικά στην αύξηση της πίστης των πελατών. Πολλές ηλεκτρονικές επιχειρήσεις έχουν αντικαταστήσει την κλασική στρατηγική marketing με την αντίστοιχη στρατηγική στα Social Media (Social Media Marketing) και αυτό γιατί η τελευταία υπόσχεται καλύτερα αποτελέσματα και μεγαλύτερη απόδοση. Οι πελάτες μπορούν πλέον να επικοινωνήσουν με την ηλεκτρονική επιχείρηση μέσω των Social Media αυξάνοντας την εμπιστοσύνη τους προς αυτήν ενώ παράλληλα οι επιχειρήσεις μπορούν να μειώσουν το κόστος και τον χρόνο των προωθητικών τους ενεργειών. [6]

Πλούσιο υλικό για την ανάλυση των δεδομένων που αφορούν προϊόντα και υπηρεσίες παράγει το λεγόμενο Social Commerce, δηλαδή η πραγματοποίηση ενεργειών και συναλλαγών ηλεκτρονικού εμπορίου στο περιβάλλον των Social

Media, των κοινωνικών δικτύων που χρησιμοποιούν τεχνολογίες Web 2.00. Οι ιστοσελίδες όπου λαμβάνει χώρα το Social Commerce είναι πλατφόρμες στις οποίες οι χρήστες μπορούν να συνεργαστούν online, να λάβουν συμβουλές από άλλους χρήστες και να αναζητήσουν προϊόντα και υπηρεσίες πριν προχωρήσουν στην αγορά τους. Σημαντικές δυνατότητες των Social Websites είναι:

- Τα κλασικά e-shops, όπως η Amazon.com κ.α., μπορούν να προσθέσουν δυνατότητες κοινωνικής δικτύωσης ώστε να επωφεληθούν από τη δύναμη των κοινωνικών μέσων και να κατανοήσουν καλύτερα τις ανάγκες του αγοραστικού τους κοινού.
- Από τη μεριά τους οι πλατφόρμες κοινωνικής δικτύωσης προσθέτουν εμπορικές λειτουργίες οι οποίες επιτρέπουν την προβολή εμπορικών διαφημίσεων ή ακόμα και τη διεξαγωγή συναλλαγών. Πλατφόρμες όπως το Facebook, Instagram και LinkedIn πλέον προσφέρουν εργαλεία προγραμματισμού για διεξαγωγή εμπορικών συναλλαγών μεταξύ των μελών τους. [3]

Παρά το γεγονός ότι κάποιες δραστηριότητες των κοινωνικών δικτύων δεν έχουν εμπορική φύση, δεν έχουν δηλαδή εμπορικά οφέλη όπως η αγορά ή η πώληση, οι πληροφορίες που μοιράζονται είναι σημαντικές και μπορούν να έχουν εμπορικές επιπτώσεις. Ακολουθούν τα μετρήσιμα αποτελέσματα τα οποία μπορούμε να εξάγουμε από την ανάλυση των δεδομένων μιας διαδικτυακής πλατφόρμας:

- Εμπιστοσύνη των καταναλωτών
- Χρησιμότητα web site
- Κέρδη / Έσοδα
- Ανάπτυξη της αγοράς
- Νέα προϊόντα και υπηρεσίες
- Η γνώμη του καταναλωτή
- Η πρόθεση για αγορά
- Ικανοποίηση του πελάτη
- Click Through Rate
- Η αντίληψη του χρήστη [3]

Οι βασικοί τομείς οι οποίοι ωφελούνται από τη χρήση των κοινωνικών δικτύων είναι:

1. Marketing / Διαφήμιση (reviews, ratings, word of mouth, recommendations)
2. Διοίκηση Επιχειρήσεων
3. Management
4. Τεχνολογική υποδομή
5. Στρατηγικές συμφωνίες με άλλες εταιρείες (Collaboration)
6. CRMs – υπηρεσίες προς τους πελάτες
7. Συναλλαγές
8. Πηγές πληροφόρησης [3]

Τα πλεονεκτήματα που λαμβάνουν οι εταιρείες ecommerce από την αξιοποίηση των πληροφοριών που προέρχονται από τα social media είναι:

1. Προσωποποίηση προϊόντων και υπηρεσιών: Επιτυγχάνεται με την προσφορά προσωποποιημένου περιεχομένου και προωθητικών ενεργειών. Η ανάλυση των προφίλ των χρηστών και της δραστηριότητας τους στα social media βοηθάει της εταιρείες να ξεχωρίσουν τους νέους από τους υφιστάμενους πελάτες, καθώς και αυτούς που έχουν βάση ενδιαφερόντων την μεγαλύτερη πιθανότητα να γίνουν μελλοντικοί πελάτες ώστε να ακολουθήσουν την κατάλληλη προωθητική πολιτική. Σύμφωνα με έρευνες η προσωποποίηση αυτή μπορεί να αυξήσει τις πωλήσεις τουλάχιστο κατά 10% και 5-8 φορές επιπλέον την απόδοση των επενδύσεων και των δαπανών του marketing (personalized email marketing, follow up κ.α.).
2. Καλύτερη εξυπηρέτηση των πελατών: Με χρήση κατάλληλων εργαλείων και αισθητήρων που ανιχνεύουν πληροφορίες, σχόλια κ.α. για τα προϊόντα και τις υπηρεσίες της επιχείρησης μπορούν να προσφέρουν καλύτερες υπηρεσίες "μετά την πώληση" στους πελάτες τους.
3. Εργαλεία ανάλυσης προβλέψεων: Η αναγνώριση γεγονότων πριν αυτά συμβούν με τη χρήση των Open Data απαιτεί εντατική εξόρυξη δεδομένων η οποία βοηθάει τις επιχειρήσεις να πραγματοποιήσουν προβλέψεις για τα έσοδά ή ζημίες τους και να αναγνωρίσουν μελλοντικά μοτίβα πωλήσεων με βάση τις πωλήσεις του παρελθόντος. [1]

2.12 Social Data Analysis

Υπάρχουν τρεις τύποι χρηστών στη διαδικασία του Data Analysis, αυτοί που παράγουν τα δεδομένα, αυτοί που τα συλλέγουν και αυτοί που τα αναλύουν. [2] Στα πλαίσια την παραγωγής και συλλογής δεδομένων έχουν αναπτυχθεί σύγχρονες τεχνολογίες με σκοπό την παραγωγή μεγάλων δομημένων οντολογιών, όπως το Google's Knowledge Graph, οι οποίες παράγονται και διατηρούνται με την ενσωμάτωση δεδομένων που προέρχονται από υψηλής ποιότητας δομημένες πηγές πληροφορίας. Η εφαρμογή σημασιολογικών ερωτημάτων στις βάσεις αυτές έχουν μεγάλο νόημα για τον τομέα των επιχειρήσεων. Η DBpedia συνεχώς εξελίσσεται με την εξαγωγή δεδομένων από τις βάσεις της Wikipedia. Πιο συγκεκριμένα, τον Οκτώβριο του 2016 περιείχε 4,18 εκατομμύρια αντικείμενα και τα 4,22 από αυτά ήταν ταξινομημένα σε δομημένες οντολογίες. Ωστόσο, η συνεχής παραγόμενη γνώση δεν μπορεί να παρακολουθηθεί από την Wikipedia. Συγκεκριμένα, η ανακάλυψη πληροφοριών που αφορούν δημοφιλή θέματα είναι πιθανό να πραγματοποιηθεί από τις οντολογίες της Wikipedia, ωστόσο λιγότερο δημοφιλή θέματα τα οποία πρόκειται να παρουσιάσουν μελλοντική ανάπτυξη δεν είναι εύκολο να ανακαλυφθούν. [14]

Οι μακροχρόνιες προβλέψεις παίζουν πολύ σημαντικό ρόλο στο ecommerce, για τον λόγο αυτό χρησιμοποιούμε μια νέα περισσότερο ισχυρή πηγή πληροφόρησης, τα Social Media.[14] Ένα ευρύ φάσμα εταιρειών και δημοσίων φορέων συμμετέχουν ενεργά στην ανάλυση των open data που προέρχονται από Social Media με τη χρήση ελεύθερων APIs (Application Program Filter Interfaces) που διαθέτουν οι διάφορες πλατφόρμες Social Media. [2] Ο μεγάλος όγκος των social media (1.8 δισ, χρήστες)

και η επικαιρότητα των δεδομένων επιτρέπουν ακόμα και σε δεδομένα μικρότερης αξίας αλλά μεγάλης πιθανής μελλοντικής αξίας να αφήσουν κάποια ίχνη ώστε να ανακαλυφθούν. Η διαδικασία αυτή εμπεριέχει αρκετές δυσκολίες καθώς τα ίχνη αυτά δεν είναι ταξινομημένα, είναι διασκορπισμένα, αμφιβόλου ποιότητας και ενδεχομένως εσφαλμένα. Ο ερευνητικός κόσμος βρίσκεται ακόμα σε πρώιμο στάδιο όσον αφορά την σύνθεση οντολογιών από πληροφορίες των social media, καθώς η DBpedia, YAGO, Google Knowledge Graphs και Facebook αποτελούνται κυρίως από αδόμητα ή ημιδομημένα δεδομένα. Για τον λόγο αυτό έχουν αναπτυχθεί τα Social Media Analytics τα οποία αναλύουν τα φαινόμενα του πραγματικού κόσμου μέσα από τα social media. Στη κατεύθυνση αυτή έχουν δημιουργηθεί χρήσιμες τεχνολογίες και εργαλεία που αναγνωρίζουν και δημιουργούν συνδέσμους με κοινωνικό περιεχόμενο. [14]

Τα social media περιλαμβάνουν χαμηλής συχνότητας δεδομένα τα οποία μπορούν να εξαχθούν με χρήση ειδικών ανεξάρτητων εργαλείων του σημασιολογικού ιστού. Η χρήση της DBpedia σαν πηγή άντλησης της πληροφορίας εμπεριέχει τους περιορισμούς που αναφέραμε παραπάνω, ωστόσο υπάρχουν εμπορικά εργαλεία τα οποία αντιστοιχίζουν τα δεδομένα των social media με τύπους της DBpedia. [14]

Ένας από τους βασικούς τρόπους αξιοποίησης των δεδομένων των Social Media είναι ο διαχωρισμός του κοινού σε κατηγορίες, τακτική η οποία εφαρμόζεται σε διάφορους τομείς. Π.χ. οι διαφημιστικές εταιρείες κατηγοριοποιούν τους χρήστες μέσα από διαδικασίες Data Analysis. Με βάση τη συμπεριφορά τους και τη δράση τους στα Social Media οι χρήστες αξιολογούνται και διαχωρίζονται σε αυτούς που αποτελούν την αγορά στόχο και στους οποίους αξίζει να εφαρμοσθούν προωθητικές ενέργειες και σε εκείνους που δεν αξίζει να ασχοληθούν οι εταιρείες. Η ανάλυση των δεδομένων των χρηστών με την μεγαλύτερη σημαντικότητα και επιρροή είναι βασική καθώς δίνει τη δυνατότητα στις εταιρείες να προσαρμόσουν τα νέα και τις δραστηριότητές τους με βάσει τα στοιχεία που έχουν για το κοινό τους. Με τα δεδομένα αυτά παράγονται διάφορες αναπαραστάσεις του κοινού όπως αυτές που παράγει ο Twitter's Trends Algorithm ο οποίος αναπαριστά τις συζητήσεις στο Twitter ανά γεωγραφικό διαμέρισμα την κάθε χρονική στιγμή. [2]

