ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ
ΣΠΟΥΔΩΝ «ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επισκόπηση μεθόδων ανάλυσης δεδομένων με χρήση Σημασιολογικών Στοιχείων και Γράφων Γνώσης

ΑΙΚΑΤΕΡΙΝΗ Δ. ΔΗΜΗΤΡΙΟΥ

ΕΠΙΒΛΕΠΩΝ: κ. Δημήτριος Ασκούνης, Καθηγητής Ε.Μ.Π.

ΟΚΤΩΒΡΙΟΣ 2019

.....

Αικατερίνη Δ. Δημητρίου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός & Μηχανικός Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο

Περίληψη

Ο όγκος των δεδομένων και των πληροφοριών που συναλλάσσονται καθημερινά οι χρήστες του Διαδικτύου είναι ασύλληπτα μεγάλος και ποικιλόμορφος. Αναλύοντας αυτά τα δεδομένα, μπορεί κανείς να συλλέξει αμέτρητες και ανεκτίμητες πληροφορίες, που σε διαφορετική περίπτωση θα ήταν μη προσβάσιμη γνώση.

Την τελευταία δεκαετία, κυρίως, η επιστήμη της ανάλυσης των Μεγάλων Δεδομένων (Big Data) έχει επικεντρωθεί σε αυτό ακριβώς το ζήτημα, προσπαθώντας με διάφορους τρόπους να βελτιώσουν διάφορες μεθόδους Ανάκτησης πληροφοριών (Information Retrieval) από το Διαδίκτυο. Αυτό καθίσταται ευκολότερο με την υιοθέτηση του σημασιολογικού ιστού (Semantic Web), ο οποίος παρουσιάζει τα δεδομένα που υπάρχουν στο διαδίκτυο σε μορφή κατανοητή από τους υπολογιστές ώστε να υποστηρίζονται δυνατότητες μηχανικής εκμάθησης.

Προς την κατεύθυνση αυτή, η παρούσα διπλωματική εργασία μελετά μεθόδους οι οποίες χρησιμοποιούν τη σημασιολογία (Semantics) και τα γραφήματα γνώσης (Knowledge Graphs) για την επεξεργασία και ανάλυση των Μεγάλων Δεδομένων.

Στο 1^{ο} μέρος της παρούσας εργασίας, παρουσιάζονται οι βασικές αρχές πάνω στη Διερεύνηση Δεδομένων (Data Mining) και στην Ανάκτηση πληροφοριών. Γίνεται αναφορά στα Διασυνδεδεμένα Ανοιχτά Δεδομένα (Linked Open Data), στη Διερεύνηση του Διαδικτύου (Web Mining) και ειδικότερα στη Διερεύνηση Δεδομένων του Σημασιολογικού Ιστού (Semantic Web Data Mining).

Στο 2^{ο} μέρος παρουσιάζονται διαφορετικές μέθοδοι για την επεξεργασία των Μεγάλων Δεδομένων. Πρώτα αναλύονται προσεγγίσεις που χρησιμοποιούν το Σημασιολογικό Ιστό για την Διερεύνηση του Διαδικτύου (Semantic web approaches to Web Mining). Στη συνέχεια διερευνάται μια μέθοδος που χρησιμοποιεί τα σημασιολογικά δεδομένα σε συνδυασμό με τις παραδοσιακές μεθόδους αναζήτησης πληροφοριών (Semantics & Traditional Search System) για την αναζήτηση αποτελεσμάτων κατά την ανάλυση των δεδομένων. Ακολουθούν μια προσέγγιση που δίνει έμφαση στη χρήση Οντολογιών για την Διερεύνηση Δεδομένων (Semantic based Ontology approach) και μια μέθοδος που χρησιμοποιεί τα σημασιολογικά δεδομένα μαζί με τα Διασυνδεδεμένα Ανοιχτά Δεδομένα (Semantics & Linked Open Data) για την εξατομίκευση αποτελεσμάτων κατά την επεξεργασία των Μεγάλων Δεδομένων. Τέλος, παρουσιάζονται μέθοδοι που χρησιμοποιούν γράφους γνώσης για

την επεξεργασία των Μεγάλων Δεδομένων: εννοιολογικοί γράφοι με μέτρο ομοιότητας (Graphs of Concepts with a Similarity Measure), γράφοι σημασιολογικής συγγένειας (Semantics Relatedness Graphs) και προσέγγιση που ενσωματώνει τους γράφους γνώσης σε συστήματα συστάσεων (Graph based Methods for Recommender Systems).

Λέξεις κλειδιά: Μεγάλα Δεδομένα, Διερεύνηση Δεδομένων, Ανοιχτά Δεδομένα, Σημασιολογικός ιστός, Γραφήματα Γνώσης.

Abstract

The amount of data and information that Internet users handle daily is innumerable and diverse. By analyzing this data, one can collect countless and invaluable information that would otherwise be inaccessible.

In the last decade in particular, the science of Big Data Analytics has focused on this very issue, trying different ways of improving the various methods of retrieving information from the Internet. This is made easier by adopting the Semantic Web, which present data available online in a comprehensible format by computers, to support machine learning capabilities.

To this end, this thesis studies methods that use Semantics and Knowledge graphs to process and analyze Big Data.

The first part of this paper presents the basic knowledge on Data Mining and Information Retrieval. Reference is made to Linked Open Data, Web Mining and Semantic Web Mining.

The second part presents different methods of processing Big Data. First, an approach that uses Semantics for Data Analysis is presented. Next, a method that uses semantic data in conjunction with traditional search system to search for results in data analysis is explored. Following, a Semantic based Ontology approach is presented as well as a method that uses semantic data along with Linked Open Data to personalize the results during Big Data Analysis is considered. Finally, methods that use Knowledge Graphs for Big Data Analysis are presented: Graphs of concepts along with a similarity measure, Semantic relatedness graphs and Graph-based Method for Recommender systems.

Key Words: Big Data, Data Mining, Linked Open Data, Semantic Web, Knowledge Graphs

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, στο πλαίσιο του Μεταπτυχιακού Προγράμματος Σπουδών «Τεχνο-οικονομικά Συστήματα» κατά τη διάρκεια του τελευταίου εξαμήνου φοίτησης.

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον Επιβλέποντα Καθηγητή κ. Δημήτριο Ασκούνη για την εμπιστοσύνη που μου έδειξε με την ανάθεση αυτής της ενδιαφέρουσας διπλωματικής εργασίας, και την κα. Ευμορφία Μπιλίρη για τη συνεχή επίβλεψη και την άρτια καθοδήγηση κατά τη διάρκεια της εκπόνησης της.

Ιδιαίτερες ευχαριστίες θα ήθελα επίσης να αποδώσω σε όλους τους φίλους και συμφοιτητές μου για την έμπρακτη συμπαράσταση και υποστήριξη που μου παρείχαν καθ' όλη τη διάρκεια των σπουδών μου.

Τέλος, θα ήθελα να εκφράσω τη βαθιά ευγνωμοσύνη μου προς την οικογένειά μου για την πολύτιμη βοήθεια, την ηθική στήριξη, την αμέριστη συμπαράσταση και τη διαρκή ενθάρρυνση.

Αικατερίνη Δ. Δημητρίου

Οκτώβριος 2019

Περιεχόμενα

1. Εισαγωγή 13
   1.1 Αντικείμενο – Σκοπός 13
   1.2 Διάρθρωση διπλωματικής εργασίας 13
2. Data Mining and Information Retrieval 15
   2.1 Εισαγωγή 15
   2.2 Information Retrieval 15
   2.2.1 Μοντέλα IR 16
   2.3 Big Data 16
   2.3.1 Χαρακτηριστικά των Big Data 17
   2.3.2 Επιπτώσεις και Προκλήσεις 18
   2.4 Data Mining 19
   2.4.1 Τύποι Διερεύνησης Δεδομένων 20
   2.4.2 Μέθοδοι Διερεύνησης Δεδομένων 21
   2.4.3 Εργαλεία Διερεύνησης Δεδομένων 21
   2.4.4 Διαδικασία Διερεύνησης Δεδομένων 23
   2.5 Linked Open Data 24
   2.5.1 Διερεύνηση Δεδομένων χρησιμοποιώντας τα LOD 27
   2.6 Semantic Web Mining 29
   2.6.1 Εισαγωγή 29
   2.6.2 Η Αρχιτεκτονική του Σημασιολογικού Ιστού 30
   2.6.3 Διερεύνηση Σημασιολογικού Ιστού 33
   2.6.4 Προκλήσεις 33
   2.7 Knowledge Graphs 34
3. Semantics Related Approaches for Data Mining 36
   3.1 Using Semantic Web and LOD during the Data mining process 36
   3.1.1 Επιλογή – Selection 36
   3.1.2 Προεπεξεργασία – Preprocessing 37
   3.1.3 Μεταμόρφωση – Transformation 38
   3.1.4 Διερεύνηση Δεδομένων – Data mining 39
   3.1.5 Ερμηνεία – Interpretation 40
   3.1.6 Recommender System case study 40
   3.2 Combination of Semantic Web with Traditional Search Systems 41

3.2.1 Μοντέλο Διερεύνησης 42
3.2.2 Ο ρόλος των Οντολογιών στην αναζήτηση 44
3.3 The Ontology Approach 44
3.3.1 Εφαρμογές 46
3.3.2 Γλώσσες και Εργαλεία Οντολογιών 47
3.4 Personalized concept-based search 47
3.4.1 Αρχιτεκτονική και Περιγραφή Συστήματος 48
3.4.2 Αποτελέσματα 50
4. Knowledge Graph related Approaches for Data Mining 52
4.1 Graph-of-concepts Method for a Microblog Semantic Context Retrieval System .. 52
4.1.1 Μέθοδος 52
4.1.2 Αξιολόγηση 54
4.2 Building a relatedness graph from Linked Open Data 55
4.2.1 Μέθοδος 55
4.2.2 Αξιολόγηση 56
4.3 Content-based Recommender System using Semantic Web Knowledge Graph Embeddings 57
4.3.1 Μέθοδος 58
4.3.2 Αξιολόγηση 58
5. Συμπεράσματα 60
6. Βιβλιογραφία 62

1. Εισαγωγή

1.1 Αντικείμενο – Σκοπός

Ο όγκος των δεδομένων και των πληροφοριών που συναλλάσσονται καθημερινά οι χρήστες του Διαδικτύου είναι ασύλληπτα μεγάλος και ποικιλόμορφος. Αναλύοντας αυτά τα δεδομένα, μπορεί κανείς να συλλέξει αμέτρητες και ανεκτίμητες πληροφορίες, που σε διαφορετική περίπτωση θα ήταν μη προσβάσιμη γνώση.

Την τελευταία δεκαετία, κυρίως, η επιστήμη της ανάλυσης των Μεγάλων Δεδομένων (Big Data) έχει επικεντρωθεί σε αυτό ακριβώς το ζήτημα, προσπαθώντας με διάφορους τρόπους να βελτιώσουν διάφορες μεθόδους Ανάκτησης πληροφοριών (Information Retrieval) από το Διαδίκτυο. Αυτό καθίσταται ευκολότερο με την υιοθέτηση του σημασιολογικού ιστού (Semantic Web), ο οποίος παρουσιάζει τα δεδομένα που υπάρχουν στο διαδίκτυο σε μορφή κατανοητή από τους υπολογιστές ώστε να υποστηρίζονται δυνατότητες μηχανικής εκμάθησης.

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η παρουσίαση μεθόδων που χρησιμοποιούν Σημασιολογικά Στοιχεία (Semantics) και Γράφους Γνώσης (Knowledge Graphs) για την ανάλυσης Μεγάλων Δεδομένων. Στόχος της παρούσας εργασίας είναι να αναδείξει τη σπουδαιότητα των διαφορετικών αυτών μεθόδων στην ανάλυση δεδομένων και τα πλεονεκτήματα αλλά και τις ελλείψεις που παρουσιάζουν.

1.2 Διάρθρωση διπλωματικής εργασίας

Εν περιλήψει, τα επιμέρους κεφάλαια τα οποία αναπτύσσονται στο πλαίσιο της παρούσας διπλωματικής εργασίας πραγματεύονται τα εξής:

Στο **2^{ο} κεφάλαιο** της παρούσας εργασίας, παρουσιάζονται οι βασικές αρχές πάνω στη Διερεύνηση Δεδομένων (Data Mining) και στην Ανάκτηση πληροφοριών. Γίνεται αναφορά στα Διασυνδεδεμένα Ανοιχτά Δεδομένα (Linked Open Data), στη Διερεύνηση του Διαδικτύου (Web Mining) και ειδικότερα στη Διερεύνηση Δεδομένων του Σημασιολογικού Ιστού (Semantic Web Data Mining).

Στο **3ο κεφάλαιο** παρουσιάζονται τέσσερις προσεγγίσεις που χρησιμοποιούν τα Σημασιολογικά στοιχεία για την Διερεύνηση Δεδομένων καθώς και τα αποτελέσματά τους. Η πρώτη ασχολείται με τη χρήση του Σημασιολογικού Ιστού για την Διερεύνηση του Διαδικτύου. Η δεύτερη χρησιμοποιεί τα σημασιολογικά δεδομένα σε συνδυασμό με τις παραδοσιακές μεθόδους αναζήτησης πληροφοριών για την αναζήτηση αποτελεσμάτων κατά την ανάλυση των δεδομένων. Η επόμενη δίνει ιδιαίτερη προσοχή στη χρήση των Οντολογιών κατά τη Διερεύνηση του Σημασιολογικού Ιστού. Τέλος, ακολουθεί μια μέθοδος που χρησιμοποιεί τα σημασιολογικά δεδομένα μαζί με τα Διασυνδεδεμένα Ανοιχτά Δεδομένα για την εξατομίκευση αποτελεσμάτων κατά την επεξεργασία των Μεγάλων Δεδομένων.

Στο **4ο κεφάλαιο** παρουσιάζονται τρεις προσεγγίσεις που χρησιμοποιούν τους Γράφους Γνώσης για τη Διερεύνηση Δεδομένων. Παρουσιάζονται Γράφοι εννοιών που χρησιμοποιούν μέτρο ομοιότητας, Γράφοι με σημασιολογικής σχετικότητα και τέλος Γράφοι Γνώσης για συστήματα συστάσεων περιεχομένου.

Τέλος, στο **5ο κεφάλαιο** συνοψίζονται τα βασικά συμπεράσματα που εξάγονται από την παρούσα διπλωματική εργασία με ιδιαίτερη βάση στη μελλοντική κατεύθυνση που μπορεί να λάβει ο τομέας της Διερεύνησης Δεδομένων.

2. Data Mining and Information Retrieval

2.1 Εισαγωγή

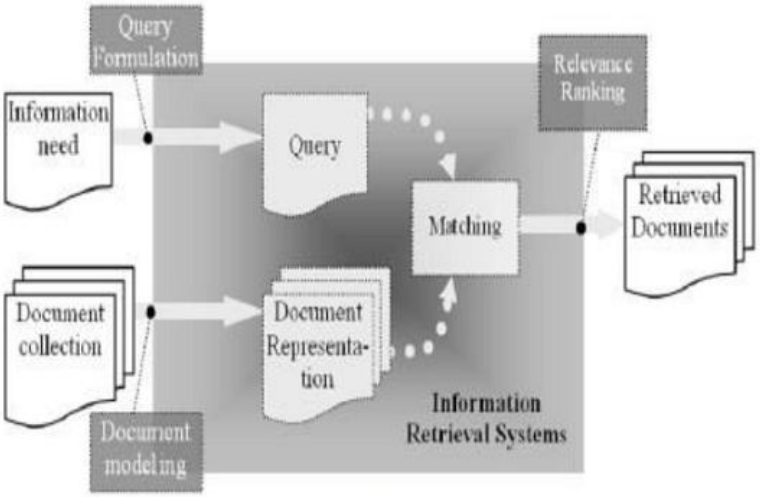
Στην σημερινή εποχή, η δημιουργία δεδομένων δεν συμβαδίζει με την απόκτηση γνώσης και τη διορατικότητα. Μόνο ένα πολύ μικρό ποσοστό των δεδομένων χρησιμοποιείται για την απόκτηση γνώσης. Η κατοχή και χρήση κατάλληλων εργαλείων ανάλυσης δεδομένων είναι το κλειδί στην κατανόηση και σωστή διαχείρισή τους. Για να επιτευχθεί αυτό, τα δεδομένα πρέπει να ταξινομηθούν, να μετασχηματιστούν, να εναρμονισθούν και να υποβληθούν σε επεξεργασία. Ο όγκος, η σύνθεση και η ποικιλομορφία των δεδομένων καθιστά την ανάλυση των διαθέσιμων δεδομένων μία δύσκολη και πολύ σύνθετη διαδικασία [1].

2.2 Information Retrieval

Η διαδικασία της Ανάκτησης Πληροφορίας (Information Retrieval) αποτελεί μια μέθοδο για την αποθήκευση, διαχείριση και αναζήτηση μεγάλου όγκου δεδομένων, τα οποία συνήθως είναι αδόμητα. Η πληροφορία η οποία αποσπάται από τα δεδομένα βρίσκεται με ποικίλους τρόπους και σε διάφορες μορφές.

Κριτήρια αξιολόγησης της κάθε διαδικασίας είναι η αποδοτικότητα, η αποτελεσματικότητα και ο χρόνος εκτέλεσης της αναζήτησης. Σημαντικό ρόλο παίζουν επίσης και η βελτιστοποίηση της διαδικασίας αναζήτησης καθώς και η σωστή κατανομή (indexing) της πληροφορίας.

Η συνήθης διαδικασία της Ανάκτησης Πληροφορίας ξεκινά με ένα απλό ερώτημα από το χρήστη, με βάση τα στοιχεία του οποίου τα αποτελέσματα κατατάσσονται αναλόγως, και τα στοιχεία στην κορυφή της κατάταξης εμφανίζονται πίσω σε αυτόν. Υποβοηθά λοιπόν τον χρήστη στην ορθή οργάνωση της πληροφορίας και στην ανάκτηση νέας σχετικής γνώσης, ανάλογα με το ερώτημα που έχει θέσει [Σχ. 3].



Σχ. 1: Διαδικασία Ανάκτησης Πληροφορίας [12].

2.2.1 Μοντέλα IR

Τα μοντέλα που χρησιμοποιούνται για την Ανάκτηση Πληροφορίας συμβάλλουν στην καθοδήγηση της αναζήτησης και του τρόπου με το οποίο θα υλοποιηθεί ένα σύστημα Ανάκτησης Πληροφορίας. Σημαντικά χαρακτηριστικά των μοντέλων είναι η εξασφάλιση της συνοχής του συστήματος καθώς και η δυνατότητα αναπαραγωγής και υλοποίησης σε πραγματικό χρόνο.

Διαδεδομένα μοντέλα είναι τα ακόλουθα: μοντέλο Boolean, μοντέλο Περιοχής, μοντέλο Διανυσματικού Χώρου, Πιθανοτικό μοντέλο, μοντέλο 2-Poisson, μοντέλα δικτύων Bayesian, Γλωσσικό μοντέλο, Google’s Page Rank μοντέλο και τέλος μοντέλο Εξελικτικού υπολογισμού[12].

2.3 Big Data

Ένας από τους πιο σημαντικούς ρόλους της Ανάκτησης Πληροφορίας είναι η απόσπαση πληροφοριών από τα Μεγάλα Δεδομένα με μεγαλύτερη ταχύτητα και αποτελεσματικότητα.

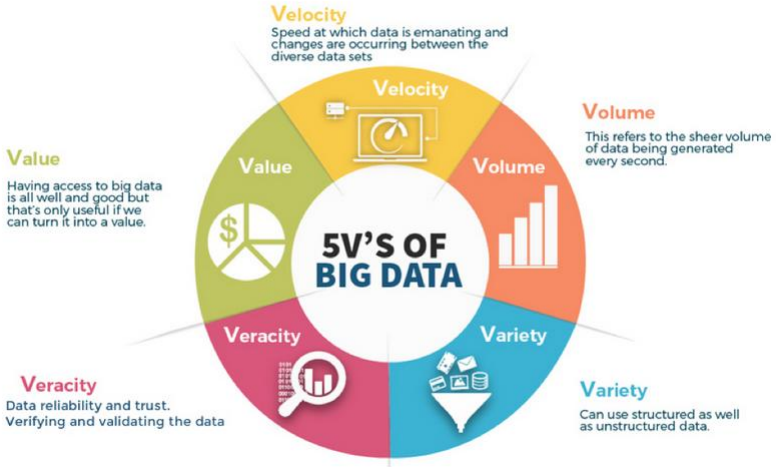
Ως Μεγάλα Δεδομένα χαρακτηρίζεται ένας τεράστιος όγκος από πληροφορία, η οποία αυξάνεται συνεχώς με ταχύτατους ρυθμούς. Σε αυτό έχει συμβάλει ιδιαίτερα η ανάπτυξη των τεχνολογιών Internet of Things (IoT), όπου η ανάλυση και η διαχείριση

των δεδομένων που παράγουν είναι καίρια στην επιτυχημένη εξέλιξή τους. Από τα Big Data 1.0 του ηλεκτρονικού εμπορίου, και τα Big Data 2.0 του Web 2.0 και των social media, βρισκόμαστε πλέον στην εποχή των Big Data 3.0 που βασίζεται κυρίως σε εφαρμογές IoT.

Συσκευές και αισθητήρες δημιουργούν και διαμοιράζονται δεδομένα σε πραγματικό χρόνο χωρίς την επέμβαση του ανθρώπου. Τα δεδομένα αυτά αποτελούν συχνά μη δομημένα δεδομένα, και πολλές πλατφόρμες, όπως το Hadoop, έχουν επικεντρώσει την έρευνά τους στην ανάλυσή τους και στην παροχή νέων διαδικασιών που μπορούν να μετατρέψουν αυτά τα Μεγάλα Δεδομένα σε ουσιαστική γνώση.

2.3.1 Χαρακτηριστικά των Big Data

Οι κυριότερες ιδιότητες των Μεγάλων Δεδομένων είναι οι ακόλουθες [12]:



Σχ. 2: Τα 5 V των Μεγάλων Δεδομένων [13].

- Όγκος – Volume: Ο όγκος των δεδομένων που δημιουργούνται καθημερινά είναι τεράστιος, ποικιλόμορφος και συνεχώς αυξανόμενος. Εκτιμάται ότι τα δεδομένα που δημιουργούνται ετησίως είναι της τάξεως των Zettabytes (10^9 Terabytes).
- Ταχύτητα – Velocity: Ο ρυθμός δημιουργίας και επεξεργασίας των δεδομένων είναι αυτός που έχει ωθήσει την αναζήτηση νέων και καλύτερων μοντέλων ανάλυσης δεδομένων. Πλέον δίνεται ιδιαίτερη σημασία στην ανάλυση πραγματικού χρόνου των Μεγάλων Δεδομένων, καθώς μπορεί να εξαχθούν πολύτιμες πληροφορίες σχεδόν τη στιγμή δημιουργίας των δεδομένων.
- Ποικιλία – Variety: Τα διαφορετικού τύπου δεδομένα που δημιουργούνται είναι συνήθως αδόμητα. Μπορούν να ενσωματώνουν Word, pdf, text ή

multimedia format, καθώς και σχεσιακά ή ημιδομημένα XML δεδομένα κ.α.. Είναι απαραίτητο αυτά τα δεδομένα να μετατραπούν σε δομημένη μορφή, ώστε να μπορούν αν αποθηκευθούν, ανακληθούν, αναλυθούν και να υποστούν επεξεργασία από τη Διερεύνηση Δεδομένων.

- Φιλαλήθεια – Veracity: Αναφέρεται στην αξιοπιστία και την κατανόηση των δεδομένων, ώστε να μπορεί να εξαχθεί ποιοτική πληροφορία και ορθό αποτέλεσμα.
- Αξία – Value: Αποτελεί τον απώτερο στόχο της Διερεύνησης Δεδομένων, να αποδώσει πρόσθετη αξία σε δεδομένα που πριν ήταν αδιάφορα.

Καθώς η τεχνολογία που ασχολείται με τα Μεγάλα Δεδομένα και την επεξεργασία τους εξελίσσεται, έτσι και οι διαστάσεις τους αλλάζουν. Εκτός από τις προαναφερθείσες πέντε βασικές ιδιότητές τους, στη βιβλιογραφία [14] γίνεται αναφορά και στις ακόλουθες:

- Μεταβλητότητα – Variability: Αναφέρεται στη μεταβολή του ρυθμού ροής των δεδομένων, η οποία έχει μεγάλη επίπτωση στο κόστος και στην ικανότητα διαχείρισης του όγκου των δεδομένων.
- Πολυπλοκότητα – Complexity: Αφορά των αριθμό των πηγών δεδομένων, καθώς όσο αυξάνεται η πολυπλοκότητά τους τόσο πιο δύσκολη είναι η συλλογή, επεξεργασία και αποθήκευση τους, χωρίς να υπάρχει επανάληψη πληροφορίας.
- Φθορά – Decay: Χαρακτηρίζει την πτώση στην αξία των δεδομένων με την πάροδο του χρόνου. Εξαιτίας αυτού, η ανάλυση σε πραγματικό χρόνο είναι ιδιαίτερα σημαντική, καθώς σε αρκετές περιπτώσεις απαιτείται άμεση επεξεργασία ώστε να παρθούν ορθές αποφάσεις (π.χ. παρακολούθηση ασθενών ή παρακολούθηση περιβαλλοντικών αλλαγών).

2.3.2 Επιπτώσεις και Προκλήσεις

Η ανάλυση των Μεγάλων Δεδομένων προσφέρει μεγάλες δυνατότητες και πλεονεκτήματα σε επιχειρήσεις και οργανισμούς που δραστηριοποιούνται στο τομέα αυτό. Εξατομικευμένο μάρκετινγκ, καλύτερη τιμολογιακή πολιτική, μείωση κόστους, βελτιωμένη εξυπηρέτηση πελατών είναι μερικοί τομείς που επηρεάζονται άμεσα.

Συμβάλει επίσης στην καλύτερη λήψη αποφάσεων και στην αύξηση της ποιότητας των προσφερόμενων προϊόντων και υπηρεσιών.

Ωστόσο, παρά τα προφανή πλεονεκτήματα, υπάρχουν ακόμη σημαντικές προκλήσεις που καλούνται να λύσουν όσοι ασχολούνται με τα Μεγάλα Δεδομένα, τόσο στο τεχνικό όσο και στο διαχειριστικό κομμάτι.

Η διατήρηση της ποιότητας των δεδομένων και η διαχείριση τους, η ασφάλεια των δεδομένων, το απόρρητο, οι επενδύσεις, και η έλλειψη εξειδικευμένων επιστημόνων είναι μερικές από αυτές.

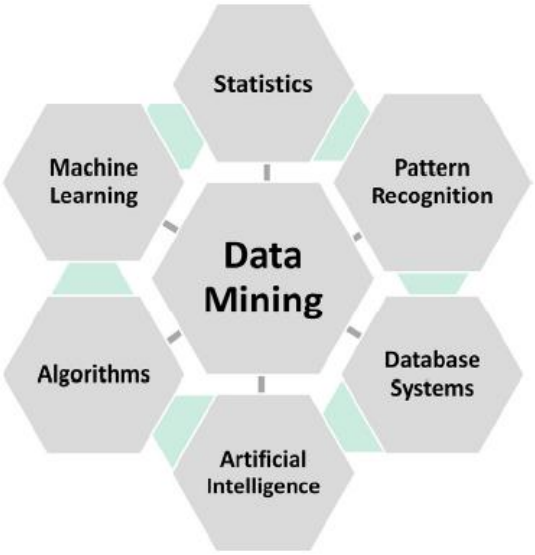
Επιπρόσθετα, οι επιχειρήσεις θα πρέπει να αντιμετωπίσουν προκλήσεις όσον αφορά την ταχύτητα επεξεργασίας, την ερμηνεία, την ποιότητα και την παρουσίαση των Μεγάλων Δεδομένων [14].

Τα δεδομένα που καλούνται να επεξεργαστούν είναι κυρίως ετερογενή με ποικίλες διαστάσεις, οι πηγές από τις οποίες προέρχονται είναι αυτόνομες με αποκεντρωμένο έλεγχο και οι σχέσεις μεταξύ τους είναι σύνθετες και εξελισσόμενες.

2.4 Data Mining

Η Διερεύνηση Δεδομένων, ή αλλιώς data mining, μπορεί να περιγραφεί ως μια μη τετριμμένη διαδικασία αναγνώρισης έγκυρων, νέων, δυνητικά χρήσιμων και εν τέλει κατανοητών προτύπων/κανόνων στα δεδομένα [2]. Ο κλάδος της Διερεύνησης Δεδομένων ασχολείται κυρίως με την ανάλυση σύνθετων δεδομένων με σκοπό την απόκτηση γνώσης, την αναζήτηση δομών, μοτίβων και κανόνων σε μεγάλες και πολύπλοκες βάσεις δεδομένων.

Το ιδιαίτερο χαρακτηριστικό είναι ότι αναζητά στοιχεία μέσα από πολλές, διαφορετικές, πιθανές και ως επί το πλείστον άγνωστες υποθέσεις [3]. Απαιτεί δε, τη ζεύξη πολλών διαφορετικών κλάδων πεδίων ανάλυσης, όπως στατιστική, αναγνώριση προτύπων, βάσεων δεδομένων, τεχνητή νοημοσύνη και αλγόριθμοι μηχανικής μάθησης [Σχ. 3]



Σχήμα 3: Κλάδοι που συμμετέχουν στην Διερεύνηση Δεδομένων [1]

2.4.1 Τύποι Διερεύνησης Δεδομένων

Γενικά, συναντώνται πολλοί διαφορετικοί τύποι Διερεύνησης Δεδομένων, με κριτήριο το αντικείμενο που μελετούν και τους στόχους τους [1]. Πιο αναλυτικά:

- Διερεύνηση Δεδομένων: Περιλαμβάνει την ανάλυση αριθμητικών και κατηγορηματικών δεδομένων σε μεγάλα και σύνθετα σύνολα δεδομένων.

- Διερεύνηση Κειμένου: Περιλαμβάνει αλγόριθμους ανάλυσης λεξιλογικού και γραμματικού περιεχομένου κειμένων. Εκτός από κείμενα μπορεί να χρησιμοποιηθεί για την ανάλυση σημειώσεων, ερευνών, γραφημάτων, σημειώσεων, φόρουμ και παρουσιάσεων.

- Διερεύνηση Ιστού: Συνίσταται στην εφαρμογή μεθόδων Διερεύνησης Δεδομένων σε πληροφορία που έχει συνταχθεί στο Διαδίκτυο. Διαφοροποιείται σε Διερεύνηση περιεχομένου, δομής και χρήσης ιστού ανάλογα με το που επικεντρώνεται κάθε φορά η μελέτη.

- Διερεύνηση Εικόνας: Έχει ως στόχο την ανάλυση και εξαγωγή μοτίβων τα οποία όμως δεν είναι εμφανή και προφανή στις εικόνες.