Η ροή των Tweets μπορεί να αναλυθεί σε πραγματικό χρόνο, και τα αποτελέσματα της ανάλυσης να επεξεργαστούν και να προσαρμοστούν για αναζήτηση και οπτικοποίηση. Στη διαδικασία συλλογής δεδομένων, οι λογαριασμοί χρηστών και τα hashtags μπορούν να εξαχθούν με τη χρήση της Twitter API. Η εφαρμογή παράγει ένα αρχείο JSON το οποίο αποθηκεύεται για μελλοντική χρήση. Το εργαλείο συλλογής των δεδομένων χρησιμοποιεί τον Hosebird client, μια βιβλιοθήκη της Java η οποία διευκολύνει την πρόσβαση στην ροή δεδομένων του API. Το Hosebird χειρίζεται την πολυπλοκότητα της αυθεντικοποίησης των μεγάλης διάρκειας συνδέσεων HTTP και την προσπάθεια επανασύνδεσης όταν η σύνδεση "πέφτει" για οποιονδήποτε λόγο. [13]

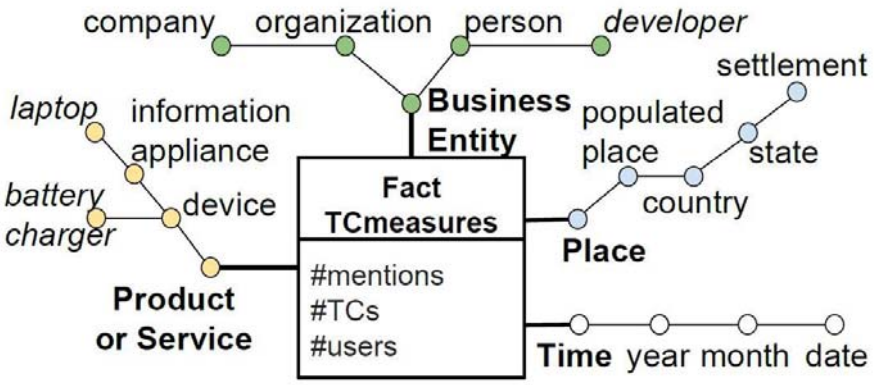
2.12.1 Social CRMs

Τα Social Media λόγω της διαδεδομένης και εύκολης χρήσης τους παράγουν ανά πάσα στιγμή ενημερωμένα δεδομένα σχετικά με μια πληθώρα θεμάτων. Τα δεδομένα αυτά παράγονται από τους ίδιους τους χρήστες, η επίσημη ορολογία τους είναι "Περιεχόμενο Παραγόμενο από τους Χρήστες" (User Generated Content - UGC), και έχουν μια πληθώρα πεδίων εφαρμογών ανάμεσα στους οποίους είναι και ο κλάδος των επιχειρήσεων με τα συστήματα CRMs (Customer Relationship Management). Έχουμε δηλαδή την ανάπτυξη των λεγόμενων Social CRMs. [7]

Οι τεχνικές της επιχειρηματικής ευφυίας, όπως το Μοντέλο Διαστάσεων και η τεχνική Online Analytical Processing - OLAP, είναι χρήσιμες στην ανάλυση της πληροφορίας που εξάγουμε από τα Social Media και αφορούν διάφορες εφαρμογές και τομείς όπως αυτών των επιχειρήσεων και των Social CRMs. Ο συνδυασμός της επιχειρηματικής ευφυΐας με τις τεχνολογίες του σημασιολογικού ιστού προσφέρουν ένα βέλτιστο επίπεδο ανάλυσης. [7]

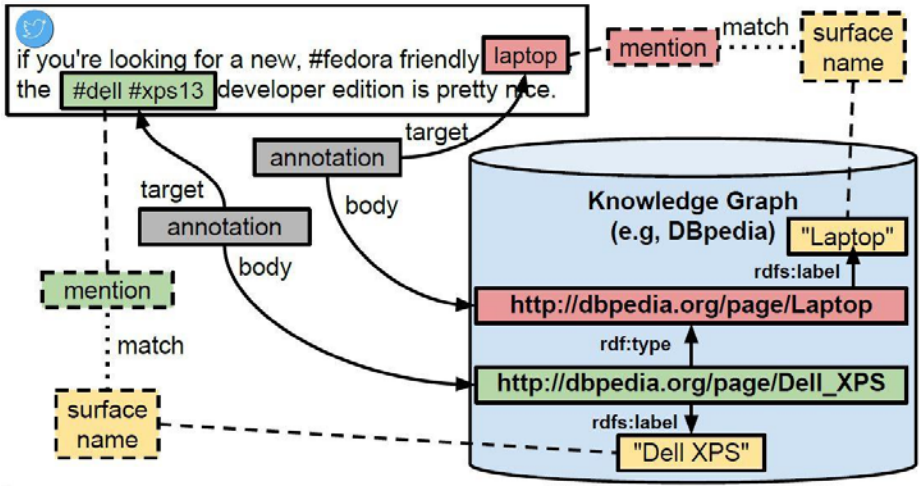
Κάθε φράση κειμένου, όπως οι αναρτήσεις των χρηστών στα social media, συνοδεύεται από πληροφορίες. Οι πιο χρήσιμες από αυτές είναι ο χρόνος (time stamp) και ο τόπος δημιουργίας της ανάρτησης (γεωγραφικές συντεταγμένες), καθώς και διάφορα μεταδεδομένα όπως η ταυτότητα του συντάκτη κλπ. Η ανάλυση του περιεχομένου που προέρχεται από τα social media μπορεί να χρησιμοποιηθεί από ποικίλες εφαρμογές για την πρόβλεψη των τάσεων σε νέα προϊόντα/υπηρεσίες, την ανάπτυξη νέων brands, το στοχευμένο marketing και τη χρήση σε συστήματα CRMs. Ο εντοπισμός και η ανάλυση τέτοιων δεδομένων είναι σε πολλές περιπτώσεις δύσκολη διαδικασία. Η αναγνώριση των σχετικών με ένα συγκεκριμένο αντικείμενο αναφορών μέσα σε ένα μεγάλο όγκο κειμένου ή μεγάλο αριθμό microtexts (π.χ. tweets) και η μεταξύ τους σύνδεση με σκοπό την παραγωγή πληροφοριών σημασιολογικού ιστού για ένα αντικείμενο αποτελεί μεγάλη πρόκληση. [7]

Η παραγωγή πληροφορίας που να έχει νόημα για τον σημασιολογικό ιστό (semantic annotation) συσχετίζει ορισμένα αντικείμενα με συγκεκριμένα δεδομένα, όπως κείμενο και πολυμέσα, συνδέοντας τα μέσω εννοιών που ανήκουν σε Γράφους Πληροφορίας (Knowledge Graph - KG) όπως οι οντολογίες, οι συλλογές από Linked Open Data (LOD) (DBpedia, Wikdata), λεξικά (WordNet) ή συνδυασμούς LODs και λεξικών (Babelnet). Πολλές τεχνικές έχουν αναπτυχθεί για την αυτοματοποίηση της διαδικασίας αυτής οι οποίες συνδυάζουν τεχνολογίες εξόρυξης κειμένου και επεξεργασίας φυσικής γλώσσας. [7]



Εικόνα 14 : Έννοιες ενός tweet που ενδιαφέρουν μια επιχείρηση

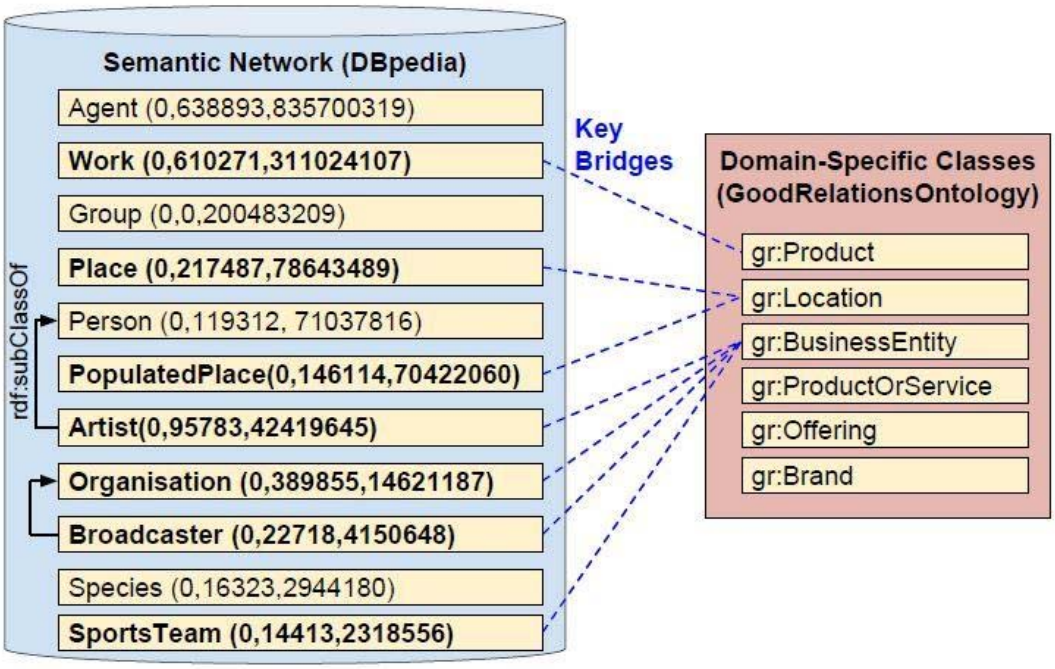
Το παραπάνω διάγραμμα αποτελεί ένα παράδειγμα όπου αναλύονται οι έννοιες που ενδιαφέρουν μια επιχείρηση και που μπορεί να υπάρχουν σε ένα tweet .Το διάγραμμα οργανώνεται με βάση τις 3 μετρικές ενός γεγονότος (Fact TCmeasures) που είναι οι αναφορές (mentions), τα text clips (TCs) και οι χρήστες (users). Οι διαστάσεις που αναλύονται προκύπτουν από τις ιεραρχίες των κλάσεων της DBpedia (Προϊόν ή υπηρεσία, η οντότητα της επιχείρησης, τόπος, χρόνος). [7]



Εικόνα 15 : Παραγωγή RDF από ένα tweet με την χρήση πληροφορίας της βάσης DBpedia-Spotlight

Στην παραπάνω εικόνα τώρα έχουμε ένα παράδειγμα διαδικασίας παραγωγής περιεχομένου σημασιολογικού ιστού σε μορφή RDF από ένα tweet με την χρήση πληροφορίας της Ανοικτής Διασυνδεδεμένης Βάσης Δεδομένων DBpedia-Spotlight. Η παραγόμενες αντιστοιχίσεις δεδομένων αφορούν τις λέξεις laptop, dell και xps13 οι οποίες συνδέονται με τα αντίστοιχα αντικείμενα του γράφου πληροφορίας (π.χ. DBpedia). Παρατηρούμε ότι υπάρχει μια σχέση ανάμεσα στις οντότητες καθώς η αναφορά “Dell XPS” αποτελεί ιδιότητα (rdfs:type) του υποκειμένου “Laptop”. [7]
Η παραγωγή περιεχομένου σημασιολογικού ιστού από περιεχόμενο δημοσιεύσεων στα Social Media εμπεριέχει κάποιες δυσκολίες οι οποίες οφείλονται κυρίως στο γεγονός ότι περιλαμβάνουν αρκετή αδόμητη πληροφορία. Μια λύση αποτελεί η

χρήση εννοιών γενικού σκοπού που προέρχονται από γνωστούς γράφους πληροφορίας, όπως η DBpedia, και το φιλτράρισμα τους χρησιμοποιώντας συνδέσμους εδραιωμένων οντολογιών όπως το πρότυπο GoodRelations Ontology (GRO) που αναφέρθηκε παραπάνω. Οι οντολογίες αυτές μοντελοποιούν σχέσεις και κλάσεις γενικού ενδιαφέροντος στον τομέα των επιχειρήσεων. [7]



Εικόνα 16 : Σύνολο συνδέσμων ανάμεσα σε κλάσεις της BDpedia και σε κλάσεις της GRO

Στην παραπάνω εικόνα βλέπουμε ένα παράδειγμα συνδέσμων που μπορούν να δημιουργηθούν ανάμεσα σε κλάσεις της BDpedia και σε κλάσεις της GRO στον τομέα των επιχειρήσεων. Το μοντέλο αυτό χρησιμοποιείται από εταιρείες όπως η Google, Yahoo!, BestBuy, Sears και Kmart. Οι κλάσεις της οντολογίας Good Relations θεωρούνται από πολλούς ερευνητές του κλάδου κατάλληλες για την ανάλυση πληροφοριών του κλάδου του ηλεκτρονικού εμπορίου. Μερικά παραδείγματα αποτελούν οι οντότητες gr:Offering - πώληση, επισκευή, ενοικίαση κλπ., gr:ProductOrService και gr:Location - κατάστημα ή διαθέσιμη προσφορά. [7]

Η τεχνική της δημιουργίας συνδέσμων ανάμεσα σε κλάσεις Ανοιχτών Διασυνδεδεμένων Δεδομένων (όπως αυτές της DBpedia) και υψηλού επιπέδου εννοιών μιας οντολογίας (όπως η GoodRelations) παίζει σημαντικό ρόλο στον εντοπισμό των αντικειμένων που παρουσιάζουν ενδιαφέρον σε έναν συγκεκριμένο κλάδο. Πειράματα που έχουν πραγματοποιηθεί με χρήση tweets το περιεχόμενο των οποίων έχει εμπλουτιστεί με πόρους από την DBpedia με τη χρήση του DBpedia-Spotlight, έδειξαν ότι ακόμα και ελάχιστοι σύνδεσμοι ανάμεσα σε κλάσεις όπως αυτοί της παραπάνω εικόνας είναι αρκετοί για να εξασφαλίσουν τη συνοχή των υπαρχόντων συνδέσμων αλλά και για να ανακαλύψουν ένα μεγαλύτερο αριθμό νέων συνδέσμων. Οι σύνδεσμοι αυτοί βοηθούν στον εντοπισμό αναφορών που

παρουσιάζουν υψηλό ενδιαφέρον για έναν συγκεκριμένο επιχειρηματικό κλάδο, όπως αυτόν του ecommerce. [7]

2.12.2 Recommendation systems με χρήση social content

Τα σύγχρονα Recommendation Systems τα οποία χρησιμοποιούνται για την πρόταση προϊόντων και υπηρεσιών στους χρήστες του διαδικτύου χρησιμοποιούν αλγορίθμους που λαμβάνουν υπόψιν τους βαθμολογίες και σχόλια προϊόντων ή υπηρεσιών τα οποία υπάρχουν σε μορφή περιεχομένου στα Social Media. Τα σχόλια αποτελούν την πιο έγκυρη πηγή πληροφόρηση καθώς είναι δυσκολότερο για έναν χρήστη να δώσει ψευδή στοιχεία από ότι είναι όταν βαθμολογεί. Επίσης, τα σχόλια είναι πιο πλούσια ως προς την πληροφορία που περιέχουν. Πολλές έρευνες έχουν αποδείξει την θετική επίδραση που έχουν οι αξιολογήσεις ενός προϊόντος στην διαδικασία λήψης απόφασης αγοράς από έναν άλλον χρήστη. [4]

Προϋπόθεση για τη σωστή και αποτελεσματική λειτουργία ενός συστήματος προτάσεων είναι η γνώση του χρήστη ο οποίος χρησιμοποιεί τον μηχανισμό αυτό. Τα αντικείμενα για τα οποία οι χρήστες εκφράζουν την άποψη τους είτε με ένα like σε μια Facebook page ή ένα follow στο Tweeter μπορούν να χρησιμοποιηθούν για να ανακαλυφθούν τα προσωπικά ενδιαφέροντα και τα αντικείμενα με τα οποία επιθυμεί να αντιστοιχιστεί ο χρήστης. Αυτή η μεγάλη ποσότητα πληροφορίας είναι η αρχή για την ανάπτυξη Recommendation Systems βασισμένα στα ενδιαφέροντα των χρηστών (Interest-Based Recommender System - IBRS). [15]

Το πρώτο στάδιο ανάπτυξης ενός Recommendation System που βασίζεται σε πληροφορίες από τα Social Media είναι η εξόρυξη των αξιολογήσεων των χρηστών και πώς η πληροφορία αυτή μπορεί να συνδυαστεί με τα Linked Open Data ώστε να αξιοποιηθεί στο μέγιστο δυνατό βαθμό. Τα δεδομένα αυτά μπορούν να περιέχουν πληθώρα πληροφοριών για τα προϊόντα ή τις υπηρεσίες που θα προτείνουμε όπως, την χώρα κατασκευής τους, το όνομα του κατασκευαστή κ.α. Τα δεδομένα των αξιολογήσεων από την άλλη σπανίως παρέχουν δομημένες πληροφορίες με συνδέσεις σε άλλα αντικείμενα όπως άλλα προϊόντα για τα οποία έχουν δείξει ενδιαφέρον όσοι έχουν αγοράσει ένα συγκεκριμένο προϊόν. Τα σχόλια αυτά σε συνδυασμό με την πληροφορία που είναι ελεύθερα προσβάσιμη στο διαδίκτυο οδηγούν στην παροχή σωστών προτάσεων προϊόντων ή υπηρεσιών από τα αντίστοιχα εργαλεία. [4]

Ωστόσο δε μπορούν να αντοιχτοιχισθούν όλες οι απόψεις που εκφράζουν οι χρήστες στα social media με κοινά χρησιμοποιούμενα tags ή λέξεις που ανήκουν σε ενα δομημένο σύνολο δεδομένων. Τα αντικείμενα αυτά συνήθως αποτελούν κομμάτι ενός ευρύτερου θέματος. Π.χ. μια δεδομένη χρονική στιγμή ο Cristian Ronaldo είχε 103 εκατομμύρια likes, ενώ κάποιοι γενικότεροι όροι έχουν πολύ λιγότερα, π.χ. το Soccer 66 εκατομμύρια και Football 46 εκατομμύρια. Ωστόσο, οι ευρύτεροι αυτοί όροι είναι πιθανό να αναφέρονται συχνότερα σε άλλα tags ή άλλου είδους περιγραφές, όπως σε ευχετήριες κάρτες, αθλητικό εξοπλισμό κ.α. δυσκολεύοντας μας να συμπεράνουμε

ποιος όρος είναι περισσότερο δημοφιλής. Για να γεφυρωθεί αυτό το κενό στα θέματα γενικού περιεχομένου χρησιμοποιούμε την βάση της DBpedia για τον εμπλουτισμό των πληροφοριών αυτών των θεμάτων τα οποία μπορούν να αποτελέσουν τα ενδιαφέροντα ενός χρήστη. Ο μηχανισμός αυτός αντιμετωπίζει το cold start problem που θα αναφέρουμε παρακάτω. [15]

Τα συστήματα προτάσεων που βασίζονται στα Lined Open Data αντλούν τις πληροφορίες που χρειάζονται από τις ανοιχτές βάσεις διασυνδεδεμένων δεδομένων όπως η DBpedia και Wikidata. Λαμβάνουν υπόψιν τους συνδέσμους μεταξύ των αντικειμένων του Ιστού των Δεδομένων και χρησιμοποιούν τις σχέσεις αυτές για να υπολογίσουν τη σημασιολογική ομοιότητα που υπάρχει μεταξύ των αντικειμένων αυτών. Οι σύνδεσμοι αυτοί μπορούν να είναι απευθείας σύνδεσμοι ή διαδρομές (paths) ανάμεσα στα αντικείμενα που εξετάζουν. [4]

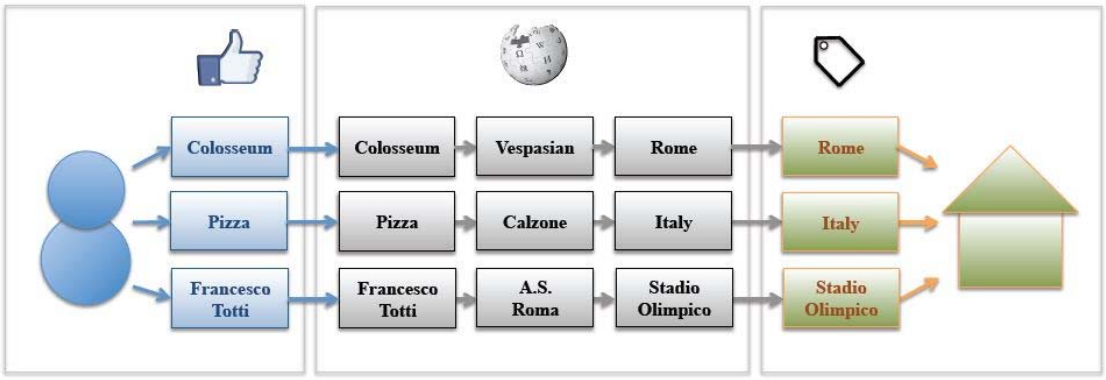
Οι Damljanovic et al. το 2012 εισήγαγαν μια καινοτόμα προσέγγιση η οποία ανακαλύπτει συσχετίσεις μεταξύ αντικειμένων μέσα από σχέσεις ιεραρχίας ή κατά μήκος συνδέσμων όπως οι οριζόντιες ή κάθετες συσχετίσεις στην DBpedia. Οι Heitmann & Hayes πρότειναν ένα Recommendation System το οποίο αξιοποιεί τα Linked Open Data για να αντιμετωπίσει τα προβλήματα των συστημάτων προτάσεων προϊόντων/ υπηρεσιών σε νέους χρήστες, συγκεντρώνοντας και συνδυάζοντας στοιχεία από διαφορετικές πλατφόρμες. Οι Musto et al. μελέτησαν το αντίκτυπο της γνώσης που είναι διαθέσιμη στο Διαδίκτυο των Δεδομένων στην συνολική απόδοση ενός αλγορίθμου προτάσεων που βασίζεται σε γράφο. Οι Vagliano et al. το 2016 παρουσίασαν έναν αλγόριθμο προτάσεων που βασίζεται στα διασυνδεδεμένα δεδομένα και εξετάζει τις σχέσεις που υπάρχουν ανάμεσα στα αντικείμενα αναλύοντας δυναμικά τις κατηγορίες στις οποίες ανήκουν καθώς και τις αναφορές που υπάρχουν σε άλλα αντικείμενα. Οι Di Noia et al. το 2012 επίσης παρουσίασαν μια προσέγγιση η οποία βασίζεται σε ένα μοντέλο που παράγει προτάσεις προερχόμενες από το περιεχόμενο των ανοιχτών διασυνδεδεμένων δεδομένων. Εισήγαγαν έναν υβριδικό αλγόριθμο ο οποίος εξάγει σημασιολογικά χαρακτηριστικά βασισμένα σε συνδέσμους από την DBpedia υπολογίζοντας προτάσεις για τους χρήστες. [4]
Οι βάσεις δεδομένων DBpedia και Wikidata αποτελούν τα σημαντικότερα σύνολα δεδομένων που περιέχουν δεδομένα προερχόμενα από πολλούς και διαφορετικούς τομείς.