- Διερεύνηση Δεδομένων Φωτογραφίας, Βίντεο και Μουσικής: Ασχολείται κυρίως με την αναγνώριση χαρακτηριστικών σε δεδομένα βίντεο,

φωτογραφίας και μουσικής. Ιδιαίτερη σημασία δίνεται στην ταχεία ανάκτηση, ταξινόμηση και παρακολούθηση περιεχομένου πολυμέσων.

- Διερεύνηση Δεδομένων Χρονοσειρών: Διερευνώνται οι χρονικές σχέσεις στα δεδομένα μέσω ειδικής εξίσωσης απόστασης, όπως η δυναμική χρονική στρέβλωση, ώστε να αναγνωρισθούν ομοιότητες κατά την εξέλιξη των χρονοσειρών.

- Διερεύνηση Χωρικών Δεδομένων: Στόχος είναι η αναγνώριση προτύπων σε μεγάλα, πολυδιάστατα σύνολα χωρικών δεδομένων όπως δημιουργούνται από τεχνικές τηλεπισκόπησης στην παρατήρηση της γης.

2.4.2 Μέθοδοι Διερεύνησης Δεδομένων

Υπάρχουν πολλές μέθοδοι Διερεύνησης Δεδομένων [4]. Οι πιο συχνά χρησιμοποιούμενες είναι οι ακόλουθες [1]:

- *Association and sequence analysis*: Με την ανάλυση συσχέτισης μπορούν να βρεθούν και να ποσοτικοποιηθούν σχέσεις μεταξύ δύο αντικειμένων.

- *Grouping and clustering*: Σκοπός της είναι η ενοποίηση όμοιων αντικειμένων και η διακριτή διαφορά μεταξύ διαφορετικών ομάδων.

- *Regression*: Η παλινδρόμηση χρησιμοποιείται για την εκτίμηση ή την πρόβλεψη μεταβλητών [3].

- *Classification*: Σκοπός της είναι η ταξινόμηση αντικειμένων για τα οποία δεν υπήρχε πρωτύτερα τέτοιου είδους πληροφορία.

2.4.3 Εργαλεία Διερεύνησης Δεδομένων

Τα πιο γνωστά εργαλεία Διερεύνησης Δεδομένων είναι τα εξής: RapidMiner, R, Orange, WEKA και KNIME.

i. RapidMiner [5], [6]: Είναι ένα open-source εργαλείο, το οποίο υποστηρίζει διάφορες μορφές δεδομένων προς επεξεργασία από τον Ιστό Δεδομένων. Συνεισφέρει στον σχεδιασμό της Διερεύνησης Δεδομένων από το χρήστη, χωρίς την απαίτηση γνώσεων προγραμματισμού, χρησιμοποιώντας προ-

εγκατεστημένες διεργασίες, οι οποίες ονομάζονται *operators*. Κάθε operator εκτελεί μια συγκεκριμένη λειτουργία στα δεδομένα, και ο χρήστης το μόνο που έχει να κάνει είναι να δημιουργήσει τις σχέσεις αλληλουχίας μεταξύ των operator που έχει επιλέξει.

ii. R [7]: Η γλώσσα R και τα εργαλεία της έχουν χρησιμοποιηθεί ευρέως στην στατιστική ανάλυση δεδομένων καθώς και στη διερεύνησή τους. Προαπαιτείται η γνώση της γλώσσας για την χρήση της, γεγονός που ισοσταθμίζεται από το πλήθος των διαθέσιμων οδηγών καθώς και με τη χρήση γραφικού περιβάλλοντος πολύ φιλικό προς τον χρήστη.

iii. Orange [8]: Είναι ένα λογισμικό, γραμμένο στη γλώσσα Python, το οποίο χρησιμοποιείται για μηχανική εκμάθηση και διερεύνηση δεδομένων, βασιζόμενο σε εξαρτήματα που ονομάζονται widgets. Τα widgets προσφέρουν σημαντικές λειτουργίες όπως εμφάνιση πίνακα δεδομένων και επιλογή λειτουργιών, ανάγνωση δεδομένων, προγνωστικά κατάρτισης και σύγκριση αλγορίθμων μάθησης, οπτικοποίηση δεδομένων κλπ..

iv. WEKA [9]: Αποτελεί ένα ελεύθερο εργαλείο, το οποίο δημιουργήθηκε από το Πανεπιστήμιο του Waikato, και είναι ιδανικό για την Διερεύνηση δεδομένων και τη μοντελοποίηση τους. Είναι γραμμένο σε γλώσσα Java και μπορεί να εκτελέσει σημαντικές εργασίες όπως διερεύνηση, επεξεργασία, ταξινόμηση, παλινδρόμηση κλπ. πάνω στα υπό εξέταση δεδομένα.

v. KNIME [10]: Είναι ένα εργαλείο μηχανικής εκμάθησης γραμμένο σε Java, το οποίο έχει αναπτυχθεί από το Πανεπιστήμιο της Konstanz και σκοπός του είναι να προσφέρει υψηλή λειτουργική διάρθρωση και είναι επομένως εύκολο στη χρήση χωρίς πρότερη γνώση προγραμματισμού. Η κάθε λειτουργία που προσφέρει αντιπροσωπεύεται από ένα “knot”, το οποίο ο χρήστης μπορεί να το συνδυάσει με τα υπόλοιπα για να φτιάξει τη διαδικασία που επιθυμεί. Ένα από τα κύρια χαρακτηριστικά του KNIME είναι η ευκολία εκμάθησής του καθώς και η ταχύτητα υλοποίησης και κλιμάκωσης.

Σχ. 4: Data mining tools [11].

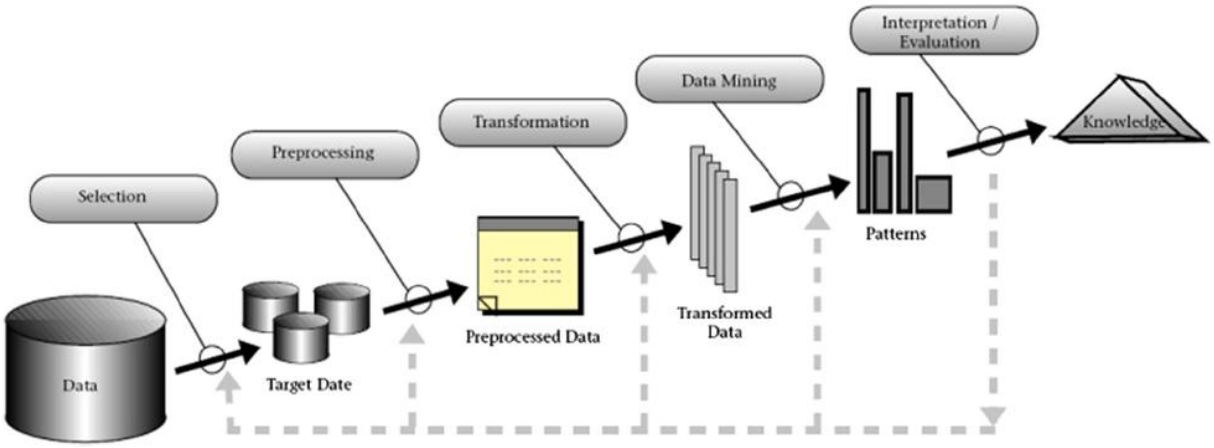
2.4.4 Διαδικασία Διερεύνησης Δεδομένων

Ανεξάρτητα με το ποια μέθοδο ή ποια εργαλεία χρησιμοποιούνται, υπάρχουν πέντε βασικά βήματα τα οποία οδηγούν από ακατέργαστα δεδομένα σε χρήσιμη γνώση αξιοποιήσιμη από το χρήστη [2], [15]. Τα πέντε βήματα είναι τα εξής:

1. Επιλογή: Το πρώτο βήμα είναι η επιλογή των δεδομένων που έχουμε σκοπό να χρησιμοποιήσουμε για τη διερεύνηση και απόκτηση γνώσης. Αυτό προϋποθέτει την κατανόηση και τον ορισμό του στόχου της Διερεύνησης Δεδομένων από τη πλευρά του τελικού χρήστη.
2. Προεπεξεργασία: Τα δεδομένα που επιλέγονται στο 1ο βήμα υποβάλλονται σε επεξεργασία με τρόπο που επιτρέπει μεταγενέστερη ανάλυση. Συνήθως εδώ διορθώνονται οι κενές τιμές, ο θόρυβος και τα σφάλματα, οι διπλές εγγραφές καθώς και ο έλεγχος, η αντιστοίχηση και η ένωση των δεδομένων που προέρχονται από διαφορετικές πηγές.
3. Μεταμόρφωση: Στη συνέχεια τα δεδομένα μετατρέπονται σε μία μορφή αναγνωρίσιμη και κατανοητή από τους αλγόριθμους διερεύνησης δεδομένων.
4. Διερεύνηση Δεδομένων: Σε αυτό το βήμα επιλέγεται και υλοποιείται η μέθοδος Διερεύνησης, αναζητούνται μοτίβα ενδιαφέροντος με μία συγκεκριμένη μορφή ή σειρά παραστάσεων, όπως σύνολα κανόνων ή δέντρα.
5. Αξιολόγηση και Ερμηνεία: Το τελευταίο βήμα εξετάζονται τα μοντέλα και μοτίβα που προέκυψαν από τη διαδικασία όσον αφορά την εγκυρότητά τους.

Ο τελικός χρήστης αξιολογεί επίσης και την αξία της γνώσης που δημιουργήθηκε για τη συγκεκριμένη εφαρμογή.

Σε κάθε βήμα υπάρχει η δυνατότητα οπισθοχώρησης και επανάληψης προηγούμενων βημάτων, σε περίπτωση που απαιτείται εκ νέου παραμετροποίηση ή ακόμη και πλήρης αλλαγή της μεθόδου που θα χρησιμοποιηθεί, έως ότου η ποιότητα των αποτελεσμάτων είναι ικανοποιητική. Στο ακόλουθο σχήμα φαίνεται γραφικά ολόκληρη η διαδικασία.



Σχ. 5: Διαδικασία Ανάκτησης Γνώσης από Δεδομένα [15].

2.5 Linked Open Data

Ο όρος Linked Open Data (LOD) - Ανοιχτά Διασυνδεδεμένα Δεδομένα, αναφέρεται σε κάθε ανοιχτή, αλληλένδετη συλλογή συνόλων δεδομένων, σε μορφή κατανοητή από μηχανές και καλύπτει πολλούς τομείς, ώστε να μπορεί να χρησιμοποιηθεί σε όλα τα στάδια της Διερεύνησης Δεδομένων [15]. Στην ουσία, τα LOD αναφέρονται σε δεδομένα τα οποία έχουν δημοσιευθεί στο Web, με συγκεκριμένη μορφή ώστε να μπορούν να προσπελασθούν από μηχανές, η σημασία τους είναι ρητά ορισμένη, είναι συνδεδεμένα με διάφορα σύνολα δεδομένων, και με τη σειρά τους μπορούν και αυτά να συνδεθούν από εξωτερικά σύνολα.

Βασική προϋπόθεση των LOD είναι τα έγγραφα να περιέχουν δεδομένα σε μορφή RDF (Resource Description Framework), η οποία είναι μία γενική μέθοδος περιγραφής πληροφοριών. Τα δεδομένα αυτά πρέπει να διέπονται από τις ακόλουθες βασικές αρχές [16]:

- Να χρησιμοποιούν URIs, Uniform Resources Identifiers, για την ονομασία αντικειμένων.
- Να χρησιμοποιούν HTTP URIs ώστε οι χρήστες να μπορούν να αναζητήσουν αυτές τις ονομασίες.
- Να παρέχεται χρήσιμη πληροφορία μέσω προτύπων RDF.
- Να περιλαμβάνονται σύνδεσμοι για άλλα URIs, ώστε να μπορούν να ανακαλύψουν επιπλέον στοιχεία.

Τα URIs προσδίδουν στα δεδομένα ένα πιο γενικό μέσο για την αναγνώριση τους. Το πρωτόκολλο HTTP παρέχει ένα κοινά αποδεκτό μηχανισμό ανάκτησης πόρων που μπορούν να απεικονισθούν ως μία σειρά από bytes ή ανάκτησης περιγραφών οντοτήτων που δεν μπορεί να πραγματοποιηθεί ο διαμοιρασμός τους μέσω του Διαδικτύου.

Συμπληρωματικό των URI και HTTP είναι το μοντέλο RDF. Παρέχει ένα γενικό μοντέλο απεικόνισης δεδομένων που βασίζεται σε γράφους με το οποίο να δομηθούν και να συνδεθούν τα δεδομένα.

Το μοντέλο RDF κωδικοποιεί τα δεδομένα σε μορφή τριπλέτων που περιλαμβάνουν υποκείμενο, σχέση και αντικείμενο. Το υποκείμενο και το αντικείμενο είναι πάντα URIs που αντιστοιχούν σε κάποιο πόρο, ή ένα URI και μία σειρά αντίστοιχα. Η σχέση καθορίζει το πώς συνδέονται μεταξύ τους, και συνήθως αναπαρίσταται και αυτή από URI.

Χρησιμοποιώντας τα URI και HTTP για τον προσδιορισμό πόρων, το HTTP πρωτόκολλο ως μηχανισμός ανάκτησης δεδομένων και το μοντέλο RDF για την αναπαράσταση της περιγραφής των πόρων, τα LOD δομούνται απευθείας με βάση την αρχιτεκτονική του Διαδικτύου. Τα Ανοιχτά Διασυνδεδεμένα Δεδομένα δημιουργούν ένα διαδίκτυο δεδομένων το οποίο περιλαμβάνει όλους τους τύπους δεδομένων, είναι προσπελάσιμο από οποιονδήποτε χρήστη και ο κάθε ένας μπορεί να

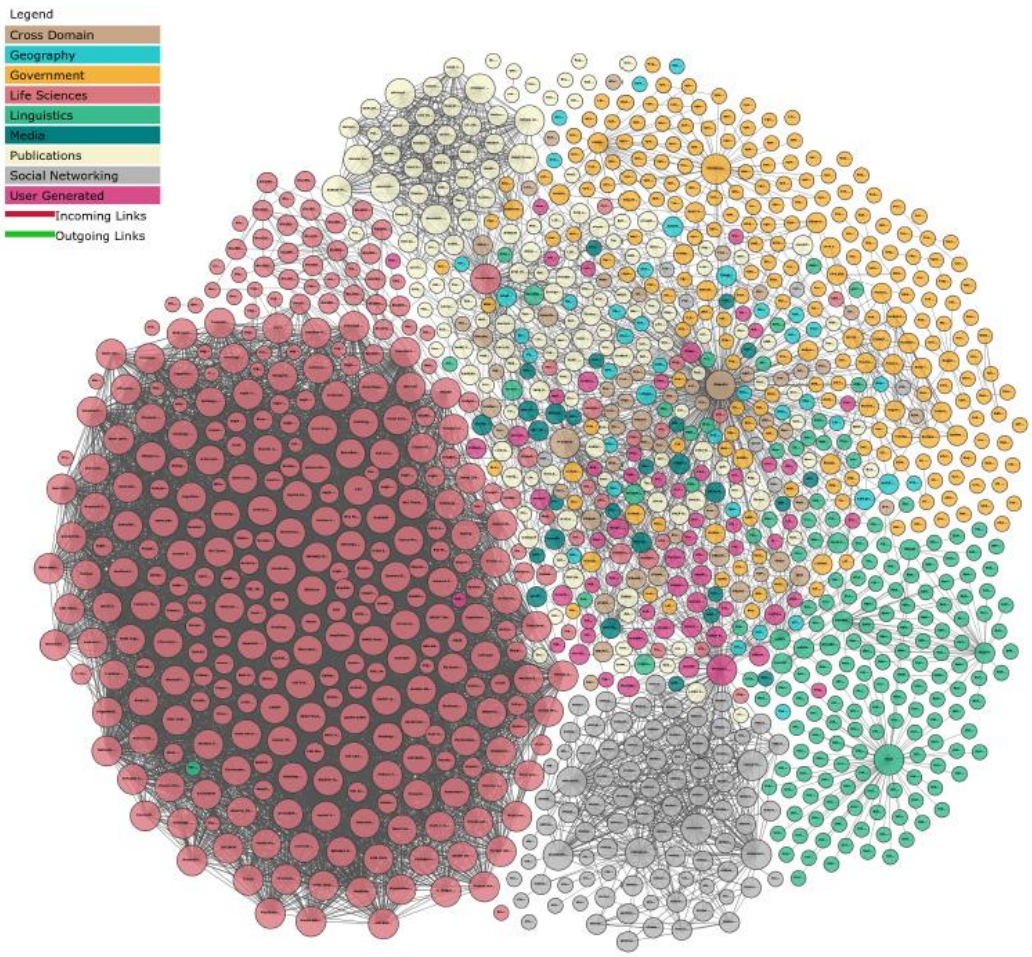
δημοσιεύσει δεδομένα, και τα δεδομένα μεταξύ τους συνδέονται μέσω των RDF συνδέσμων, δημιουργώντας ένα παγκόσμιο γράφο δεδομένων. Ο κάθε χρήστης μπορεί να δημιουργήσει και να δημοσιεύσει και το δικό του Λεξιλόγιο, δηλαδή συλλογή από κλάσεις και ιδιότητες σε RDF μορφή ώστε να περιγράψουν αντικείμενα και τη σχέση τους, χρησιμοποιώντας την γλώσσα RDF Vocabulary Definition Language (RDFS) και την γλώσσα Web Ontology Language (OWL). Δεν υπάρχει περιορισμός στην επιλογή Λεξιλογίου και το καθένα μπορεί να συνδεθεί με τα υπόλοιπα ορίζοντας έτσι αντιστοιχίες μεταξύ διαφορετικών Λεξιλογίων [16].