Περιορισμοί που υπάρχουν σε όλες τις παραπάνω προσεγγίσεις είναι οι παρακάτω:

1. Εξάρτηση από το θέμα: Τα περισσότερα συστήματα βασίζονται σε βάσεις δεδομένων και social media τα οποία επικεντρώνονται σε συγκεκριμένη θεματολογία. Η αποδέσμευση από τη θεματολογία επιτρέπει την επαναχρησιμοποίηση του μηχανισμού και την μελλοντική βελτίωση του.
2. Εξάρτηση από τη γλώσσα: Αντίστοιχα με την εξάρτηση από τη θεματολογία η εξάρτηση από τη γλώσσα δημιουργεί τους ίδιους περιορισμούς. Ένα κλασικό πρόβλημα των μηχανών προτάσεων είναι η ύπαρξη συνωνύμων, κυρίως όταν

βασιζόμαστε σε χρήση tags. Πολλές σελίδες με διαφορετικό όνομα (π.χ. Football, Soccer) αναφέρονται στο ίδιο θέμα. Το facebook έχει κάνει προσπάθειες να ενώσει σελίδες με το ίδιο περιεχόμενο αλλά που έχουν αναπτυχθεί σε διαφορετική γλωσσά σε ενιαίες σελίδες διευκολύνοντας τους χρήστες να εντοπίσουν τις σελίδες που τους ενδιαφέρουν. Στην περίπτωση αυτή η αναζήτηση πραγματοποιείται είναι ανεξάρτητη από τη γλώσσα που χρησιμοποιείται στη σελίδα. Ωστόσο το πρόβλημα ακόμα υπάρχει και χρήζει αντιμετώπισης.

3. Εξάρτηση από την πλατφόρμα social media: Υπάρχουν διαφορετικές πλατφόρμες κάθε μία από τις οποίες επικεντρώνεται σε ένα συγκεκριμένο σκοπό (Facebook, LinkedIn, Twitter,Instagram, Pinterest). Πχ. ένας αλγόριθμος αναζήτησης θέσεων εργασίας είναι πιθανό να χρησιμοποιήσει δεδομένα από το LinkedIn για να διαμορφώσει τις προτάσεις προς τους χρήστες. Ωστόσο, η απεξάρτηση από το είδος της πλατφόρμας είναι σημαντική προκειμένου να φτιάξουμε συστήματα προτάσεων πλήρως ανεξαρτητοποιημένα.
4. Το πρόβλημα της "κρύας" εκκίνησης (cold start problem) [15]



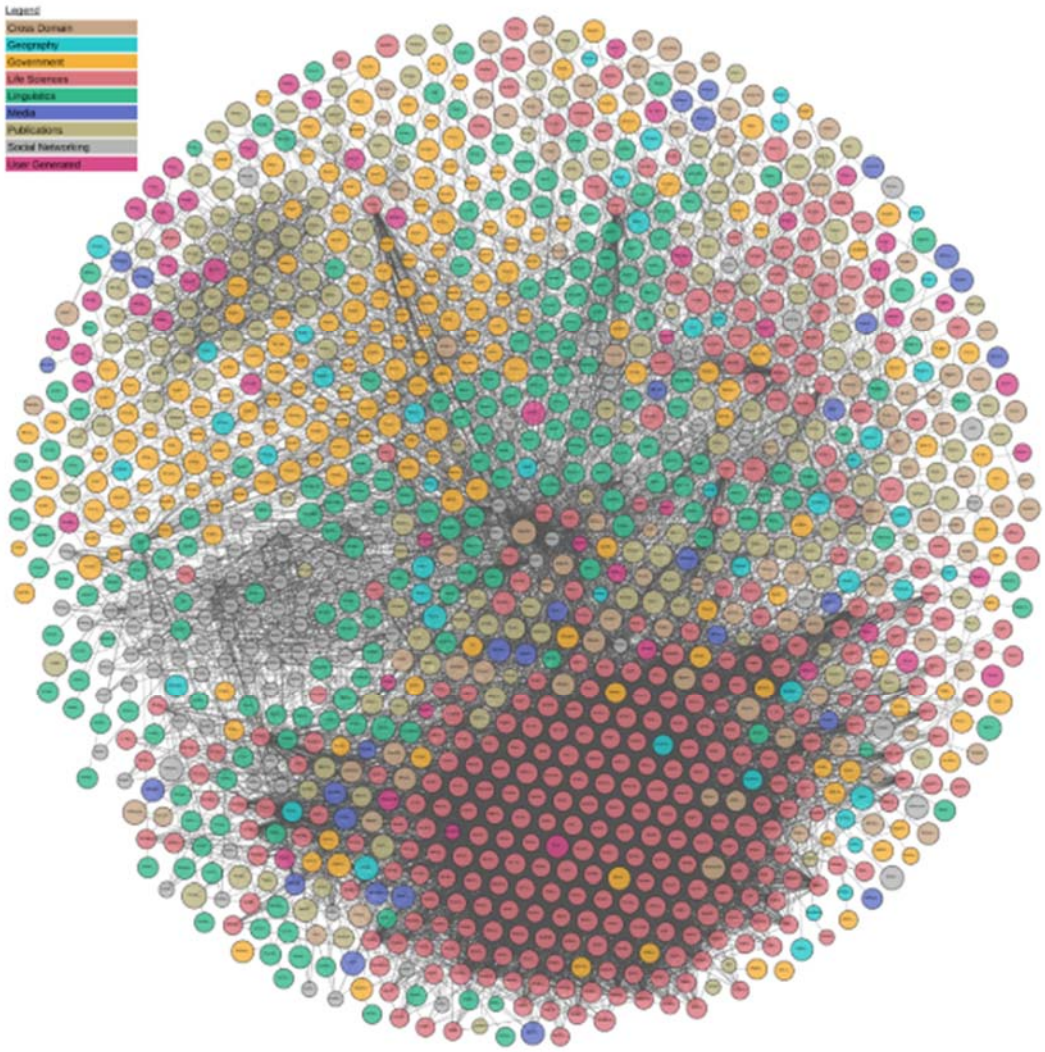
Εικόνα 17 : Πως οι προτιμήσεις των χρηστών στα social media συνδέονται με δεδομένα από τα ανοιχτά διασυνδεδεμένα δεδομένα [15]

2.13 Οι πιο γνωστές Datasets με social media content

Οι Datasets που παραθέτω παρακάτω προέκυψαν έπειτα από εξαντλητική αναζήτηση στο " " που φιλοξενείται στη διεύθυνση https://lod-cloud.net. Πρόκειται για ένα εκτεταμένο διάγραμμα νέφους ανοιχτών διασυνδεδεμένων δεδομένων στο οποίο συμμετέχουν datasets που υπάρχουν δημοσιευμένες σε μορφή διασυνδεδεμένων δεδομένων. Μέχρι και τον Μάιο του 2020 περιλάμβανε 1.301 βάσεις με 16.283 συνδέσμους. Οι διάφορες βάσεις στο νέφος έχουν χωριστεί ως προς τη θεματολογία τους σε εννέα κατηγορίες. Η κάθε μια απεικονίζεται με διαφορετικό χρώμα στο νέφος:

- Cross Domain
- Γεωγραφία
- Κυβερνητικά δεδομένα

- Επιστήμες ζωής
- Γλωσσολογία
- Μέσα μαζικής ενημέρωσης
- Δημοσιεύσεις
- Κοινωνική δικτύωση
- Δεδομένα που δημιουργήθηκαν από χρήστες



Εικόνα 18 : Το διασυνδεδεμένο νέφος (The Linked Open Data Cloud)

Όπως βλέπουμε στην παραπάνω απεικόνιση του Διασυνδεδεμένου Νέφους, με γκρι εμφανίζονται οι κόμβοι που αντιστοιχούν σε βάσεις δεδομένων με περιεχόμενο από τα κοινωνικά δίκτυα. Ακολουθούν οι 6 μεγαλύρερες βάσεις στης κατηγορίας αυτής:

- Social Link: περιλαμβάνει συνολικά 80.101.425 τριπλέτες και οι 643.235 από αυτές συνδέονται με την Dbpedia. Είναι μια ελεύθερης πρόσβασης βάση ανοιχτών διασυνδεδεμένων δεδομένων η οποία "ταιριάζει" λογαριασμούς χρηστών στο twitter με τις αντίστοιχες οντότητες της Dbpedia

(περιλαμβάνονται διαφόρων γλωσσών οντότητες). Συνδέοντας με αποτελεσματικό τρόπο τον κόσμο του Twitter και το νέφος των ανοιχτών διασυνδεδεμένων δεδομένων, η SocialLink επιτρέπει την ανταλλαγή γνώσεις μεταξύ τους. Από τη μία, οι επαγγελματίες του Σημασιολογικού Ιστού μπορούν να συλλέξουν τεράστιες ποσότητες πολύτιμων και ενημερωμένων πληροφοριών που διατίθενται δωρεάν στο Twitter, και από την άλλη, οι ερευνητές των Social Media μπορούν να αξιοποιήσουν τα δεδομένα της DBpedia για να επεξεργαστούν μια αποδοτικά τα θορυβώδη, ημιδομημένα δεδομένα του Twitter.

- sears.com: διαθέτει 75.000.000 τριπλέτες που συνδέονται με 100 τριπλέτες της dbpedia. Πάνω από 15.000.000 σελίδες αντικειμένων της διαθέτουν συνδέσμους εδραιωμένων οντολογιών όπως το πρότυπο GoodRelations σε.
- Linked Crunchbase: 5.000.000 τριπλέτες με συνδέσμους στην dbpedia
- VIVO: διαθέτει 4,514,025 τριπλέτες που συνδέονται με 58 τριπλέτες της dbpedia. Η συγκεκριμένη βάση δημιουργήθηκε από τον NIH (National Institutes of Health) με σκοπό να αποτελέσει ένα είδους "Facebook" σημασιολογικού ιστού για τους επιστήμονες. Χρησιμοποιεί τις τεχνολογίες Σημασιολογικού Ιστού για την μοντελοποίηση επιστημόνων και παρέχει τη δυνατότητα αναζήτησης και ανακάλυψης ερευνητών και συνεργατών στις Ηνωμένες Πολιτείες. Η οντολογία VIVO εκτός από επιστήμονες, επίσης μοντελοποιεί δημοσιεύσεις, πηγές, επιχορηγήσεις, τοποθεσίες και υπηρεσίες. Στην οντολογία VIVO core version 1.0 υπάρχουν 236 κλάσεις με 278 αντικείμενα και 222 ιδιότητες αντικειμένων, ενσωματώνοντας κλάσεις από δημοφιλείς οντολογίες, όπως η BIBO, Dublin Core, Event, FOAF, geopolitical και SKOS. Τα δεδομένα της βασίζονται σε υψηλής ποιότητας δεδομένα προσωπικού χαρακτήρα για τα συνεργαζόμενα πανεπιστήμια υγείας τα οποία προέρχονται κυρίως από ετήσιες εκθέσεις της κάθε σχολής, με πρόσθετες πληροφορίες από δημοσιεύσεις σε Scopus και PubMed. Το δεδομένα της οντολογίας αυτής αφορούν κυρίως επιστήμονες που ειδικεύονται στην έρευνα και τη διδασκαλία.
- Debian Package Tracking System: περιλαμβάνει 1,500,000 τριπλέτες – Τα δεδομένα για όλα τα πακέτα Debian είναι διαθέσιμα σε μορφή RDF μέσω ερωτημάτων στο Debian QA website το οποίο χρησιμοποιεί το πρότυπο ADMS.SW.
- Social Semantic Web Thesaurus: διαθέτει 20,000 τριπλέτες που συνδέονται σε 300 τριπλέτες της dbpedia, 200 τριπλέτες της freebase και 100 τριπλέτες στο geonames-semantic-web. Περιλαμβάνει δεδομένα για ανθρώπους, οργανισμούς, εφαρμογές, τεχνολογίες και οτιδήποτε άλλο σχετίζεται με τον Σημασιολογικό Ιστό.