Σχετικά με την ανάπτυξη εφαρμογών βασιζόμενες στα LOD, τα πλεονεκτήματα που προσδίδουν είναι τα εξής:

- ✓ Υπάρχει αυστηρός διαχωρισμός ανάμεσα στη μορφή και στην παρουσία των δεδομένων.
- ✓ Τα δεδομένα είναι αυτό-προσδιοριζόμενα. Μέσω των URIs μπορούν να βρεθούν οι σύνδεσμοι που επεξηγούν τα δεδομένα σε περίπτωση που το Λεξιλόγιο δεν είναι κατανοητό.
- ✓ Η χρήση του HTTP ως μηχανισμού ανάκτησης δεδομένων και του RDF ως μοντέλου αναπαράστασης δεδομένων απλοποιούν την πρόσβαση στα δεδομένα σε σχέση με άλλες Web APIs.

Οι εφαρμογές που βασίζονται στα LOD μπορούν να ανακαλύπτουν νέες πληροφορίες και πηγές σε πραγματικό χρόνο ακολουθώντας τους συνδέσμους RDF.

Επί του παρόντος, περίπου 1000 σετ δεδομένων είναι αλληλοσυνδεδεμένα μέσω του σύννεφου LOD, με το πλήθος των συνδέσμων να ενώνουν ταυτόσημες οντότητες σε δύο σύνολα δεδομένων. Η δομή του σύννεφου φαίνεται στο [Σχ. 6].



Σχ. 6: Το διάγραμμα των LOD[20].

Τρεις είναι οι πιο συνήθεις εφαρμογές που χρησιμοποιούνται για την αναζήτηση των LOD: DBPedia [21], YAGO [22] και Wikidata [23], οι οποίες χρησιμοποιούν δεδομένα από τη Wikipedia ώστε να δημιουργήσουν την βάση LOD δεδομένων τους.

2.5.1 Διερεύνηση Δεδομένων χρησιμοποιώντας τα LOD

Για να μπορεί να εκφραστεί καλύτερα η πληροφορία μιας ιστοσελίδας, χρησιμοποιείται εδώ και χρόνια ο όρος οντολογίες (Ontologies), δηλαδή αυστηρά καθορισμένες έννοιες και οι σχέσεις μεταξύ τους για κάθε domain[15]. Οι οντολογίες διακρίνονται σε:

- Οντολογίες Ιστοσελίδων: Εκφράζουν βασικές γνώσεις σχετικά με τον τομέα των δεδομένων στα οποία θα υλοποιηθεί η Διερεύνηση.

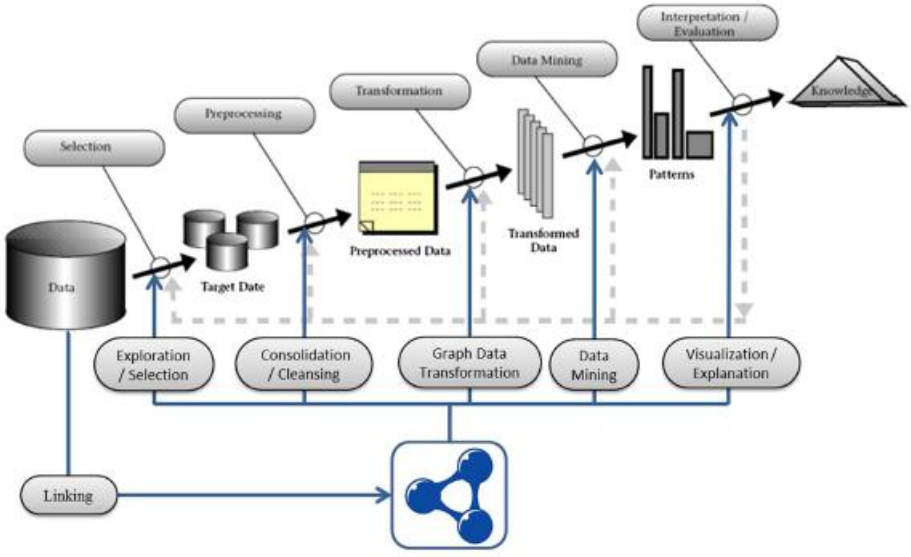
- Οντολογίες για Διερεύνηση Δεδομένων: Καθορίζουν πληροφορίες για τη διαδικασία, τα βήματά της, τον αλγόριθμό της καθώς και όλες τις πιθανές παραμέτρους.

- Οντολογίες μεταδεδομένων: Περιγράφουν πληροφορίες για τα μεταδεδομένα, όπως προέλευση, ιδιότητες κλπ., και τις διαδικασίες που χρησιμοποιήθηκαν για την δημιουργία συγκεκριμένων σετ δεδομένων.

Αυτές οι οντολογίες μαζί με τα Ανοιχτά Διασυνδεδεμένα Δεδομένα συμβάλουν σε όλα τα στάδια της Διερεύνησης Δεδομένων και την ενισχύουν.

Τα LOD βασίζονται, όπως ανωτέρω αναφέρθηκε, στην ιδέα ότι όλα τα δεδομένα είναι αλληλοσυνδεδεμένα και προσβάσιμα στο Διαδίκτυο μέσω σημασιολογικών δικτύων. Σημασιολογικό δίκτυο είναι μια οπτική αναπαράσταση της γνώσης, όπου η τριπλέτα του RDF μοντέλου αναπαρίσταται γραφικά.

Στο ακόλουθο σχήμα φαίνεται πώς αλλάζει η διαδικασία της ανάκτησης γνώσης όταν εισαχθούν LOD δεδομένα.



Σχ. 7: Διαδικασία ανάκτησης γνώσης από δεδομένα βασισμένη σε LOD[15].

Όπως φαίνεται από το σχήμα, το πρώτο βήμα είναι η επιλογή των δεδομένων. Μετά την επιλογή μπορούν να εξερευνηθούν όλες οι υφιστάμενες σχέσεις του αρχικού σετ, το οποίο έχει συνδεθεί με τα αντίστοιχα αντικείμενα του LOD σετ.

Στο βήμα της προεπεξεργασίας, μπορούν να χρησιμοποιηθούν αρκετές τεχνικές για την ενοποίηση και τον «καθαρισμό» των δεδομένων, όπως αντιστοίχηση, συγχώνευση, κανονικοποίηση τιμών κλπ..

Στη συνέχεια τα δεδομένα μετασχηματίζονται, ώστε να αντιπροσωπεύουν δεδομένα που να μπορούν να επεξεργασθούν με οποιονδήποτε διαθέσιμο αλγόριθμο ανάλυσής τους.

Μόλις ολοκληρωθεί η μετατροπή, επιλέγεται και εκτελείται ο αλγόριθμος Διερεύνησης Δεδομένων, τα αποτελέσματα του οποίου παρουσιάζονται στον τελικό χρήστη για αξιολόγηση.

Το μεγαλύτερο πλεονέκτημα της χρήσης LOD έναντι «παραδοσιακών» διαδικασιών Διερεύνησης Δεδομένων έγκειται στο γεγονός ότι με «απλές» μεθόδους Διερεύνησης, τα μοτίβα και η πρόσθετη γνώση μπορεί να προέλθει μόνο από κλειστά σετ δεδομένων, όπου όλη η πληροφορία που απαιτείται για την ανάλυση σχετίζεται με ήδη γνωστές σχέσεις. Ο ρόλος του αναλυτή επηρεάζει σε μεγάλο βαθμό το αποτέλεσμα, καθώς εάν επιλέξει ακατάλληλες μεταβλητές για το μοντέλο, τότε υπάρχει ο κίνδυνος να μην αναγνωρισθούν σχέσεις και μοτίβα λόγω της ελλιπής αρχικής επιλογής. Χρησιμοποιώντας LOD δίνεται στον χρήστη η δυνατότητα να συλλάβει και να καθορίσει άγνωστα, έως τώρα, μοτίβα και σχέσεις στα δεδομένα, χρησιμοποιώντας την αντιστοίχιση τους με σετ δεδομένων που διατίθενται στο Διαδίκτυο.

2.6 Semantic Web Mining

2.6.1 Εισαγωγή

Η Διερεύνηση Δεδομένων και η ανακάλυψη άγνωστων πληροφοριών γίνεται μέσω του Web 3.0. Εδώ οι πληροφορίες παρουσιάζονται με σαφή και διαρθρωμένο τρόπο και επιτρέπουν την συνεργασία μηχανής και ανθρώπου. Τα δεδομένα είναι διασυνδεδεμένα μέσω οντολογιών, εύκολα αναγνώσιμα από τις μηχανές, και είναι δυνατός ο διαμοιρασμός τους αλλά και η επεξεργασία τους από αυτοματοποιημένα εργαλεία καθώς και ανθρώπους.

Το δίκτυο του Σημασιολογικό Ιστού (Semantic Web) διαθέτει διάφορα επίπεδα. Η περιγραφή της πληροφορίας και του τύπου γίνεται μέσω των RDF και RDF Schema, τα ερωτήματα τίθενται σε γλώσσα κατάλληλη να προσπελάσει RDF δεδομένα (π.χ.SPARQL), και το Λεξιλόγιο καθορίζει την κοινή γνώση και περιγράφει τις σημασιολογικές σχέσεις μεταξύ των δεδομένων. Είναι σημαντικό η πληροφορία να είναι αναγνωρίσιμη από μηχανές και εργαλεία διότι μπορεί να καθοδηγήσει την αναζήτηση σε εφάμιλλη γνώση και να βελτιώσει την ακρίβεια και την ποιότητα[17].

Η Διερεύνηση Ιστού (Web mining) στοχεύει στην ανακάλυψη γνώσης σχετικά με την έννοια και τη χρήση των δεδομένων του Διαδικτύου, η οποία, λόγω της ποικιλομορφίας που συναντάται σε αυτά και της κυρίως αδόμητης φύσης τους, αποτελεί μια δύσκολη διαδικασία. Συνδυάζει διερεύνηση κειμένου και δεδομένων και τα εφαρμόζει στο Διαδίκτυο, το οποίο αποτελεί τη μεγαλύτερη πηγή πληροφορίας παγκοσμίως.

Υπάρχουν τρεις κατηγορίες Διερεύνησης Ιστού: α) Διερεύνηση Περιεχομένου Ιστού, β) Διερεύνηση Δομής Ιστού, και τέλος γ) Διερεύνηση Χρήσης ιστού. Η πρώτη αναφέρεται στην άντληση γνώσης από τις ιστοσελίδες και άλλα αντικείμενα του διαδικτύου. Η δεύτερη χρησιμοποιεί τη Δομή των συνδέσεων μεταξύ των Ιστοσελίδων και των λοιπών στοιχείων για την απόκτηση πληροφορίας. Τέλος, η Διερεύνηση Χρήσης Ιστού ασχολείται με την επεξεργασία των μοτίβων χρήσης των Ιστοσελίδων των χρηστών του Διαδικτύου.

Η πρώτη κατηγορία είναι αυτή που θα μας απασχολήσει, καθώς ο Σημασιολογικός Ιστός παρέχει ένα ευέλικτο πλαίσιο για την ανάκτηση γνώσης βάσει περιεχομένου.

2.6.2 Η Αρχιτεκτονική του Σημασιολογικού Ιστού

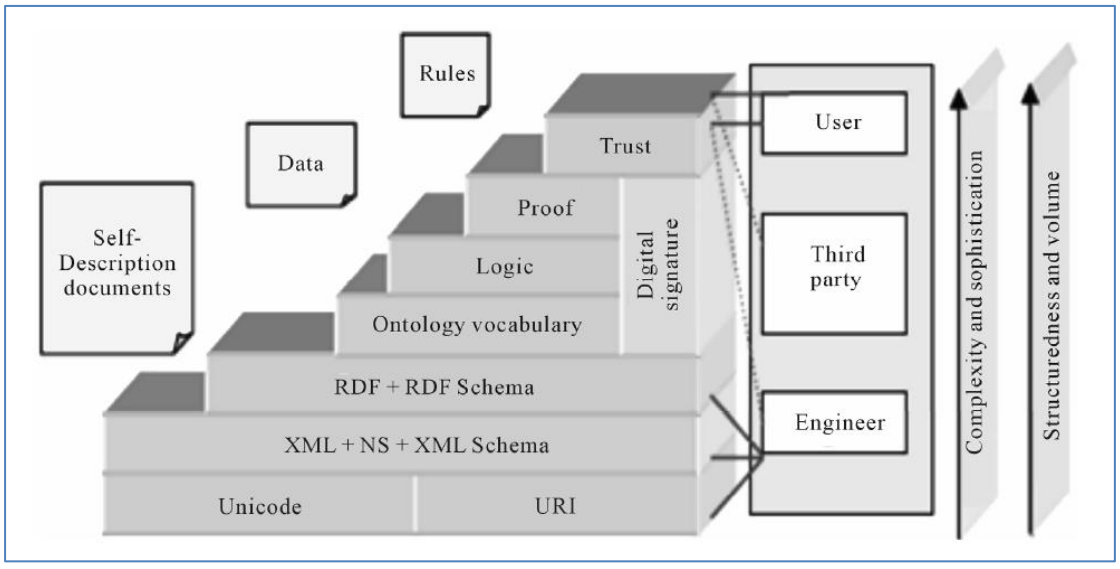
Ο κύριος σκοπός του Σημασιολογικού Ιστού είναι να αποδώσει σημασία στα δεδομένα που προέρχονται από διαφορετικές πηγές του Διαδικτύου, ώστε να μπορούν οι μηχανές να ερμηνεύσουν και να κατανοήσουν αυτά τα εμπλουτισμένα δεδομένα, να απαντήσουν με ακρίβεια και να ικανοποιήσουν τα αιτήματα του χρήστη, επιτρέποντας τη διαχείριση της γνώσης με αυτόματο τρόπο[18,19].