2.14 Περιορισμοί στην επεξεργασία των δεδομένων από τα social media

Ένα είδος online απάτης που εφαρμόζεται στα δεδομένα των social media και έχει ως σκοπό την παραπληροφόρηση και την χειραγώγηση του κοινού ονομάζεται Απάτη των Κοινωνικών Μέσων (Social Media Fraud). Είναι η διαδικασία της δημιουργίας likes, follows, views ή κάθε είδους άλλης αλληλεπίδρασης με στόχο την εικονική αύξηση των ακολούθων ενός λογαριασμού στα social media προκειμένου να αυξηθεί η επιρροή του. Οι εταιρείες marketing για να αυξήσουν την αναγνωρισιμότητά ενός brand, συνεχώς ανακαλύπτουν νέες τεχνικές για την απόκτηση likes, followers ή σχολίων στους λογαριασμούς των πελατών τους στα social media. Ο αριθμός των παραπάνω μετρικών αποτελεί σημαντικό παράγοντα επιρροής και δημοσιότητας ενός λογαριασμού και η απόκτηση τους είναι μια ιδιαίτερα χρονοβόρα και συχνά όχι προσβάσιμη από όλους διαδικασία. [17]

Ωστόσο, πολλές εταιρείες προκειμένου να δημιουργήσουν την ψευδαίσθηση της αύξησης της δημοτικότητας των social media account τους με πολύ χαμηλό κόστος αγοράζουν likes, follows ή views. Με μία απλή αναζήτηση στις μηχανές αναζήτησης μπορεί κάποιος να βρει τέτοιου είδους υπηρεσίες "απάτης" οι οποίες παρέχουν π.χ. 10.000 twitter followers με κόστος από 40-216 $. Τιμή η οποία είναι ελάχιστη σε σχέση με τον χρόνο και τα χρήματα που χρειάζονται για να αποκτήσει κάποιος τους followers αυτούς. Οι τεχνικές που χρησιμοποιούν είναι 1) η διακύβευση πραγματικών υπαρχόντων profiles και 2) η δημιουργία μαζικών νέων εικονικών profiles. Στην πρώτη περίπτωση οι χρήστες δελεάζονται με την υπόσχεση ότι κάνοντας κλικ σε κάποιο link θα αποκτήσουν δωρεάν followers ωστόσο με την τεχνική αυτή οι υπηρεσίες "απάτης" εκμεταλλευόμενοι τα διαπιστευτήρια των λογαριασμών τους στα social media τους χρησιμοποιούν για να δημιουργήσουν likes, follows ή views σε άλλους χρήστες. Επίσης, μεγάλος αριθμός fake profiles δημιουργούνται από τις λεγόμενες click farms ή botnets. Οι πρώτες αποτελούνται από εργάτες χαμηλότερων κοινωνικά τάξεων οι οποίοι δημιουργούν ψεύτικα προφίλ έναντι ελάχιστης πληρωμής. Τα botnets είναι ομάδες δικτύων υπολογιστών οι οποίοι ελέγχονται απομακρυσμένα από third party οργανισμούς (bot masters). Χρησιμοποιούν ψεύτικα συστήματα και λογαριασμούς οι οποίοι δημιουργούνται αυτόματα και χρησιμοποιούνται για να παράγουν αποκλειστικά likes, follows και views. Αυτά τα δίκτυα είναι γνωστά με την ονομασία social bots, και αποτελούνται από λογαριασμούς social media οι οποίοι ελέγχονται από software που μιμείται πραγματικούς χρήστες και δημιουργεί ψεύτικες ενέργειες στα social media accounts.

Για την καταπολέμηση της εξάπλωσης των fake accounts οι πλατφόρμες social media έχουν αναπτύξει τεχνικές όπως ο έλεγχος CAPTCHAS, η επιβεβαίωση μέσω τηλεφωνικού αριθμού και η IP blacklist. Τέλος, ερευνητές έχουν αναπτύξει εργαλεία και εφαρμογές με τη βοήθεια τεχνικών machine learning και πληροφοριών διασυνδεσιμότητας διαδικτύου για τον εντοπισμό συμπεριφορών που οδηγούν σε απάτες, όπως αυτές που αναφέρθηκαν παραπάνω. Ένα παράδειγμα είναι το εργαλείο "CopyCatch" το οποίο ανιχνεύει τα page likes που έγιναν από spammers στο Facebook βασιζόμενο στον χρόνο που δημιουργήθηκαν τα likes αυτά. Ακόμη ένα εργαλείο είναι το "Synchrotap", το οποίο ανιχνεύει λογαριασμούς οι οποίοι

εμφανίζουν συγχρονισμένα ύποπτη συμπεριφορά όπως το ανέβασμα πολλών spam φωτογραφιων από έναν μικρό αριθμό IP διευθύνσεων ή την απότομη αύξηση του αριθμού followers οι οποίοι προέρχονται από μια συγκεκριμένη ομάδα χρηστών. [17]

Οι τεχνικές εξαπάτησης που αναφέρθηκαν παραπάνω ανεξάρτητα από το πόσο αβλαβής μπορεί να μοιάζουν εκ πρώτης όψεως, έχουν κεντρίσει το ενδιαφέρον των ερευνητών για τους παρακάτω λόγους:

1. Η ανάπτυξη αλγορίθμων για τον εντοπισμό τέτοιου είδους παράνομων ενεργειών στα social media κοστίζει στις εταιρείες αρκετό χρόνο και χρήμα.
2. Διαστρεβλώνει τα δεδομένα και αυξάνει με τεχνητά μέσα την επιρροή συγκεκριμένων social media accounts οδηγώντας σε παραπληροφόρηση των χρηστών. Όταν οι χρήστες εξαπατώνται χάνουν την εμπιστοσύνη τους σε ολόκληρο το οικοσύστημα των social media.
3. Τα botnets τα οποία προκαλούν ένα είδος απάτης στα social media μολύνουν την υποδομή του διαδικτύου στο σύνολό της. Η μόλυνση εκατοντάδων συστημάτων από botnets των οποίων η τελική επίδραση φαίνεται ακίνδυνη μπορεί να γίνει ιδιαίτερα επιβλαβής μακροπρόθεσμα. [17]

Πέρα από τις απάτες οι οποίες επηρεάζουν το τεχνικό κομμάτι της λειτουργίας των social media και διαστρεβλώνουν το περιεχόμενο τους υπάρχει και ένας βαθύτερος ηθικός προβληματισμός ο οποίος προκύπτει από την εκμετάλλευση του περιεχομένου των Social Media και αφορά την αρχή της ιδιωτικότητας. Παρά το γεγονός ότι τα Social Media χαρακτηρίζονται για τη διαδραστικότητα τους, η εξόρυξη δεδομένων από αυτά είναι μια μη διαδραστική διαδικασία καθώς τα δεδομένα αντλούνται και αναλύονται χωρίς να μπορεί ο χρήστης να παρέμβει ή να αλληλοεπιδράσει. Για το λόγο αυτόν είναι πολύ σημαντικό να εφαρμοστούν κατάλληλες πολιτικές για την προστασία των δεδομένων τους. Η διαδικασίες Social Media Data Mining οδηγούν σε όλο και λιγότερη ιδιωτικότητα, χαρακτηριστική είναι η δήλωση του ιδρυτή του Facebook Mark Zuckerberg το 2010 ότι η ιδιωτικότητα στο μέλλον δεν θα αποτελεί τον κανόνα. Την άποψη αυτή ασπάζονται όλες οι εταιρείες κοινωνικής δικτύωσης καθώς κερδίζουν χρήματα πουλώντας το περιεχόμενο των λογαριασμών των χρηστών τους. [2]

Μια βέλτιστη πρακτική την οποία εισήγαγε η έννοια των Open Social Media Data και η οποία εξασφαλίζει την έννοια της δημοκρατικότητας είναι η ελεύθερη πρόσβαση από όλους τους χρήστες στα εργαλεία εξόρυξης δεδομένων καθώς και τα ίδια τα δεδομένα. Οι τεχνικές εξόρυξης δεδομένων θα πρέπει αρχικά να είναι νόμιμες και διαφανείς, να έχουν δημόσια υποστήριξη και το κοινό να συμμετέχει στον τρόπο ρύθμισης τους. Τα Open Data δίνουν τη δυνατότητα επαναχρησιμοποίησης ενός συνόλου δεδομένων τα οποία προσφέρονται με ανοιχτή πρόσβαση για το κοινό καλό και θα πρέπει να συνδυάζονται με εξειδικευμένα εργαλεία ανάλυσης. Είναι απαραίτητος ο προσδιορισμός του ποιος τα έκανε δημόσια προσβάσιμα, σε ποιους είναι ανοιχτά και σε πιο βαθμό είναι επεξεργάσιμα. Στο σημείο αυτό αξίζει να αναφέρουμε τον νέο γενικό κανονισμό της Ευρωπαϊκής Ένωσης για την προστασία

των προσωπικών δεδομένων (General Data Protection Regulation - GDPR) ο οποίος τέθηκε σε εφαρμογή στις 25 Μάϊου 2018. [2]

3 Οντολογία eClassOWL στον κλάδο του ecommerce: Παραδείγματα και εφαρμογές ερωτημάτων

Η οντολογία η οποία θα παρουσιαστεί στην ενότητα αυτή μέσα από παραδείγματα ερωτημάτων με τη γλώσσα turtle είναι η eClassOWL. Αποτελεί μέρος της W3C Web Ontology Language (OWL) και χρησιμοποιείται για να περιγράψει τύπους και ιδιότητες προϊόντων και υπηρεσιών στο σημασιολογικό ιστό. Σε συνδυασμό με την οντολογία GoodRelations αναπαριστά έννοιες του κλάδου του ecommerce όπως προσφορές, ζήτηση προϊόντων, τιμές, πληρωμές, επιλογές αποστολής κ.α.

Η δημιουργία της οντολογίας eClassOWL ξεκίνησε από τον Martin Hepp το 2003 και τώρα φιλοξενείται και συντηρείται από το "E-Business and Web Science Research Group" στο πανεπιστήμιο Bundeswehr του Μόναχο.

Υπάρχουν αρκετά πρότυπα τα οποία χρησιμοποιούνται για την περιγραφή προϊόντων, όπως το UNSPSC,RNTD κ.α. Όμως το πιο διαδεδομένο είναι το eClass,καθώς καλύπτει τη μεγαλύτερη γκάμα προϊόντων και υπηρεσιών (παραπάνω από 30.000 είδη) με τις περισσότερες ιδιότητες προϊόντων (πάνω από 5.000). [30]

3.1 eClassOWL και GoodRelations

Για τη δημοσίευση meta-data για τα προϊόντα ή τις υπηρεσίες ενός ηλεκτρονικού καταστήματος, τα οποία θα είναι αναγνωρίσιμα από μηχανές, χρειάζονται δύο ειδών παράμετροι:

- Το είδος και τα χαρακτηριστικά του προϊόντος (πχ. "οθόνη τηλεόρασης 20 ιντσών")
- Τις λεπτομέρειες τις προσφοράς (πχ. Τιμή και τρόποι πληρωμής)

Βάση των παραπάνω προϋποθέσεων η οντολογία GoodRelations είναι κατάλληλη και χρησιμοποιείτε ευρέως από πολλές εταιρείες του κλάδου του ecommerce όπως οι BestBuy, Yahoo, O'Reilly κ.α. Η οντολογία eClassOWL είναι πλήρως συμβατή με την GoodRelations.
Ακολουθεί ένα απλό παράδειγμα για το πως μπορείς να περιγράψεις ένα αντικείμενο με την eClassOWL 5.1.4. [30]

Το eClass 5.1.4 ID για την κλάση "Pencil" είναι AKF303003
Το eClass 5.1.4 ID για την ιδιότητα "Length" είναι BAF559001
Το eClass 5.1.4 ID για την ιδιότητα "Design of tip state" είναι BAG073001
Το eClass 5.1.4 ID για την τιμή "pointed" είναι BAC386001

- Σύνταξη Turtle:

@prefix eco: <http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#> .
@prefix gr: <http://purl.org/goodrelations/v1#> .

```
@prefix foo: <http://www.mystore.com/semanticweb/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# myPencil is a pencil with a pointed tip and a length of 150 mm

foo:myPencil
   # Class "Pencil": 24-24-01-01 [ AKF303003 ]
   a eco:C_AKF303003-gen ;
   # Length [BAF559001] 150 mm
   eco:P_BAF559001 foo:QuantitativeValueFloat_1;
   # Design of tip state [BAG073001] = "pointed" [BAC386001]
   eco:P_BAG073001 eco:V_BAC386001.

foo:QuantitativeValueFloat_1
   a gr:QuantitativeValueFloat ;
   gr:hasUnitOfMeasurement "MMT"^^xsd:string ;
   gr:hasValueFloat "150.0"^^xsd:float .
```

- Κλάσεις:

Στην οντολογία eClassOWL η κάθε κατηγορία προϊόντος ή υπηρεσίας προσδιορίζεται από δύο κλάσεις. Μία "γενική" (generic) κλάση η οποία αναπαριστά το είδος του προϊόντος ή υπηρεσίας, και μια κλάση "ταξινομίας" (taxonomi) η οποία αναπαριστά ένα πλήθος αντικειμένων τα οποία πιθανών σχετίζονται με το προϊόν.

Στην πράξη ο κανόνας που ακολουθούμε είναι ο εξής:
Όταν περιγράφουμε ένα αντικείμενο χρησιμοποιούμε την "γενική" κλάση για πραγματικά προϊόντα ή μοντέλα προϊόντων και την κλάση "ταξονομίας" για οτιδήποτε άλλο, που δεν είναι προϊόν ή μοντέλο προϊόντος.

Όταν υποβάλουμε ερωτήματα στα δεδομένα RDF, χρησιμοποιούμε τη "γενική" έννοια όταν θέλουμε τα επιστρεφόμενα δεδομένα να προέρχονται από μια συγκεκριμένη κλάση προϊόντων, ενώ την έννοια της "ταξινομίας" αν θέλουμε να επεκτείνουμε το ερώτημα ώστε να συμπεριλάβει όλες τις συσχετιζόμενες κατηγορίες που αποτελούν δευτερεύοντες κόμβους της eClass. Όπως ήδη αναφέρθηκε η οντολογία eClassOWL είναι πλήρως συμβατή με την οντολογία GoodRelations. Στα επόμενα παραδείγματα όλες οι κλάσεις που αναγράφονται ως "gen" είναι rdfs:subClassOf της κλάσης gr:ProductOrService της οντολογίας GoodRelations. [30]

- Ιδιότητες:

Το μετα-μοντέλο GoodRelations, το οποίο αποτελεί το εννοιολογικό θεμέλιο για το eClassOWL, υποστηρίζει τρεις τύπους ιδιοτήτων προϊόντων:

- ποσοτικές ιδιότητες, για τα τυπικά χαρακτηριστικά προϊόντων, υποστηρίζει τιμές με αριθμητικό εύρος.
- ποιοτικές ιδιότητες, για χαρακτηριστικά προϊόντος με προκαθορισμένες τιμές αξίας
- ιδιότητες τύπου δεδομένων, που χρησιμοποιούνται σε περιορισμένα χαρακτηριστικά τα οποία αποτελούν συμβολοσειρές τύπων δεδομένων, όπως ημερομηνία, ώρα, ημερομηνία και ώρα ή boolean.

Ο τομέας της οντολογίας GoodRelations που περιλαμβάνει όλες τις ιδιότητες της eClassOWLείναι ο gr:ProductOrService. [30]

- Τιμές:

Όλες οι προκαθορισμένες τιμές της eCl@ss αναπαρίστανται ως περιπτώσεις της κλάσης gr: QualitativeValue της eClassOWL. [30]

- URI Schema

Όλα τα URIs της eClassOWL προκύπτουν συνδιάζοντας το βασικό URI της eClassOWL 5.1.4 ("http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#") και ένα prefix για τις κλάσεις ("C_"), για τις ιδιότητες (P_) και τις τιμές (V_), το πρωτότυπο αναγνωριστικό της eCl@ss (e.g. "AKF303003"), και την κατάληξη "-gen" ή "-tax" για τον διαχωρισμό των “γενικών” και κλάσεων “ταξονομίας”.

Όπως ήδη αναφέρθηκε υπάρχουν δύο κλάσεις στην eClassOWL, μια “γενική” (generic) κλάση η οποία αναπαριστά το είδος του προϊόντος ή υπηρεσίας, και μια κλάση “ταξινομίας” (taxonomi) η οποία αναπαριστά ένα πλήθος αντικειμένων τα οποία πιθανών σχετίζονται με το προϊόν.

Τα URIs προκύπτουν από το αναγνωριστικό eCl@ss ως εξής:

- Προϊόντα ("generic" Class)
  http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#C_<ID>-gen

- Προϊόντα και συσχετιζόμενα αντικείμενα ("taxonomic" Class)
  http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#C_<ID>-tax

- Ιδιότητες
  http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#P_<ID>

- Ποιοτικές Αξίες
  http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#V_<ID> [30]

3.2 Χρήση και παραδείγματα

Η οντολογία eCl@ss είναι πολύ μεγάλη, επομένως είναι προτιμότερη η αναζήτηση στο Web Site της για να βρούμε τις κατάλληλες κλάσεις και ιδιότητες για την περιγραφή ενός αντικειμένου παρά η αναζήτηση της ίδιας της οντολογίας. Η τρέχουσα έκδοση η οποία είναι λειτουργική είναι η eCl@ss 5.1.4. [30]

Για να περιγράψουμε ένα μολύβι μήκους 150 mm με μυτερή άκρη. Ακολουθούμε τα παρακάτω βήματα:

1. Βρίσκουμε το URI της κλάσης

Πρώτα θα πρέπει να βρούμε το ID της κατάλληλης κλάσης. Με μια αναζήτηση του όρου “μολύβι” στο Web Site της eCl@ss επιστρέφονται ένα πλήθος υποψήφιων κλάσεων. Η ακόλουθη κλάση μοιάζει να ταιριάζει περισσότερο σε αυτό που θέλουμε:

Classification: 24-24-01-01 [AKF303003]

Ο οκταψήφιος κωδικός "24-24-01-01" αντικατοπτρίζει την θέση της κλάσης στην ιεραρχία της οντολογίας, ενώ ο κωδικός "AKF303003" είναι ένα σταθερό αναγνωριστικό της κλάσης αυτής. Οι κωδικοί αυτοί πάντα αποτελούνται από αλφαριθμητικούς χαρακτήρες.
Τώρα πρέπει να αποφασίσουμε αν θέλουμε να χρησιμοποιήσουμε την “γενική” ή την κλάση “ταξονομίας”. Υπάρχει ο απλός κανόνας που αναφέραμε προηγουμένως:
Αν θέλουμε να περιγράψουμε το ίδιο το προϊόν ή ένα μοντέλο προϊόντος χρησιμοποιούμε την “γενική” κλάση.
Για οτιδήποτε άλλο, πχ ένα τιμολόγιο ή μια φωτογραφία που σχετίζεται με ένα προϊόν, χρησιμοποιούμε την κλάση “ταξονομίας”. [30]

Στη συνέχεια συνθέτουμε το URI της κλάσης προϊόντος συνδυάζοντας το βασικό URI της eClassOWL 5.1.4 ("http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#"), ένα πρόθεμα για τις κλάσεις ("C_"), το αρχικό αναγνωριστικό της κλάσης (AKF303003"), και ένα επίθεμα, το "-gen" (generic) για τα πραγματικά προϊόντα και το "-tax" (taxonomic) για οτιδήποτε άλλο.
Στο παράδειγμα μας, το URI που προκύπτει είναι το:

http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#C_ AKF303003-gen ή eco:C_AKF303003-gen

Αν ορίσουμε το πρόθεμα "eco" για το url "http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#".

2. Βρίσκουμε τα URIs των ιδιοτήτων

Στη συνέχεια πρέπει να βρούμε όλα τα IDs των ιδιοτήτων που θέλουμε να χρησιμοποιήσουμε. Στο παράδειγμα μας, θέλουμε να χρησιμοποιήσουμε τις παρακάτω δύο ιδιότητες τις οποίες εντοπίσαμε με αναζήτηση στην οντολογία:

- Την ιδιότητα "Length", για την οποία το eCl@ss ID είναι BAF559001.
- Την ιδιότητα "Design of tip state", για την οποία το eCl@ss ID είναι BAG073001.

Συνθέτουμε τα URIs των ιδιοτήτων αυτών με τον ίδιο τρόπο που συνθέσαμε το URI της κλάσης, συνδυάζοντας το βασικό URI της eClassOWL 5.1.4 ("http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#"), ένα πρόθεμα για τις ιδιότητες ("P_"), και τα αρχικά αναγνωριστικά της κλάσης eCl@ss για τις ιδιότητες αυτές ("BAF559001" and "BAG073001"). [30]
Τα URIs που προκύπτουν είναι τα παρακάτω:

http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#P_BAF559001 and
http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#P_BAG073001
ή

eco:P_BAF559001 and
eco:P_BAG073001.
Ο τομέας της GoodRelations για όλες τις ιδιότητες της eClassOWL είναι ο gr:ProductOrService.