Ο στόχος του είναι να αναπτυχθούν δυναμικά πρότυπα και τεχνολογίες που θα συμβάλλουν στην αυτόματη επεξεργασία των δεδομένων του Διαδικτύου από

μηχανές ώστε να υποστηρίξουν πλουσιότερη ανακάλυψη, ενσωμάτωση, πλοήγηση και αυτοματοποίηση των εργασιών. Παρέχει έτσι ένα κοινό πλαίσιο που επιτρέπει την κοινή χρήση και επαναχρησιμοποίηση των δεδομένων σε όρια εφαρμογών, επιχειρήσεων και κοινοτήτων[19].

Αποσκοπεί στη λύση δύο σημαντικών προβλημάτων, τον περιορισμό στην πρόσβαση σε δεδομένα στο Διαδίκτυο και την ανάθεση κατηγοριών προβλημάτων. Παρέχει πρόσβαση σε δεδομένα κατά όλο το μήκος του Διαδικτύου και επιτρέπει την ανάθεση έργων ώστε να λυθούν γρήγορα προβλήματα όπως η ενοποίηση πληροφορίας κ.α.[18].

Στο ακόλουθο σχήμα [Σχ. 8] παρουσιάζεται μία προτεινόμενη αρχιτεκτονική του Σημασιολογικού Ιστού. Το Δίκτυο αυτό χωρίζεται σε επτά επίπεδα: URI, XML – NS – XML Schema, RDF & RDF Schema, Ontology Vocabulary, Logic, Proof και Trust[18, 19].



Σχ. 8: Τα επίπεδα της αρχιτεκτονικής του Σημασιολογικού Ιστού [18].

Το επίπεδο **URI** είναι υπεύθυνο για την διαδικασία κωδικοποίησης των πόρων και την αναγνώρισή της. Ορίζει με σαφήνεια ένα αναγνωριστικό για να αντιπροσωπεύει κάθε πόρο με ομοιόμορφο τρόπο, προσδιορίζοντας σχήματα αναπαράστασης συμπεριλαμβανομένων τάξεων, ιδιοτήτων και ατόμων. Η σαφήνεια καθιστά δυνατή την συγκέντρωση της πλήρους πληροφορίας ενός αντικειμένου και κάνει ευκολότερη την ενσωμάτωση πηγών που έχουν δημιουργηθεί ανεξάρτητα.

Το επίπεδο **XML** ασχολείται με το διαχωρισμό του περιεχομένου των δεδομένων, της μορφής τους και της γλωσσικής επίδοσης τους, καθώς και με την αναπαράστασή τους χρησιμοποιώντας μια κοινή γλώσσα μορφοποίησης. Χρησιμοποιεί μεταδεδομένα για να περιγράψει τον τύπο και τη μορφή των δεδομένων. Είναι η βάση για την οργάνωση των δεδομένων στο διαδίκτυο, από άποψη μορφής και όχι σημασιολογίας. Συμβάλει στην περιγραφή των δεδομένων ώστε να είναι κατανοητά από τα ανώτερα επίπεδα και να γίνουν διαλειτουργικά.

Το **RDF & RDF Schema** επίπεδο καθορίζει σημασιολογικά την πληροφορία και τον τύπο της. Παρέχει ένα τυποποιημένο πλαίσιο για την υποβολή δηλώσεων σχετικά με τους πόρους και τις ιδιότητές τους. Μαζί με το XML επίπεδο δημιουργούν το βασικό επίπεδο γλώσσας μορφοποίησης των δεδομένων. Για την έκφραση νοήματος, χρησιμοποιείται το Λεξιλόγιο της RDF Schema γλώσσας, η οποία με τη χρήση κλάσεων και ιδιοτήτων, ενισχύει τον ορισμό και παρέχει πρόσθετα περιγραφικά χαρακτηριστικά. Και σε αυτό το επίπεδο δημιουργούνται μεταδεδομένα για τα ανώτερα στρώματα.

Το επίπεδο **Ontology Vocabulary** επικεντρώνεται στην αποκάλυψη της σημασιολογίας μεταξύ των πληροφοριών, καθορίζοντας την κοινή γνώση και τις σημασιολογικές σχέσεις ανάμεσα σε διαφορετικά είδη πληροφορίας, δημιουργώντας έτσι έναν δίκτυο νοήματος. Οι Οντολογίες είναι χρήσιμες για να εκπροσωπούν σαφώς τα αντικείμενα και τις σχέσεις μεταξύ τους, για να γίνεται κατανοητό το νόημά τους από τις μηχανές αλλά και για να διευκολύνεται η ανταλλαγή πληροφοριών.

Το επίπεδο **Logic** προσδίδει αξίες και τεκμήρια αξιωμάτων στην πληροφορία, θέτοντας τη βάση για έξυπνες υπηρεσίες. Χρησιμοποιείται ως ένα πλαίσιο για την εξαγωγή νέων συμπερασμάτων και πώς αυτά τα συμπεράσματα θα πρέπει να εκφραστούν για την υλοποίηση του Σημασιολογικού Ιστού.

Τέλος, τα δύο τελευταία επίπεδα, **Proof & Trust**, ασχολούνται με τον εμπλουτισμό της ασφάλειας του δικτύου, χρησιμοποιώντας μηχανισμούς κρυπτογράφησης και ψηφιακής υπογραφής για να καταγράφουν και να αναγνωρίζουν αλλαγές στα δεδομένα. Στόχος τους είναι να επαληθεύουν την αξιοπιστία των αποτελεσμάτων και να παρέχουν ένα μηχανισμό εμπιστοσύνης μεταξύ πληροφορίας και χρήστη.

2.6.3 Διερεύνηση Σημασιολογικού Ιστού

Ο όρος Διερεύνηση Σημασιολογικού Ιστού (Semantic Web Mining), περιγράφει ακριβώς αυτή την νέα μορφή που λαμβάνουν τα δεδομένα. Συνδυάζει το Σημασιολογικό Ιστό, και την προσπάθεια να γίνει ο τεράστιος όγκος δεδομένων επεξεργάσιμα από μηχανές, με τη Διερεύνηση Ιστού, και την διαδικασία εξαγωγής γνώσης που δεν είναι εμφανής. Αντί για Διερεύνηση δεδομένων, η Διερεύνηση Σημασιολογικού Ιστού επικεντρώνεται στην άντληση γνώσης από το Διαδίκτυο[17].

Η Διερεύνηση Σημασιολογικού Ιστού επικεντρώνεται στην εξόρυξη γνώσης από το Διαδίκτυο. Λόγω του τεράστιου όγκου διαθέσιμων σημασιολογικών δεδομένων, είναι αναγκαία η αναθεώρηση και εφαρμογή των μεθόδων διερεύνησης για την απόκτηση γνώσεων και πληροφοριών και για τον εμπλουτισμό των δεδομένων που χρησιμοποιούνται.

2.6.4 Προκλήσεις

Μια από τις σημαντικότερες προκλήσεις που πρέπει να αντιμετωπίσουν όσοι ασχολούνται με τη Διερεύνηση του Σημασιολογικού Ιστού είναι η αναγνώριση των ενδιαφερόντων στοιχείων από τα ημι-δομημένα δεδομένα, διότι πρώτον, οι περισσότεροι αλγόριθμοι είναι σχεδιασμένοι να επεξεργάζονται ομογενή δεδομένα, δεύτερον, η RDF τριπλέτα (υποκείμενο, σχέση, αντικείμενο) προσδίδει περιπλοκή, και τρίτον, οι διαφορετικές πιθανές δομές κλάσεων της Web Ontology Language (OWL) καθιστούν τα δεδομένα περισσότερο ετερογενή.

Ένα ακόμη πρόβλημα προκύπτει όταν εφαρμόζονται κλασσικοί αλγόριθμοι δέντρων απόφασης σε δεδομένα από Οντολογίες, και προσπαθούν να επιλέξουν ορθά τις μεταβλητές. Ο τρόπος με τον οποίο δομούνται οι Οντολογίες μπορούν να οδηγήσουν σε αναρίθμητες ιδιότητες και τιμές, και επομένως είναι αδύνατον να αποδοθούν μεταβλητές σε μη μετρίσιμα στοιχεία.

Τέλος, το πρόβλημα των σπάνιων τιμών μπορεί να οδηγήσει στην αναγνώριση μοτίβων μη σχετικών με το θέμα της αναζήτησης.

Δεν έχουν βρεθεί ακόμη ικανοποιητικές λύσεις για τα ζητήματα αυτά, ωστόσο παρουσιάζονται κάποιες προτάσεις στη βιβλιογραφία [18]. Η βελτιστοποίηση της

επεξεργασίας των δεδομένων και η οργάνωσή τους με βάση τις απαιτήσεις του χρήστη μπορεί να συμβάλλει στην καλύτερη αναγνώριση της χρήσιμης γνώσης. Η εισαγωγή επιπλέον πληροφορίας σχετικά με τις σχέσεις των εννοιών και των ρόλων των αντικειμένων της OWL και η χρήση μεθόδου αυτόματης επιλογής μεταβλητών προτείνεται για την αντιμετώπιση των ζητημάτων που προκύπτουν σχετικά με τις Οντολογίες των σημασιολογικών δεδομένων. Τέλος, η γενίκευση των τιμών του υποκειμένου ή του αντικειμένου της τριπλέτας RDF μπορεί να βοηθήσει στην μείωση της επίπτωσης που έχουν οι σπάνιες τιμές στα αποτελέσματα.

2.7 Knowledge Graphs

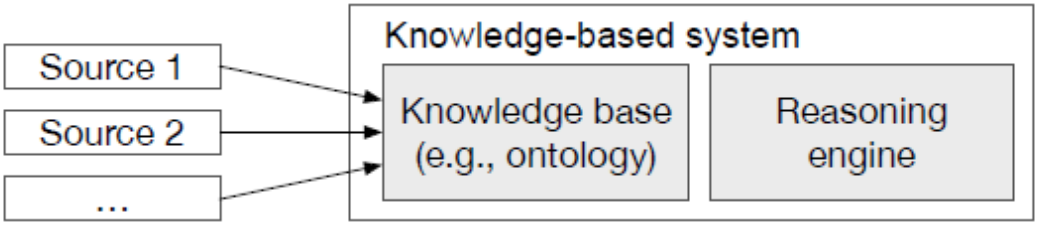
Όπως προηγουμένως αναφέρθηκε, η περιγραφή της πληροφορίας στον Σημασιολογικό Ιστό γίνεται με τη χρήση των τριπλέτων RDF. Κάθε σετ τέτοιων τριπλέτων ονομάζεται γράφος RDF. Αυτή η γραφική αναπαράσταση γνώσης περιλαμβάνει οντότητες, που αποτελούν τους κόμβους του γράφου, οι οποίες συνδέονται με τις σχέσεις, που αποτελούν τις ακμές του γράφου. Κάθε οντότητα χαρακτηρίζεται από ένα URI, και συνήθως η κάθε μία διαθέτει τύπο, ο οποίος δηλώνεται με σχέση «είναι»[20].

Στο [20] δίνονται επίσης και δύο ορισμοί του RDF γράφου ως εξής:

- Ο RDF γράφος, είναι ένας γράφος G=(V,E) όπου V είναι το σύνολο κορυφών και E το σύνολο των κατευθυνόμενων ακμών, κάθε κορυφή ή ακμή έχει μοναδική ετικέτα, όπου κάθε κορυφή του V χαρακτηρίζεται από ένα μοναδικό αναγνωριστικό και κάθε ακμή του E φέρει ετικέτα από ένα πεπερασμένο σύνολο ετικετών ακμής.
- Ο Γράφος Γνώσης, γενικότερα, περιγράφει οντότητες του πραγματικού κόσμου και τις σχέσεις μεταξύ τους, οργανωμένες σε γράφημα. Επιπρόσθετα, καθορίζει πιθανές κατηγορίες και σχέσεις οντοτήτων σε ένα σχήμα, επιτρέπει την πιθανή αλληλεπίδραση αυθαίρετων οντοτήτων μεταξύ τους και τέλος καλύπτει διαφορετικούς τομείς.

Πρέπει να τονιστεί ότι οι Γράφοι Γνώσεων δεν είναι οντολογίες. Η διαφορά τους έγκειται αρχικά στην ποσότητα, ο γράφος είναι μια πολύ μεγάλη οντολογία, και στη

συνέχεια σε επιπλέον απαιτήσεις, ο γράφος απαιτεί την εφαρμογή μηχανισμού λογικής με βάση την οποία αντλείται η νέα γνώση. Αυτή η εφαρμογή του μηχανισμού λογικής είναι και που κάνει τον Γράφο Γνώσης να διαφοροποιείται και από τις βάσεις γνώσεων [24].



Σχ. 9 : Αρχιτεκτονική Γράφου Γνώσης

Η διαφορά των Γράφων Γνώσης με το Σημασιολογικό Ιστό δεν είναι τόσο προφανής. Σε μικρή κλίματα, οι Γράφοι μπορούν εύκολα να αναγνωρισθούν καθώς ασχολούνται με συγκεκριμένο περιεχόμενο. Ωστόσο σε μεγαλύτερη κλίμακα είναι δυσκολότερος ο διαχωρισμός τους. Πολλές φορές ο Σημασιολογικός Ιστός μπορεί να ερμηνευτεί ως ο πιο ολοκληρωμένος Γράφος Γνώσης, ή αντίστοιχα ένας Γράφος Γνώσης που περιλαμβάνει το σύνολο του Διαδικτύου μπορεί να ερμηνευτεί ως ένα αυτοδύναμο Σημασιολογικό Δίκτυο [24].

3. Semantics Related Approaches for Data Mining

Τα τελευταία χρόνια πολλές προσεγγίσεις έχουν προταθεί που να συνδυάζουν δεδομένα του Σημασιολογικού Ιστού με τη διαδικασία Διερεύνησης Δεδομένων και την αναζήτηση γνώσης. Στο κεφάλαιο που ακολουθεί παρουσιάζονται τέσσερις αντιπροσωπευτικές μέθοδοι χρήσης δεδομένων με σημασιολογικές λεπτομέρειες για την άντληση γνώσης από αυτά.

3.1 Using Semantic Web and LOD during the Data mining process

Στην περιεκτική τους έρευνα [15], οι P. Ristoski και H. Paulheim παρέχουν μια επισκόπηση των προσεγγίσεων που συνδυάζουν δεδομένα από το Σημασιολογικό Ιστό με τη Διερεύνηση Δεδομένων και την διαδικασία Ανάκτησης Γνώσης.

Επικεντρώνονται στα παρακάτω κριτήρια:

- Η μέθοδος είναι σχεδιασμένη ώστε να βελτιώνει την διαδικασία Ανάκτησης Γνώσης σε τουλάχιστον ένα βήμα.
- Η μέθοδος χρησιμοποιεί τουλάχιστον ένα σετ δεδομένων από το Σημασιολογικό Ιστό.

Οι συγγραφείς παρουσιάζουν τη βασική ιδέα του πώς να χρησιμοποιήσεις τα δεδομένα αυτά σε κάθε βήμα και αξιολογούν τα ευρήματα της έρευνάς τους.

3.1.1 Επιλογή – Selection

Ο χρήστης πρέπει αρχικά να προσδιορίσει επακριβώς, ποιό είναι το πεδίο των δεδομένων, τι γνώση περιέχουν και ποια πιθανή πρόσθετη πληροφορία θα μπορούσε να εξαχθεί από αυτά. Ως αποτέλεσμα αυτού, ο χρήστης μπορεί να αναγνωρίσει το σκοπό της Διερεύνησης Δεδομένων και να επιλέξει δείγμα δεδομένων που θα είναι πιο κατάλληλα για την επίτευξή του.

Με τη χρήση του Σημασιολογικού Ιστού, και ιδιαίτερα τη χρήση των LOD και των Οντολογιών για το συγκεκριμένο πεδίο, ο χρήστης επιτυγχάνει καλύτερη αναπαράσταση και εξερεύνηση των δεδομένων. Η σύνδεση - *linking* του αρχικού σετ

δεδομένων με το αντίστοιχο του Σημασιολογικού Ιστού αποτελεί το πρώτο βήμα. Μια σύνδεση ή χαρτογράφηση με το LOD σετ υλοποιείται, και αυτό συμβάλει στην αυτόματη εξαγωγή επιπλέον βασικής γνώσης. Επιτρέπει έτσι την αναδιοργάνωση των εννοιών και των πληροφοριών των δεδομένων, καθορίζοντας τύπους και σχέσεις μεταξύ τους. Η χρήση εργαλείων για την καλύτερη απεικόνιση και εξερεύνηση των LOD δεδομένων συνιστάται από τους συγγραφείς [15].

Υπάρχουν τρεις ευρύτερες κατηγορίες που περιγράφουν την σύνδεση των τοπικών δεδομένων με τα LOD, ανάλογα με την δομή του αρχικού σετ:

a. Χρήση LOD για την ερμηνεία σχεσιακών βάσεων δεδομένων.
   Για την αντιμετώπιση των διαφορετικών σχημάτων που συνήθως συναντώνται σε σχεσιακές βάσεις δεδομένων, οι συγγραφείς [15] προτείνουν τη χρήση του εργαλείου D2RQ[25], το οποίο προσφέρει πρόσβαση βασισμένη σε RDF μορφή στο περιεχόμενο των σχεσιακών βάσεων δεδομένων, χωρίς να το αναπαράγει σε RDF μορφή.
b. Χρήση LOD για την ερμηνεία ημι-δομημένων δεδομένων.
   Οι πίνακες απεικόνισης που χαρακτηρίζουν τα ημι-δομημένα δεδομένα είναι χρήσιμοι για την κατανόησή τους από τους ανθρώπους αλλά όχι για την αυτόματη επεξεργασία τους από τις μηχανές. Προτείνεται η χρήση εφαρμογών LOD για να μετατραπεί η πληροφορία που περιέχεται στους πίνακες σε RDF τριπλέτα.
c. Χρήση LOD για την ερμηνεία αδόμητων δεδομένων.
   Η άντληση πληροφορίας από αδόμητα δεδομένα, τα οποία είναι συνήθως σε μορφή κειμένου, μπορεί να βελτιστοποιηθεί με τη σύνδεση των οντοτήτων του κειμένου με αυτές από τη LOD βάση, χρησιμοποιώντας εργαλεία όπως το DBpedia Spotlight.

3.1.2 Προεπεξεργασία – Preprocessing

Μετά τον προσδιορισμό και τη σύνδεση των δεδομένων με τα αντίστοιχα από το Σημασιολογικό ιστό, έπεται η χρήση Οντολογιών και Σημασιολογικών δεδομένων για την βελτίωση της ποιότητάς τους, εκτελώντας έλεγχους εγκυρότητας και καθαρισμό. Οι ακραίες τιμές και ο θόρυβος, οι κενές τιμές και οι τιμές εκτός των περιορισμών εντοπίζονται και μορφοποιούνται.

Υπάρχουν δύο τρόποι εφαρμογής των Οντολογιών για το στάδιο αυτό:

a. Domain-independent

   Η οντολογία περιέχει οδηγίες και προδιαγραφές για την εκτέλεση εργασιών καθαρισμού και ελέγχου. Καθοδηγεί τον χρήστη υποδεικνύοντας πιθανές λειτουργίες που πρέπει να εκτελεστούν στα δεδομένα.

b. Domain-specific

   Η συγκεκριμένη κατηγορία οντολογιών παρέχει ειδικές γνώσεις που απαιτούνται για τον καθαρισμό και την επικύρωση των δεδομένων, συνήθως με αυτόματο τρόπο.

3.1.3 Μεταμόρφωση – Transformation

Σε αυτό το σημείο γίνεται η προετοιμασία για τη δημιουργία καλύτερων δεδομένων για τη Διερεύνησή τους. Η δημιουργία και η επιλογή χαρακτηριστικών των δεδομένων υλοποιούνται σε αυτό το στάδιο, ώστε να διασφαλίζεται τόσο η ποιότητα όσο και η επεκτασιμότητα και το πλήθος των νέων χαρακτηριστικών.

a. Δημιουργία χαρακτηριστικών

   Ο εμπλουτισμός των δεδομένων με χαρακτηριστικά βασισμένα στα LOD βελτιώνει τη διαδικασία Διερεύνησης δεδομένων ενώ εξωτερικεύει το κόστος δημιουργίας και διατήρησης αυτής της γνώσης. Για την πρόσβαση των LOD με τους αλγόριθμους Διερεύνησης, πρέπει να υλοποιηθούν αρκετές μεταμορφώσεις, η οποίες δημιουργούν προτεινόμενα χαρακτηριστικά, ώστε να μπορέσουν οι γράφοι των LOD να συνδεθούν με τα χαρακτηριστικά διανύσματα των αλγόριθμων. Οι συγγραφείς [15] προτείνουν αρκετές εφαρμογές ανάλογα με το είδος των δεδομένων προς επεξεργασία.

b. Επιλογή χαρακτηριστικών

   Η δημιουργία χαρακτηριστικών έχει ως αποτέλεσμα την αύξηση των χαρακτηριστικών των δεδομένων σε τέτοιο βαθμό, όπου να απαιτείται επιλεκτική επιλογή τους. Ο στόχος είναι να προσδιοριστούν τα χαρακτηριστικά εκείνα τα οποία σχετίζονται περισσότερο με το σκοπό της Διερεύνησης. Οι συγγραφείς [15] προτείνουν μια προσέγγιση η οποία εκμεταλλεύεται την ιεραρχία για την επιλογή χαρακτηριστικών σε συνδυασμό με πρότυπα κριτήρια όπως το κέρδος ή η συσχέτιση.

Υπάρχουν μέθοδοι με ή χωρίς επίβλεψη, όπου στην πρώτη περίπτωση, ο χρήστης πρέπει να καθορίσει ένα ερώτημα για τη δημιουργία χαρακτηριστικών, όταν γνωρίζει ποια θα μπορούσαν να είναι πολύτιμα. Οι κλασσικές μέθοδοι δημιουργίας έχουν ως αποτέλεσμα χαρακτηριστικά που μπορούν να ερμηνευθούν εύκολα από τον άνθρωπο, αλλά είναι και εφαρμόσιμα για περιγραφική Διερεύνηση. Οι μέθοδοι Kernel συχνά παρέχουν καλύτερες μεθόδους πρόβλεψης, αλλά με το κόστος της απώλειας της ερμηνείας των αποτελεσμάτων τους.

Ο αριθμός και η επεκτασιμότητα των χαρακτηριστικών που δημιουργούνται από τα Linked Open Data είναι ένα κρίσιμο πρόβλημα και συνίσταται η επιλογή κατάλληλου υποσυνόλου βασιζόμενο σε σχηματικές πληροφορίες, από τα ίδια τα LOD. Αυτοί οι αλγόριθμοι παρέχουν καλύτερο συνδυασμό μεταξύ της μείωσης του όγκου και τις προγνωστικής απόδοσής τους σε σχέση με προσεγγίσεις που δεν βασίζονται σε σχηματικές πληροφορίες.

3.1.4 Διερεύνηση Δεδομένων – Data mining

Πολύ συχνά χρησιμοποιούνται οντολογίες που υποστηρίζουν τον χρήστη κατά τη διαδικασία Διερεύνησης και διασφαλίζουν ότι ο επιλεγμένος αλγόριθμος είναι ικανός να χειριστεί τα συγκεκριμένα δεδομένα.

Η πλατφόρμα RapidMiner[5], χρησιμοποιεί εσωτερικά σημασιολογικές περιγραφές των operators για να βοηθήσει τον χρήστη στην αποφυγή λαθών, όπως όταν συνδυάζει την προεπεξεργασία δεδομένων με τους operators μηχανικής μάθησης. Εδώ η λογική δεν ελέγχει μόνο την εγκυρότητα μιας διαδικασίας αλλά είναι υπεύθυνη και για τη διόρθωση μιας μη έγκυρης.

Προσεγγίσεις που χρησιμοποιούν σημασιολογικές πληροφορίες απευθείας σε ένα αλγόριθμό Διερεύνησης για να επηρεάσουν τα αποτελέσματά του είναι πολύ σπάνιες. Υπάρχουν κάποιες προτάσεις που χρησιμοποιούν ως υπόβαθρο σημασιολογικές πληροφορίες, κυρίως για την αναγνώριση μοτίβων που είναι ευκολότερο να κατανοηθούν από τους χρήστες.

3.1.5 Ερμηνεία – Interpretation

Τα μοτίβα που ανακαλύπτονται στο προηγούμενο βήμα, πρέπει να ερμηνευτούν και να γίνουν κατανοητά, συνήθως με τη χρήση κάποιας βασικής πληροφορίας, η οποία όμως δεν είναι εύκολο να προσδιοριστεί.

Τα δεδομένα του Σημασιολογικού Ιστού μπορούν να συμβάλουν σε αυτό το βήμα, και ιδιαίτερα για περιγραφικά στοιχεία. Αυτά συνήθως περιλαμβάνουν υποομάδες ή συμπλέγματα, ή μοντέλα κανόνων τα οποία χρησιμοποιούνται για την περιγραφή ενός σετ δεδομένων.

Η πληροφορία που χρησιμοποιείται από τα δεδομένα LOD και τις οντολογίες μπορεί να βοηθήσει περαιτέρω στην ανάλυση των ευρημάτων, επεξηγώντας π.χ. τα τυπικά χαρακτηριστικά μιας υποομάδας, και επομένων αιτιολογώντας την ομαδοποίηση που επέλεγε ο αλγόριθμος Διερεύνησης.

Τα σύνολα δεδομένων που χρησιμοποιούνται εδώ είναι μικτά, και σύνολα γενικού σκοπού, όπως η DBpedia [21], συχνά προτιμώνται. Ωστόσο, μπορούν να αξιοποιηθούν και σύνολα τα οποία είναι ιδιαίτερα εξειδικευμένα, καθώς και οι διασυνδέσεις μεταξύ τους.

3.1.6 Recommender System case study

Τέλος, οι συγγραφείς [15] αποφάσισαν να εφαρμόσουν την ανωτέρω διαδικασία στον τομέα των συστημάτων συστάσεων. Τα συστήματα συστάσεων αποτελούν μια σημαντική μέθοδο φιλτραρίσματος μεγάλο όγκου πληροφορίας και προϊόντων για τους χρήστες.

Το αρχικό σύνολο δεδομένων περιλαμβάνει βιβλία και τις κριτικές των χρηστών. Η σύνδεση των δεδομένων υλοποιήθηκε μέσω της DBpedia, όπου η μάρκα και το έτος κυκλοφορίας χρησιμοποιήθηκαν ως παράμετροι για την αντιστοίχηση με την οντότητα της DBpedia. Το αποτέλεσμα είναι η δημιουργία συνόλου με τις οντότητες της DBpedia αλλά και τις κριτικές των χρηστών. Επιπρόσθετες πληροφορίες αντλούνται από τους συνδέσμους *owl:sameAs*, όπου αναζητούνται στοιχεία για τα βιβλία από άλλες LOD βάσεις δεδομένων, και έπειτα ενοποιούνται σε ένα καθαρό σύνολο δεδομένων. Για την δημιουργία χαρακτηριστικών χρησιμοποιήθηκε μία μίξη από αυτόματη και χειροκίνητη προσέγγιση, ώστε να περιορίσουν τον αριθμό και να

αυξήσουν την σχετικότητα τους. Τέλος, για την αξιολόγηση των αποτελεσμάτων επικεντρώθηκαν στην ικανότητα παραγωγής σχετικών αποτελεσμάτων αλλά και στην ικανότητα να επεξηγούν αποτελεσματικά τα αποτελέσματα αυτά.

Τα συμπεράσματα στα οποία καταλήγουν οι συγγραφείς [15], με βάση την έρευνα αλλά και την μελέτη περίπτωσης που εκπόνησαν. είναι τα ακόλουθα:

- Η DBpedia χρησιμοποιείται σχεδόν σε κάθε προσέγγιση, γεγονός που τονίζει την χρησιμότητα μιας γενικής χρήσης πηγή γνώσης αλλά ταυτόχρονα περιορίζει την εμπειρία σχετικά με τη γενική εφαρμογή των προσεγγίσεων σε άλλες πηγές LOD.
- Συχνά χρησιμοποιούνται προσαρμοσμένες οντολογίες και σύνολα δεδομένων αντί για επαναχρησιμοποίηση ανοιχτών συνόλων από το Διαδίκτυο.
- Οι σύνδεσμοι μεταξύ των συνόλων δεδομένων δεν έχουν αξιοποιηθεί στο έπακρον και υπάρχει μεγάλο περιθώριο βελτίωσης στον συγκεκριμένο τομέα.
- Τα ρητά σχήματα, οι οντολογίες και η συλλογιστική σε αυτά σπάνια συνδυάζονται με τη Διερεύνηση Δεδομένων και την ανακάλυψη γνώσης.
- Η γνώση από το Σημασιολογικό Ιστό αφορά κυρίως τα δεδομένα υπό επεξεργασία και σπάνια τον τομέα Διερεύνησης.

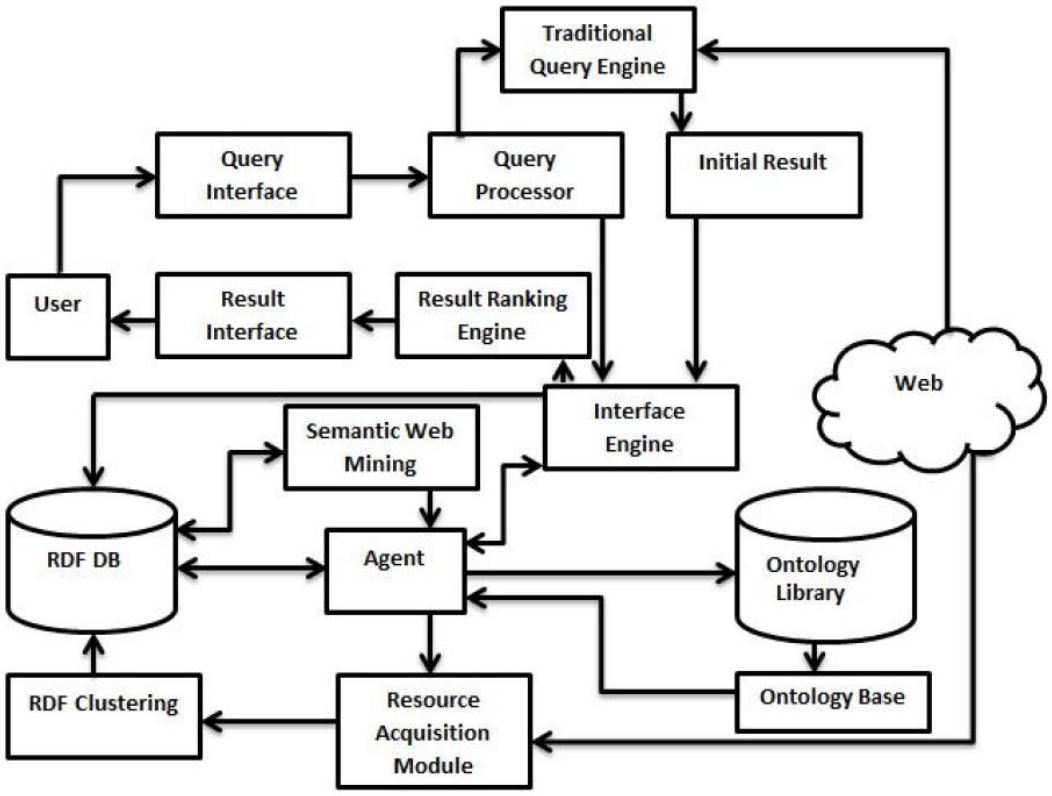
Υπάρχουν αρκετά τμήματα τα οποία χρήζουν περαιτέρω έρευνας. Οι συνολικές δυνατότητες της χρήσης του Σημασιολογικού Ιστού στη Διερεύνηση Δεδομένων ακόμη δεν έχουν αποκαλυφθεί.