3. Κωδικοποίηση τιμών

Στο τελευταίο βήμα πρέπει να κωδικοποιήσουμε τις τιμές των δύο παραπάνω ιδιοτήτων. Για να το κάνουμε αυτό πρέπει να βρούμε τα κατάλληλα εύρη τιμών για κάθε ιδιότητα. Ο πιο άμεσος τρόπος για να το κάνουμε αυτό είναι η αναζήτηση στο αρχείο HTML της οντολογίας. Εναλλακτικά, μπορούμε να κάνουμε κλικ στο ID της ιδιότητας στην οποία θέλουμε να αντιστοιχίσουμε τη τιμή, όταν κάνουμε αναζήτηση στο Web. Αν η λίστα περιλαμβάνει “Αξίες”, τότε έχουμε να κάνουμε με μια ποιοτική ιδιότητα. Στο παράδειγμα μας, αυτό ισχύει για την ιδιότητα "BAG073001", με τις ακόλουθες τιμές να ορίζονται στην eCl@ss:

BAC386001 - pointed
BAD004001 - unsharpened

Επομένως σημειώνουμε το ID BAC386001 για την ιδιότητα "pointed".

Στην περίπτωση αυτή, μπορούμε απευθείας να συνθέσουμε το URI συνδυάζοντας το βασικό URI της eClassOWL 5.1.4 ("http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#"), ένα πρόθεμα για τις τιμές (values) ("V_"), και το αρχικό αναγνωριστικό της κλάσης eCl@ss ("BAC386001").
Το uri που προκύπτει είναι το παρακάτω:

http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#V_BAC386001
ή
eco:V_BAC386001.

Τέλος, συνδέουμε το URI του προϊόντος με την ιδιότητα eco:P_BAG073001 ("Design of tip state") και στη συνέχεια με το URI της τιμής eco:V_BAC386001.

Για τις ποσοτικές ιδιότητες θα πρέπει να δημιουργήσουμε περιπτώσεις της κλάσεις GoodRelations gr:QuantitativeValueFloat (για τιμές τύπου float) ή gr:QuantitativeValueInteger (για τιμές τύπου integer).

Αυτές συνθέτουν την τιμή της μύτης του μολυβιού και την μονάδα μέτρησης της τιμής. Το πρότυπο UN/CEFACT Common Code χρησιμοποιείται στις περιπτώσεις αυτές. [30]

Στο παράδειγμα μας, η τιμή "150 mm" κωδικοποιείται ως εξής ("MMT" είναι ο κωδικός σύμφωνα με το πρότυπο UN/CEFACT):

```
foo:QuantitativeValueFloat_1
   a gr:QuantitativeValueFloat ;
   gr:hasUnitOfMeasurement "MMT"^^xsd:string ;
   gr:hasValueFloat "150.0"^^xsd:float .
```

Το μόνο που έχουμε να κάνουμε είναι να συνδέσουμε το URI του προϊόντος με την ιδιότητα eco:P_BAF559001 ("Length") στην τιμή foo:QuantitativeValueFloat_1.

Σύμφωνα με όσα αναφέρθηκαν παραπάνω, ένα απλό παράδειγμα με χρήση της γλώσσας turtle είναι το παρακάτω:

```
@prefix eco: <http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#> .
@prefix gr: <http://purl.org/goodrelations/v1#> .

# myPencil είναι ένα μολύβι με μυτερή μύτη και μήκος 150mm
# Το eClass 5.1.4 ID για την κλάση "Pencil" (μολύβι) είναι AKF303003
# Το eClass 5.1.4 ID για την ιδιότητα "Length" (μήκος) είναι BAF559001
# Το eClass 5.1.4 ID για την ιδιότητα "Design of tip state" (σχέδιο μύτης) είναι BAG073001
# Το eClass 5.1.4 ID για την τιμή "pointed" (μυτερός) είναι BAC386001

foo:myPencil
   # Κλάση "Pencil": 24-24-01-01 [ AKF303003 ]
   a eco:C_AKF303003-gen ;
   # Length [BAF559001] 150 mm
```

```
   eco:P_BAF559001 foo:QuantitativeValueFloat_1 ;
   # Design of tip state [BAG073001] = "pointed" [BAC386001]
   eco:P_BAG073001 eco:V_BAC386001 .

foo:QuantitativeValueFloat_1
   a gr:QuantitativeValueFloat ;
   gr:hasUnitOfMeasurement "MMT"^^xsd:string ;
   gr:hasValueFloat "150.0"^^xsd:float .
```

Μερικά ακόμη παραδείγματα:

Παράδειγμα 1: Το μοντέλο ενός μολυβιού

```
foo:Pencil_150
   # Class "Pencil": 24-24-01-01 [ AKF303003 ]
   a eco:C_AKF303003-gen, gr:ProductOrServiceModel ;
   rdfs:label "Heavy-duty pencil, 150 mm long, pointed tip"@en ;
   # Length [BAF559001] 150 mm
   eco:P_BAF559001 foo:QuantitativeValueFloat_1 ;
   # Design of tip state [BAG073001] = "pointed" [BAC386001]
   eco:P_BAG073001 eco:V_BAC386001 .

foo:QuantitativeValueFloat_1
   a gr:QuantitativeValueFloat ;
   gr:hasUnitOfMeasurement "MMT"^^xsd:string ;
   gr:hasValueFloat "150.0"^^xsd:float .
```

Παράδειγμα 2: Μια προσφορά ενός καταστήματος (π.χ. "Miller Stationery LLC", αγοράστε ένα μολύβι για 1€

```
foo:Seller
   a gr:BusinessEntity ;
   gr:legalName "Miller Stationery LLC"@en ;
   gr:offers foo:myOffering1 .

foo:myOffering1
   a gr:Offering ;
   rdfs:comment "We sell one pencil for 1 EURO"@en ;
   gr:includes foo:somePencils ;
   gr:hasBusinessFunction gr:Sell ;
   gr:validFrom "2021-01-01T00:00:00+01:00"^^xsd:dateTime ;
   gr:validThrough "2022-12-31T00:00:00+01:00"^^xsd:dateTime ;
   gr:hasPriceSpecification foo:Price_1 .

foo:Price_1
   a gr:UnitPriceSpecification ;
```

```
   gr:hasCurrency "EUR"^^xsd:string ;
   gr:hasCurrencyValue "1"^^xsd:float ;
   gr:hasUnitOfMeasurement "C62"^^xsd:string .
```

Παράδειγμα 3: Προσφορά 5 μολύβια με 2€

```
foo:Seller
   gr:offers foo:myOffering2 .

foo:myOffering2
   a gr:Offering ;
   rdfs:comment "We sell five pencils for 2 EURO"@en ;
   gr:includesObject foo:TypeAndQttySpec ;
   gr:hasBusinessFunction gr:Sell ;
   gr:validFrom "2010-01-01T00:00:00+01:00"^^xsd:dateTime ;
   gr:validThrough "2012-12-31T00:00:00+01:00"^^xsd:dateTime ;
   gr:hasPriceSpecification foo:Price_2 .

foo:TypeAndQttySpec
   a gr:TypeAndQuantityNode ;
   gr:amountOfThisGood "5.0"^^xsd:float ;
   gr:hasUnitOfMeasurement "C62"^^xsd:string ;
   gr:typeOfGood foo:somePencils .

foo:Price_2
   a gr:UnitPriceSpecification ;
   gr:hasCurrency "EUR"^^xsd:string ;
   gr:hasCurrencyValue "2"^^xsd:float ;
   gr:hasUnitOfMeasurement "C62"^^xsd:string .
```

3.3 SPARQL Queries

Στην ενότητα αυτή, ακολουθούν παραδείγματα χρήσης της eClassOWL για τη σύνταξη ερωτημάτων στον Ιστό των Διασυνδεδεμένων Δεδομένων. Σημειώνω ότι η οντολογίες GoodRelations και eClassOWL συνήθως χρησιμοποιούνται συνδυαστικά. [30]

Όλα τα ερωτήματα χρειάζονται τις ακόλουθες δηλώσεις προθεμάτων:

```
PREFIX eco: <http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#>
PREFIX gr: <http://purl.org/goodrelations/v1#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
```

Παράδειγμα 1: Βρίσκει όλα τα μολύβια που έχουν μήκος τουλάχιστον 100 mm.

```
SELECT ?pencil, ?length WHERE
{
   ?pencil a eco:C_AKF303003-gen.
   ?pencil a gr:ActualProductOrServiceInstance.
   ?pencil eco:P_BAF559001 ?value.
   ?value gr:hasMaxValue ?length.
   ?value gr:hasUnitOfMeasurement "MMT"^^xsd:string.
   # Υπάρχει και η παρακάτω εναλλακτική:
   # ?value gr:hasUnitOfMeasurement ?unit.
   # FILTER (regex(?unit, "^MMT")
   FILTER (?length >=100)
}
LIMIT 10
```

Παράδειγμα 2: Βρίσκει όλα τα μοντέλα προϊόντων μολυβιών που έχουν μυτερή άκρη.

```
SELECT ?model WHERE
{
   ?model a eco:C_AKF303003-gen.
   ?model a gr:ProductOrServiceModel.
   ?model eco:P_BAG073001 eco:V_BAC386001.
   # Design of tip state [BAG073001] = "pointed" [BAC386001]
}
LIMIT 10
```

Παράδειγμα 3: Βρίσκει όλες τις προσφορές που περιλαμβάνουν τουλάχιστον ένα μολύβι με μυτερή άκρη.

```
SELECT ?company, ?offer, ?currency, ?amount, (?amount/?qtty) as
?amount_per_piece WHERE
{
   ?company gr:offers ?offer.
   ?offer a gr:Offering.
   ?offer gr:hasBusinessFunction gr:Sell.
   ?offer gr:hasPriceSpecification ?price.
   ?price a gr:UnitPriceSpecification.
   ?price gr:hasCurrency ?currency.
   ?price gr:hasCurrencyValue ?amount.
   ?offer gr:includesObject ?o.
   ?o a gr:TypeAndQuantityNode.
   ?o gr:amountOfThisGood ?qtty.
```

```
    ?o gr:hasUnitOfMeasurement "C62"^^xsd:string.
    ?o gr:typeOfGood ?type.
    ?type a eco:C_AKF303003-gen.

    ?type a gr:ActualProductOrServiceInstance.
    ?type eco:P_BAG073001 eco:V_BAC386001.
    FILTER (?qtty >=1)
}
```

4 Επίλογος

4.1 Σύνοψη και συμπεράσματα

Μέσα από την ερεύνα που πραγματοποίησα με στόχο τη διερεύνηση της διαθέσιμης βιβλιογραφίας που αφορά το κομμάτι των ανοιχτών διασυνδεδεμένων δεδομένων κατέληξα στο συμπέρασμα ότι βρίσκεται ακόμη σε αρκετά πρώιμο στάδιο. Ελάχιστοι είναι η ερευνητές και ακόμη λιγότεροι οι ειδικοί του ψηφιακού marketing οι οποίοι γνωρίζουν την εν λόγω τεχνολογία και την αξιοποιούν για την ανάπτυξη των ηλεκτρονικών επιχειρήσεων τους. Ωστόσο, όπως ανέφερα ήδη, υπάρχουν μεγάλες επιχειρήσεις όπως η Rolls Royce, Amazon, Google και Netflix οι οποίες κατά καιρούς έχουν αναπτύξει καινοτόμα μοντέλα υπηρεσιών τα οποία βασίζονται στα Big Open Data. [1] Ένα ακόμη κομμάτι στο οποίο αξιοποιούνται τα ανοιχτά διασυνδεδεμένα δεδομένα είναι για τη βελτίωση των μηχανών αναζήτησης (google) αλλά και στην ανάπτυξη recommendation systems, τα οποία παίζουν πολύ σημαντικό ρόλο και επηρεάζουν δραστικά την δημοτικότητα ενός ηλεκτρονικού καταστήματος. Στη βιβλιογραφία που χρησιμοποίησα, όλοι οι ερευνητές καταλήγουν στο κοινό συμπέρασμα ότι τα social media επηρεάζουν σε μεγάλο βαθμό την online καταναλωτική συμπεριφορά. Στη κατεύθυνση αυτή έχουν αναπτυχθεί εργαλεία για την ανάλυση των δεδομένων των social media, social CRMs καθώς και recommendation systems με χρήση social content. Ωστόσο, δεν υπάρχουν ακόμα αρκετές datasets στο νέφος ανοιχτών διασυνδεδεμένων δεδομένων (The Linked Open Data Cloud) τις οποίες μπορούμε να αξιοποιήσουμε για ανάπτυξη εφαρμογών και μελλοντική έρευνα.