3.2 Combination of Semantic Web with Traditional Search Systems

Οι S. Kabir, S. Ripon, M. Rahman και T.Rahman [17], παρουσιάζουν ένα ευφυές μοντέλο Διερεύνησης Ιστού, όπου το ερώτημα των χρηστών αναζητείται ακολουθώντας τον παραδοσιακό τρόπο, δηλαδή χρησιμοποιώντας τις δημοφιλείς μηχανές αναζήτησης όπως το Google. Ο ευφυής πράκτορας ελέγχει τα δεδομένα που αναζητούνται και εξάγει μόνο αυτά που σχετίζονται σημασιολογικά με τις παραμέτρους αναζήτησης των χρηστών.

3.2.1 Μοντέλο Διερεύνησης

Το μοντέλο που προτείνουν φαίνεται σχηματικά στην ακόλουθη εικόνα [Σχ. 10]



Σχ. 11: Προτεινόμενο μοντέλο Διερεύνησης δεδομένων με χρήση σημασιολογικού πράκτορα[17]

Το πρώτο βήμα του μοντέλου είναι η προώθηση του αιτήματος του χρήστη στον επεξεργαστή ερωτημάτων (Query Processor) μέσω της αντίστοιχης διεπαφής.

Ακολουθεί η παράλληλη κλήση τόσο της παραδοσιακής μηχανής αναζήτησης όσο και του ευφυούς πράκτορα μέσω της μηχανής διεπαφής, με το αίτημα του χρήστη ως παράμετρο. Δίνεται η δυνατότητα στο χρήστη να διακόψει τη διαδικασία διερεύνησης ακαριαία, εάν το επιθυμεί. Η μηχανή ερωτήματος είναι μια υπηρεσία που λαμβάνει μια περιγραφή ενός αιτήματος αναζήτησης, το αξιολογεί και το εκτελεί κα επιστρέφει τα αποτελέσματα στο χρήστη. Η υπηρεσία αυτή δρα ως ένα ενδιάμεσο επίπεδο μεταξύ χρηστών και πηγών δεδομένων, ερμηνεύοντας τα αιτήματα χωρίς να καταιγίζονται από λεπτομέρειες σχετικά με τον τρόπο πρόσβασης στις πηγές δεδομένων. Σε αυτό το στάδιο οι παραδοσιακές μηχανές αναζήτησης επιστρέφουν τα

αποτελέσματά τους στη μηχανή διεπαφής και αυτή με τη σειρά της τα προωθεί στη βάση δεδομένων RDF.

Στη συνέχεια, για την αναζήτηση που βασίζεται στον ευφυή πράκτορα, είναι απαραίτητο να δημιουργηθεί μια οντολογία και οι διάφορες έννοιές της, με βάση διάφορα στοιχεία που θα συλλεχθούν από το Διαδίκτυο από κάποιο αλγόριθμο ομαδοποίησης. Αυτή θα αποθηκευθεί στη βιβλιοθήκη οντολογιών για μελλοντική χρήση.

Στο επόμενο στάδιο, εάν η βάση δεδομένων RDF διαθέτει τα επιθυμητά αποτελέσματα, τα αποστέλλει στο χρήστη μέσα της διεπαφής. Σε αντίθετη περίπτωση, εκτελείται αναζήτηση όλων των πιθανών σχέσεων μεταξύ των παραμέτρων που αιτήθηκε ο χρήστης και των οντοτήτων που υπάρχουν στη βιβλιοθήκη οντολογιών, και ο πράκτορας δημιουργεί μια βάση οντολογίας με τις σχέσεις για την οντολογία που δημιουργήθηκε.

Η βάση οντολογίας διαθέτει όλους τους πιθανούς κόμβους που αφορούν το αίτημα και έχουν συλλεχθεί από τον πράκτορα και από τη γνώση που υπάρχει στη βάση.

Η μονάδα Απόκτησης πόρων (Resource Acquisition Module), συλλέγει πληροφορίες σχετικά με το αίτημα από το Διαδίκτυο. Οι κόμβοι των πόρων με τα πιο σχετικά χαρακτηριστικά αποθηκεύονται στη βάση δεδομένων RDF.

Η μονάδα Διερεύνησης Σημασιολογικού Ιστού επεξεργάζεται, στη συνέχεια, τα δεδομένα της βάσης RDF για βελτιστοποίηση των αποτελεσμάτων, τα οποία και αποστέλλονται στον πράκτορα, ο οποίος για να ενισχύσει τη συνάφεια των αποτελεσμάτων εκτελεί διάφορες διαδικασίες φιλτραρίσματος.

Τέλος, τα αποτελέσματα ταξινομούνται και εμφανίζονται στο χρήστη, και παρουσιάζουν όλες τις δυνατές πτυχές από τις οποίες ανακτάται η επιθυμητή γνώση.

Το μοντέλο που αναπτύχθηκε από τους συγγραφείς[17], χρησιμοποιεί έναν ευφυή πράκτορα για την διερεύνηση μεγάλου όγκου δεδομένων. Σε περίπτωση χρήσης πολλαπλών πρακτόρων, ιδιαίτερη προσοχή πρέπει να δοθεί στο συντονισμό και στην επικοινωνία μεταξύ τους.

3.2.2 Ο ρόλος των Οντολογιών στην αναζήτηση

Οι οντολογίες παίζουν πολύ σημαντικό ρόλο στη συγκεκριμένη μέθοδο, μιας και περιλαμβάνουν όλες τις εννοιολογικές γνώσεις για τα αντικείμενα του πεδίου αναζήτησης, και τις αποθηκεύουν στη βιβλιοθήκη οντολογιών.

Όταν ο χρήστης εκτελεί ένα αίτημα, τότε ο πράκτορας αναζητά τη βιβλιοθήκη για όλους τους πιθανούς κόμβους σχέσεων που είναι σχετικοί με τις παραμέτρους της αναζήτησης. Αυτή η προσπέλαση της βιβλιοθήκης είναι δυνατή διότι όλα τα σύνολα δεδομένων είναι διασυνδεδεμένα και σαφώς ορισμένα στο Σημασιολογικό ιστό.

Επομένως εμφανίζεται ένα μεγαλύτερο εύρος αποτελεσμάτων και ο χρήστης έχει τη δυνατότητα επιλογής ανάλογα με τις απαιτήσεις του, συγκριτικά με τη παραδοσιακή αναζήτηση.

Η χρήση οντολογιών δίνει καλύτερα αποτελέσματα όταν η αναζητούμενη γνώση δεν είναι πλήρως κατανοητή.

Η επιτυχής ανακάλυψη της διαδρομής προορισμού από τον πράκτορα σύμφωνα με το αίτημα του χρήστη, μέσω των οντολογιών, παρέχει διάφορες δυνατότητες στην διαδικασίας αναζήτησης, όπως αυτοματοποίηση, τεχνητή νοημοσύνη, ολοκλήρωση, ικανότητα επικοινωνίας μηχανής με μηχανή κλπ..

Η σημασία της χρήσης των Οντολογιών για τη βελτιστοποίηση της Διερεύνησης Δεδομένων φαίνεται επίσης και στην ακόλουθη έρευνα.

3.3 The Ontology Approach

Οι K. Sridevi, R. Umarani [19], επικεντρώνουν την έρευνα τους στη χρήση των Οντολογιών κατά τη Διερεύνηση Δεδομένων από το διαδίκτυο.

Οι Οντολογίες μπορούν να χρησιμοποιηθούν ως βασικές σημασιολογικές δομές για την ανάκτηση γνώσης από το διαδίκτυο. Η Διερεύνηση Περιεχομένου στο διαδίκτυο μπορεί να ενισχυθεί, επομένως, ως διαδικασία χρησιμοποιώντας πιο εκφραστικά ερωτήματα αναζήτησης, που περιλαμβάνουν όρους αναζήτησης σε έννοιες ή/και σχέσεις.

Τα παραδοσιακά θέματα που καλύπτονται από τη Διερεύνηση Περιεχομένου μετασχηματίζονται με τις Οντολογίες ως εξής:

- ✓ Ontology based Web page classification – Κατάταξη Ιστοσελίδας βάσει Οντολογιών

  Οι ιστοσελίδες ταξινομούνται ως υποδείγματα εννοιών και «ζεύγη» ιστοσελίδων ως παραδείγματα σχέσεων.
- ✓ Ontology-based Web clustering – Ομαδοποίηση Ιστού βάσει Οντολογιών

  Χρησιμοποιούνται HTML στοιχεία που αντιστοιχούν σε έννοιες ως χαρακτηριστικά για την εξαγωγή πιο ακριβών αποτελεσμάτων.
- ✓ Ontology-based Web extraction – Εξαγωγή από το Διαδίκτυο βάσει Οντολογιών

  Εξάγονται τόσο στοιχεία HTML ως υποδείγματα εννοιών όσο και τα σχετικά ζεύγη και οι σχέσεις αυτών.
- ✓ Ontology-based Web site structure mining – Διερεύνηση δομής Ιστοσελίδας βάσει Οντολογιών

  Προσδιορίζεται το μοτίβο σύνδεσης εννοιών από τις ιστοσελίδες για τη βελτίωση του σχεδιασμού τους.

Εκτός από τη δημιουργία των Οντολογιών, οι ακόλουθες λειτουργίες μπορούν να υλοποιηθούν πάνω σε αυτές:

- ❖ Συγχώνευση: Η δημιουργία μιας νέας Οντολογίας από διασυνδέοντας τις υπάρχουσες. Συνήθως η νέα περιλαμβάνει όλες τις πληροφορίες των αρχικών, ωστόσο σε περίπτωση που οι συγχωνευόμενες δεν είναι πλήρως σύμφωνες, εισάγει επιλεγμένη γνώση από τις αρχικές, ώστε το αποτέλεσμα να είναι συνεπές. Η νέα οντολογία μπορεί να εισάγει νέες έννοιες και σχέσεις που θα λειτουργούν ως γέφυρα μεταξύ των όρων των αρχικών.
- ❖ Χαρτογράφηση: Είναι η έκφραση του τρόπου με τον οποίο μεταφράζονται οι δηλώσεις από τη μια οντολογία στην άλλη. Συχνά σημαίνει την μετάφραση μεταξύ εννοιών και σχέσεων.
- ❖ Ευθυγράμμιση: Είναι η χαρτογράφηση ανάμεσα σε οντολογίες προς και τις δύο κατευθύνσεις, κατά την οποία μπορούμε να μεταβάλουμε τις αρχικές οντολογίες, εισάγοντας έννοιες και σχέσεις, ώστε να είναι δυνατή μια τέτοια μετάφραση.

- ❖ Εκκαθάριση: Είναι η χαρτογράφηση όπου όλες οι έννοιες μιας οντολογίας έχουν αντιστοιχία στις έννοιες της δεύτερης.
- ❖ Ενοποίηση: Ευθυγραμμίζει τις έννοιες και τις σχέσεις των οντολογιών ώστε το συμπέρασμα σε μια μπορεί να χαρτογραφηθεί σε συμπέρασμα στην άλλη οντολογία.,
- ❖ Ενσωμάτωση: Είναι η διαδικασία κατά την οποία αναζητούνται τα ίδια χαρακτηριστικά σε δύο οντολογίες καθώς αναπτύσσεται μια νέα που θα επιτρέπει την μετάφραση μεταξύ των αρχικών και την διαλειτουργικότητα δύο συστημάτων που βασίζονται αντίστοιχα στις αρχικές οντολογίες. Ανάλογα με τον όγκο των αλλαγών μεταξύ των δύο οντολογιών, το επίπεδο της ενσωμάτωσης μπορεί να κυμαίνεται από ευθυγράμμιση έως και ενοποίηση.
- ❖ Κληρονομικότητα: Είναι η λειτουργία κατά την οποία μια οντολογία λαμβάνει όλες τις έννοιες, τις σχέσεις τους περιορισμούς αλλά και τα αξιώματα μιας άλλης, χωρίς να δημιουργείται ασυνέπεια από την επιπλέον γνώση της μιας.

Οι Οντολογίες προσφέρουν έναν αποτελεσματικό τρόπο μείωσης του όγκου της υπερφόρτωσης πληροφοριών, κωδικοποιώντας τη δομή ενός συγκεκριμένου τομέα και προσφέροντας ευκολότερη πρόσβαση στις πληροφορίες για τους χρήστες. Η έρευνα του Σημασιολογικού Ιστού επεκτείνεται για να βελτιώσει τη μοντελοποίηση Οντολογίας, τις μεθοδολογίες και πρακτικές επαναχρησιμοποίησης, την εξαγωγή Οντολογίας και τη σύγκριση, χαρτογράφηση, αξιολόγηση συγχώνευσης και μέτρηση αξιοπιστίας αυτών.

3.3.1 Εφαρμογές

Η εφαρμογή της Διερεύνησης Ιστού βάσει Οντολογιών μπορεί να χωρισθεί σε δύο κατηγορίες:

1) Βελτιωμένη αναζήτηση διαδικτυακών δεδομένων

   Με την εισαγωγή και χρήση Οντολογιών, τα δεδομένα μπορούν να αναπροσαρμόζονται βάσει των εννοιών και των σχέσεών τους, και να υποστηρίζουν πιο εκφραστικά ερωτήματα αναζήτησης, που με τη σειρά τους στοχεύουν σε πιο ακριβή/στοχευμένη αναζήτηση πληροφορίας, μειώνοντας τον όγκο των μη σχετικών αποτελεσμάτων.

2) Καλύτερες δυνατότητες Περιήγησης

   Οι Ιστοσελίδες αντίστοιχα μπορούν να αναζητηθούν με βάση τις έννοιες και τις σχέσεις που περιέχουν.

3.3.2 Γλώσσες και Εργαλεία Οντολογιών

Οι συγγραφείς [19] προτείνουν γλώσσες Οντολογιών που χρησιμοποιούνται για την κατασκευή τους, επιτρέπουν την κωδικοποίηση των γνώσεων σχετικά με τους τομείς στους οποίους αναφέρεται η οντολογία, και συχνά περιλαμβάνουν κανόνες συλλογισμού που υποστηρίζουν την επεξεργασία της περιεχόμενης γνώσης. Χρησιμοποιούν γλώσσα σήμανσης για την κωδικοποίησή τους, η οποία συνηθέστερα είναι η XML.

Γλώσσες για την δημιουργία οντολογιών είναι οι: DAML+OIL, Ontology Inference Layer, RDF, RDF Schema και η OWL.