4.2 Όρια και περιορισμοί της έρευνας / Προκλήσεις

Πρόκληση αποτελεί η κατανόηση και η εμπιστοσύνη στα Linked Open Data από τα στελέχη των επιχειρήσεων που δραστηριοποιούνται στον κλάδο του ecommerce. Τα αποτελέσματα της εξόρυξης και ανάλυσης των δεδομένων θα πρέπει να παρουσιάζονται με κατανοητό τρόπο μέσα από κατάλληλα εργαλεία απεικόνισης (πίνακες, αναφορές, διαγράμματα κ.α.)

Ο πρώτος περιορισμός εντοπίζεται στη σημασία των δεδομένων τα οποία λόγω των ετερογενών πηγών προέλευσής τους, των διαφορετικών μορφών (formats) που τα αντιπροσωπεύουν και της έλλειψης συνδέσμων μεταξύ τους δυσκολεύουν την αξιοποίηση τούς στο ecommerce. Ακόμη μια δυσκολία υπάρχει στην ανακάλυψη νέων δεδομένων από τα υπάρχοντα. Οι επιπλοκές αυτές μπορούν να εμφανιστούν όπου υπάρχουν δομημένα ή αδόμητα δεδομένα τα οποία προέρχονται από μη έγκυρες πηγές πληροφόρησης. Υπάρχει πλήθος μελετών του σημασιολογικού ιστού οι οποίες επικεντρώνονται στην ανακάλυψη συνδέσμων μεταξύ των δεδομένων και την ανάπτυξη οντολογιών για την ποσοτικοποίηση των δεδομένων, ωστόσο έχουν ελάχιστα αποτελέσματα στον τομέα του ecommerce. [5]

Κατά καιρούς έχουν εκφραστεί ανησυχίες σχετικά με την εκτεταμένη χρήση της εξόρυξης και ανάλυσης δεδομένων και των επιπτώσεων που θα έχει στην κοινωνία γενικότερα. Μερικές από τις εμφανείς επιπτώσεις είναι η αύξηση της παρακολούθησης και μείωση της ιδιωτικότητας, αυξάνονται οι κοινωνικές διακρίσεις και προσφέρονται νέοι τρόποι χειραγώγησης του κοινού. [2]

Μια μεγάλη πρόκληση η οποία έχει παρουσιαστεί με την ραγδαία αύξηση του περιεχομένου των social media είναι η "ανακάλυψη της αλήθειας", δηλαδή η αναγνώριση των έγκυρων και έμπιστων πηγών πληροφόρησης απαλλαγμένων από αφιλτράριστα και κατακερματισμένα δεδομένα τα οποία δημιουργούν παραπληροφόρηση και συγκρουόμενα νοήματα. Για την επίλυση του προβλήματος έχουν αναπτυχθεί πολυάριθμες προσεγγίσεις που βασίζονται σε τεχνικές machine learning, εξόρυξης δεδομένων και network sensing communities. [20]

4.3 Μελλοντικές επεκτάσεις

Θα πρέπει να μελετήσουμε τη σχέση ανάμεσα στην ποιότητα των πληροφοριών του σημασιολογικού ιστού και στον βαθμό ικανοποίησης που λαμβάνουν οι χρήστες. Στην κατεύθυνση αυτή υπάρχουν δύο σημαντικές προκλήσεις που θα πρέπει να αντιμετωπιστούν ώστε να προχωρήσουμε στην λήψη αποφάσεων οι οποίες θα βελτιώσουν την αποτελεσματικότητα των επιχειρήσεων με τη χρήση τεχνολογιών σημασιολογικού ιστού:

- Ο σχεδιασμός προηγμένων αλλά φιλικών προς τον χρήστη εργαλείων ανάλυσης τα οποία θα μπορούν εύκολα να ενσωματωθούν στο περιβάλλον των επιχειρηματικών διαδικασιών μιας εταιρείας του κλάδου του ecommerce.
- Η υιοθέτηση τεχνικών που επιτρέπουν την παραγωγή συνδυασμών συνόλων δεδομένων τα οποία ήταν μη προσβάσιμα από διαφορετικές πηγές και τα οποία πλέον θα μπορούν να αλληλοσυνδεθούν κατάλληλα για να παράγουν κάποιο αποτέλεσμα. [5]

5 Βιβλιογραφία

1. Shahriar Akter & Samuel Fosso Wamba. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research, Electron Markets (2016) 26:173–194 DOI 10.1007/s12525-016-0219-0
2. Helen Kennedy & Giles Moss (2015) Known or knowing publics? Social media data mining and the question of public agency, International Journal of Electronic Commerce, December 2011
3. Liang, Ting-Peng & Turban, Efraim. (2011). Introduction to the Special Issue Social Commerce: A Research Framework for Social Commerce. International Journal of Electronic Commerce. 16. 5-13. 10.2307/23106391.
4. Iacopo Vagliano, Diego Monti, Ansgar Scherp & Maurizio Morisio (2017) Content Recommendation through Semantic Annotation of User Reviews and Linked Data – An Extended Technical Report, Proceedings of the Ninth International Conference on Knowledge Capture, Dec. 4 - 6, 2017
5. Necula, Sabina-Cristiana & Pavaloaia, Vasile & Strimbei, Catalin & Dospinescu, Octavian. (2018). Enhancement of E-Commerce Websites with Semantic Web Technologies. Sustainability. 10. 1955. 10.3390/su10061955.
6. A Riyanto & F A Renaldi (2018) Effect of Social Media on E-Commerce Business, IOP Conf. Series: Materials Science and Engineering 407 (2018) 012033 doi:10.1088/1757-899X/407/1/012033
7. César Pereira Júnior, Vilmar & Fileto, Renato & Souza, Willian & Wittwer, Matthias & Reinhold, Olaf & Alt, Rainer. (2018). A Semantic BI Process for Detecting and Analyzing Mentions of Interest for a Domain in Tweets. 197-204. 10.1145/3243082.3243100.
8. Peska, Ladislav & Vojtáš, Peter. (2013). Enhancing Recommender System with Linked Open Data. 483-494. 10.1007/978-3-642-40769-7_42.
9. Paolo Tomeo, Ignacio Fernández-Tobías, Tommaso Di Noia & Iván Cantador (2016), Exploiting Linked Open Data in Cold-start Recommendations with Positive-only Feedback, CERI '16, June 14-16, 2016, Granada, Spain 2016 ACM. ISBN 978-1-4503-4141-7/16/06…$15.00, DOI: http://dx.doi.org/10.1145/2934732.2934745.
10. Nogales, Alberto. (2013). Exploring the Potential for Mapping Schema.org. Microdata and the Web of Linked Data.
11. Lisa Wenige & Johannes Ruhland (2016), Flexible On-the-Fly Recommendations, Springer International Publishing Switzerland 2016 W. Abramowicz et al. (Eds.): BIS 2016, LNBIP 255, pp. 43-54, 2016. DOI: https://doi.org/10.1007/978-3-319-39426-8 4
12. Vesyropoulos, N., Georgiadis, C. & Pimenidis, E. (2016) Utilizing linked open data for web service selection and composition to support e-commerce transactions, Computational Collective Intelligence: 8th International Conference, ICCCI 2016, Halkidiki, Greece, September 28-30, 2016.

13. Diana Maynard, Ian Roberts, Mark A. Greenwood, Dominic Rout, Kalina Bontcheva (2015), University of Sheffield Department of Computer Science Regent Court
14. Marco Brambilla Stefano Ceri Emanuele Della Valle Riccardo Volonterio Felix Xavier Acero Salazar (2017) Extracting Emerging Knowledge from Social Media, International World Wide Web Conference Committee (IW3C2), April 3–7, 2017
15. Victor de Graaff, Anne van de Venis, Maurice van Keulen & Rolf A. de By (2015), Vienna, Austria, September 20, 2015
16. Kalina Bontcheva a, Dominic Rout (2012) Making Sense of Social Media Streams through Semantics: a Survey, Department of Computer Science, University of Sheffield
17. Masarah Paquet-Clouston, Olivier Bilodeau & David Décary-Hétu (2017), Can We Trust Social Media Data? Social Network Manipulation by an IoT Botnet, July 28-30, 2017, Toronto
18. Alexopoulos, Charalampos & Zuiderwijk, Anneke & Charalabidis, Yannis & Loukis, E. & Janssen, Marijn. (2014). Designing a Second Generation of Open Data Platforms: Integrating Open Data and Social Media. 230-241. 10.1007/978-3-662-44426-9_19.
19. Azad, Hiteshwar. (2016). Linked Open Data Search Engine. 10.1145/2905055.2905075.
20. Zhang, Daniel Yue & Wang, Dong & Vance, Nathan & Zhang, Yang & Mike, Steven. (2018). On Scalable and Robust Truth Discovery in Big Data Social Media Sensing Applications. IEEE Transactions on Big Data. PP. 1-1. 10.1109/TBDATA.2018.2824812.
21. Bernard Marr, How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read, May 21,2018, Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=5771627f60ba
22. INFOGRAPHIC, DOMO, Data Never Sleeps 5.0, Available at: https://www.domo.com/learn/infographic/data-never-sleeps-5
23. Schema.org, Available at: https://schema.org/
24. GoodRelations, W3C, 5 March 2015, Available at: https://www.w3.org/wiki/GoodRelations
25. Olaf Kopp (2017), Unwrapping the Secrets of SEO: How Does Google's Knowledge Graph Work?, SEO & Content Marketing Blog, Available at: https://blog.searchmetrics.com/us/2017/11/16/unwrapping-the-secrets-of-seo-how-does-googles-knowledge-graph-work/
26. Víctor Saquicela, Luis Manuel Vilches Blázquez, Óscar Corcho (2011), Lightweight Semantic Annotation of Geospatial RESTful Services, Available at:https://www.semanticscholar.org/paper/Lightweight-Semantic-Annotation-of-Geospatial-Saquicela-Bl%C3%A1zquez/ac92529a8e2921edca7dd4af86f732eb935e1e92

27. DBedia Ontology, Available at: https://wiki.dbpedia.org/services-resources/ontology
28. YAGO (database), Wikipedia, 27 July 2021, Available at: https://en.wikipedia.org/wiki/YAGO_(database)
29. Lee Feigenbaum (2019), SPARQL By Example A Tutorial, Available at: https://www.w3.org/2009/Talks/0615-qbe/
30. Martin Hepp & Andreas Radinger, eClassOWL - The Web Ontology for Products and Services, Available at: http://www.heppnetz.de/projects/eclassowl/#usage-examples

6 Κατάλογος εικόνων