Οι συγγραφείς [19] παρουσιάζουν επίσης και κάποια εργαλεία για την ανάπτυξη Οντολογιών, τα οποία βοηθούν στην μορφοποίηση τους, όπως OntoEdit, OilEd, SWOOP και Protégé.

3.4 Personalized concept-based search

Σε αυτή την έρευνα οι συγγραφείς, M. Sah και V. Wade [26], παρουσιάζουν έναν καινοτόμο, εξατομικευμένο μηχανισμό αναζήτησης, που επικεντρώνεται στις έννοιες του Σημασιολογικού Ιστού, και βασίζεται στην κατηγοριοποίηση των αποτελεσμάτων.

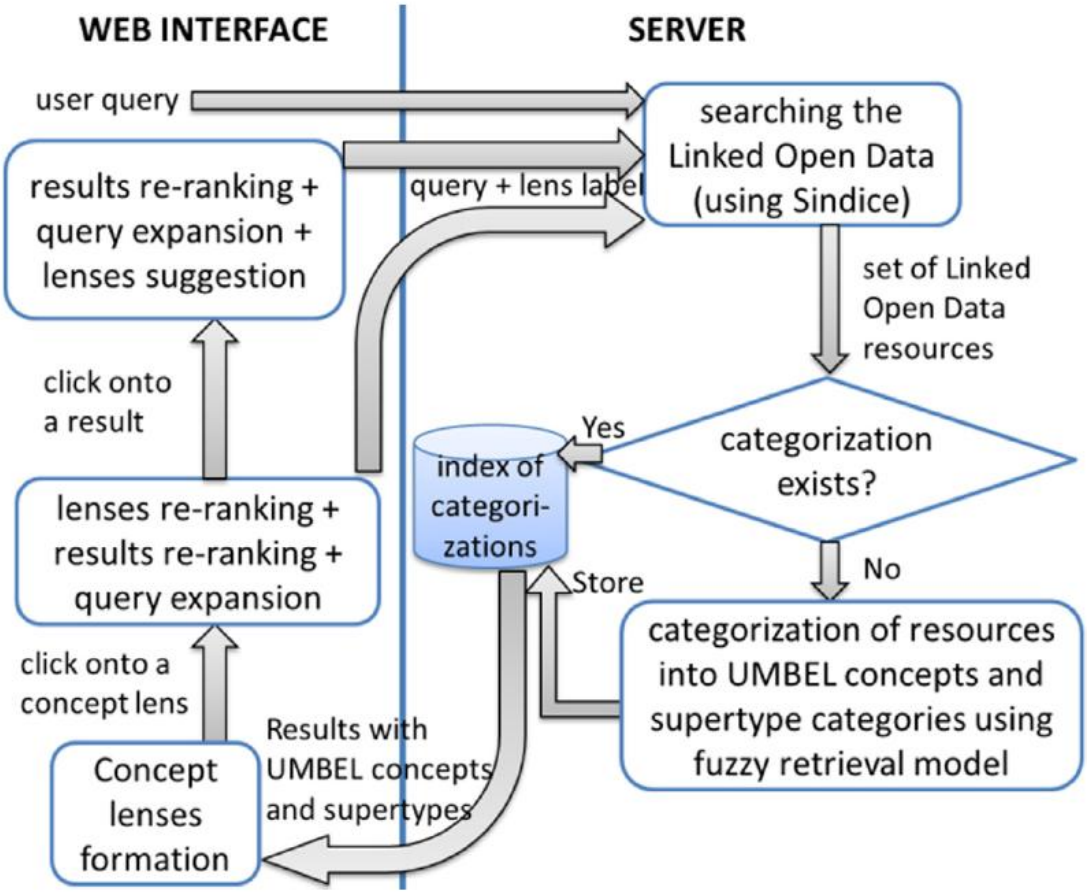
Η ιδιαιτερότητα του μηχανισμού που παρουσιάζεται έγκειται στα ακόλουθα σημεία, όπως τα υποδεικνύουν οι ίδιοι οι συγγραφείς [26]:

- Η ιδέα του εξατομικευμένου, και που βασίζεται στις έννοιες, μηχανισμού αναζήτησης για την διερεύνηση του Σημασιολογικού Ιστού είναι καινοτόμα και δεν έχει διερευνηθεί πρωτύτερα.
- Εισάγουν ένα νέο και ισχυρό αλγόριθμο κατηγοριοποίησης, χρησιμοποιώντας ένα νέο σύστημα ανάκτησης δεδομένων.

- Τα αποτελεσμάτων-εννοιών κατηγοριοποιούνται και, ανάλογα με τη θέση στην κατηγοριοποίηση και τη συνάφεια τους, ομαδοποιούνται και παρουσιάζονται μαζί σε ομάδες, οι οποίες ονομάζονται *concept lenses*.
- Προτείνεται η χρήση της κατηγοριοποίησης των αποτελεσμάτων ως ένα εργαλείο εξατομίκευσης των concept lenses και αναδιοργάνωσης αυτών και των αποτελεσμάτων, επέκτασης της αναζήτησης καθώς και πρότασης νέων lenses. Χρησιμοποιούν το λεξιλόγιο UMBEL[29], για την κατηγοριοποίηση, την παρουσίαση των εννοιών και την εξατομίκευση των αποτελεσμάτων.

3.4.1 Αρχιτεκτονική και Περιγραφή Συστήματος

Η αρχιτεκτονική του συστήματος παρουσιάζεται γραφικά στο ακόλουθο σχήμα [Σχ. 11].



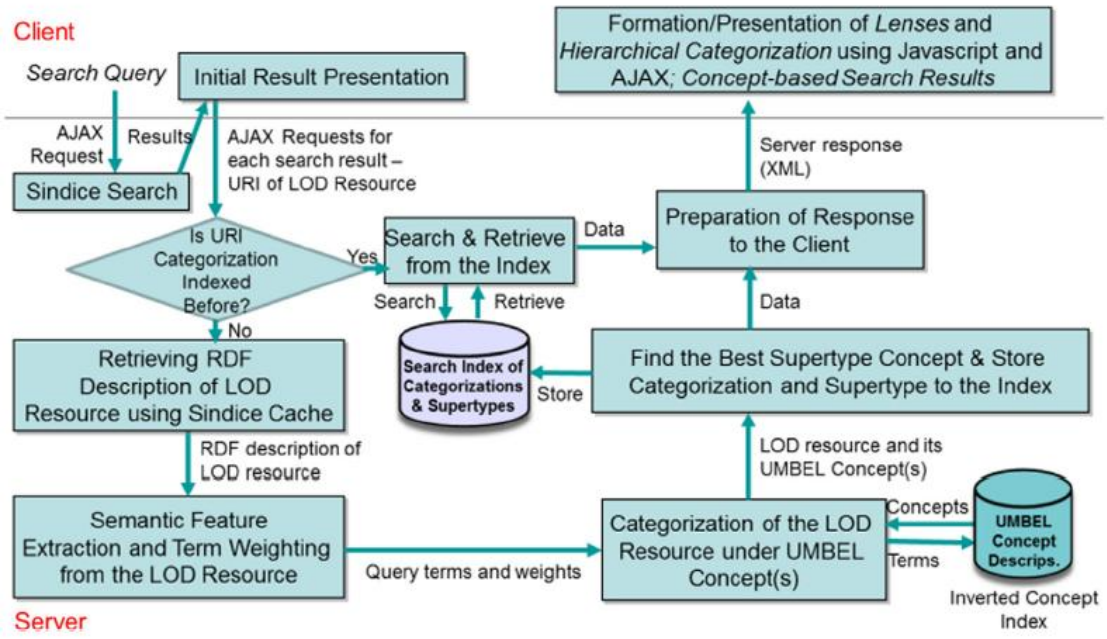
Σχ. 11: Διάγραμμα ροής του συστήματος αναζήτησης [26].

Ο μηχανισμός διαθέτει δύο κύρια μέρη: 1) χρησιμοποιώντας το αίτημα του χρήστη, το σύστημα αναζητά τα Διασυνδεδεμένα Ανοιχτά Δεδομένα και το Σημασιολογικό

Ιστό, και κατηγοριοποιεί τα ανακτημένα αποτελέσματα αναζήτησης στην πλευρά του Διακομιστή, και 2) ανάλογα με την αλληλεπίδραση του χρήστη, στην πλευρά του χρήστη-client τα αποτελέσματα εξατομικεύονται ανά χρήστη.

Οι χρήστες έχουν τη δυνατότητα να εισάγουν αιτήματα είτε με λέξεις κλειδιά είτε με τη χρήση URIs, εφόσον το γνωρίζουν. Με την είσοδο του αιτήματος, ο μηχανισμός αναζητά στο Σημασιολογικό Ιστό με τη χρήση ενός API αναζήτησης των LOD, στη συγκεκριμένη περίπτωση του Sindice API [30], και τα πρώτα αποτελέσματα της αναζήτησης παρουσιάζονται στο χρήστη χωρίς καμία κατηγοριοποίηση.

Στη συνέχει ακολουθεί η κατηγοριοποίηση των εννοιών που έχουν ανακτηθεί από την αναζήτηση, το διάγραμμα ροής των οποίων είναι το ακόλουθο:



Σχ. 12: Κατηγοριοποίηση εννοιών [26].

Για κάθε έννοια, και το αντίστοιχο URI, που ανακτάται, αρχικά λαμβάνεται η RDF περιγραφή της και με βάσει τα στοιχεία που περιέχει, αναζητούνται και διερευνώνται νέα σημασιολογικά δεδομένα από τα LOD. Ανάλογα με τη σπουδαιότητά τους προσδίδεται μια μεταβλητή βάρους σε αυτά. Στη συνέχεια γίνεται η αντιστοίχιση των νέων στοιχείων με τις έννοιες του UMBEL Λεξιλογίου, του οποίου οι έννοιες αναπαριστώνται χρησιμοποιώντας ένα σημασιολογικό μοντέλο κατάταξης. Τέλος, οι λαμβανόμενοι όροι από τα LOD και οι UMBEL έννοιες που έχουν προσδιοριστεί, χαρτογραφούνται με βάση ένα μοντέλο ανάκτησης, ταξινομούνται και αποθηκεύονται

στο ευρετήριο για μελλοντική χρήση, και αποστέλλονται πίσω στο χρήστη. Τα αποτελέσματα με τις ίδιες έννοιες ομαδοποιούνται σε concept lenses. Η πιο σημαντική κατηγοριοποίηση χρησιμοποιείται για τη δημιουργία των concept lenses, και οι υπόλοιπες χρησιμοποιούνται για την εξατομίκευση και συγκρίσεις σημασιολογικής ομοιότητας. Οι χρήστες μπορούν να εξερευνήσουν τα αποτελέσματα από διαφορετικές εννοιολογικές οπτικές γωνίες.

Μόλις ο χρήστης επιλέξει ένα concept lens για εξερεύνηση, τα αποτελέσματα αμέσως εξατομικεύονται και αναπροσαρμόζονται με βάση την επιλογή αυτή και με βάση μια συνδυασμένη σημασιολογική και συντακτική ομοιότητα με το επιλεγμένο concept lens. Ταυτόχρονα, στον ίδιο τον επιλεγμένο concept lens, περιλαμβάνονται τα πιο συναφή αποτελέσματα, με βάση την αναδιάταξή τους λόγω της επιλογής και την επέκταση του αιτήματος από τα σημασιολογικά στοιχεία του. Τέλος, μαζί με τον επιλεγμένο concept lens, συνιστώνται σχετικοί concept lenses προς εξερεύνηση.

Η αλληλεπίδραση του χρήστη με τα αποτελέσματα οδηγεί σε διαδραστική εξατομίκευση. Κάθε νέο “click” οδηγεί ανακατανομή των αποτελεσμάτων, επέκταση του αιτήματος με νέα δεδομένα και περαιτέρω προτάσεις concept lenses. Η εξατομίκευση αυτή είναι χρήσιμη σε αναζήτηση σύνθετης γνώσης, όπως είναι η ανάκτηση πληροφοριών σε άγνωστους τομείς όπου απαιτείται καλή ταξινόμηση των αποτελεσμάτων και βηματική προσπέλασή τους. Η κάθε κατηγοριοποίηση και εξατομίκευση διαρκεί μέχρι να ολοκληρωθεί η αναζήτηση γνώσης και να κλείσει η συνεδρία από το χρήστη.

Στην έρευνά τους [26], οι συγγραφείς παρουσιάζουν εκτεταμένη επιχειρηματολογία με τους λόγους που χρησιμοποίησαν το κάθε εργαλείο, π.χ. Λεξιλόγιο UMBEL, Sindice API, κλπ., και παρουσιάζουν και τα μαθηματικά μοντέλα πίσω από την κατηγοριοποίηση των αποτελεσμάτων με το μοντέλο ανάκτησής τους και την εξατομίκευση των αποτελεσμάτων ανάλογα με τις επιλογές των χρηστών.

3.4.2 Αποτελέσματα

Για την αξιολόγηση των αποτελεσμάτων και της απόδοσης του μηχανισμού αναζήτησης οι συγγραφείς [26] εκπόνησαν εκτεταμένα πειράματα.

Η αξιολόγηση της απόδοσης της κατηγοριοποίησης έγινε με βάση τα ακόλουθα κριτήρια: 1) την απόδοση κατηγοριοποίησης διαφορετικών χαρακτηριστικών των LOD, 2) την ακρίβεια κατηγοριοποίησης του μοντέλου ανάκτησης σε σχέση με το μοντέλο Vector Space και 3) την αποδοτικότητα κατηγοριοποίησης της επίδοσης του συστήματος.

Η αξιολόγηση της απόδοσης του συστήματος και της ικανοποίησης των χρηστών έγινε με βάση μελέτες χρηστών όπου συγκρίθηκε το προτεινόμενο μοντέλο έναντι ενός μη προσαρμοστικού βασικού συστήματος αναζήτησης.

Τέλος, μελετήθηκε η πρακτική εφαρμογή της μεθόδου καθώς και η σύγκρισή της με ήδη υπάρχουσες μηχανές αναζήτησης των LOD.

Τα αποτελέσματα των παραπάνω αξιολογήσεων είναι πολύ θετικά. Τα αποτελέσματα έδειξαν ότι ο μηχανισμός διαθέτει υψηλά επίπεδα ακρίβειας στην κατηγοριοποίηση (~90%) και αποδίδει πολύ καλύτερα από το μοντέλο Vector Space. Οι χρήστες προτίμησαν το προτεινόμενο μοντέλο από τα κλασσικά συστήματα αναζήτησης τα οποία δεν προσαρμόζονται στις επιλογές τους. Η αναζήτηση παρουσίασε συνέπεια στη ταχύτητα και ανακτήθηκαν οι επιθυμητές πληροφορίες σε λιγότερα στάδια και αιτήματα. Τέλος, οι χρήστες έκριναν ότι η χρησιμότητα του εξατομικευμένου μηχανισμού αναζήτησης είναι μεγαλύτερη από τις υπάρχουσες μηχανές αναζήτησης, γεγονός που ενθαρρύνει τους συγγραφείς για μελλοντική ανάπτυξη του μοντέλου τους.

4. Knowledge Graph related Approaches for Data Mining

Το κεφάλαιο που ακολουθεί επικεντρώνεται στη χρήση των Knowledge Graphs, ή αλλιώς Γράφων Γνώσης, για την αναπαράσταση των εννοιών των Σημασιολογικών Ιστών, εισάγοντας τη σχέση μεταξύ των εννοιών στη διαδικασία Διερεύνησης δεδομένων και προσδίδοντας της μια επιπλέον διάσταση.

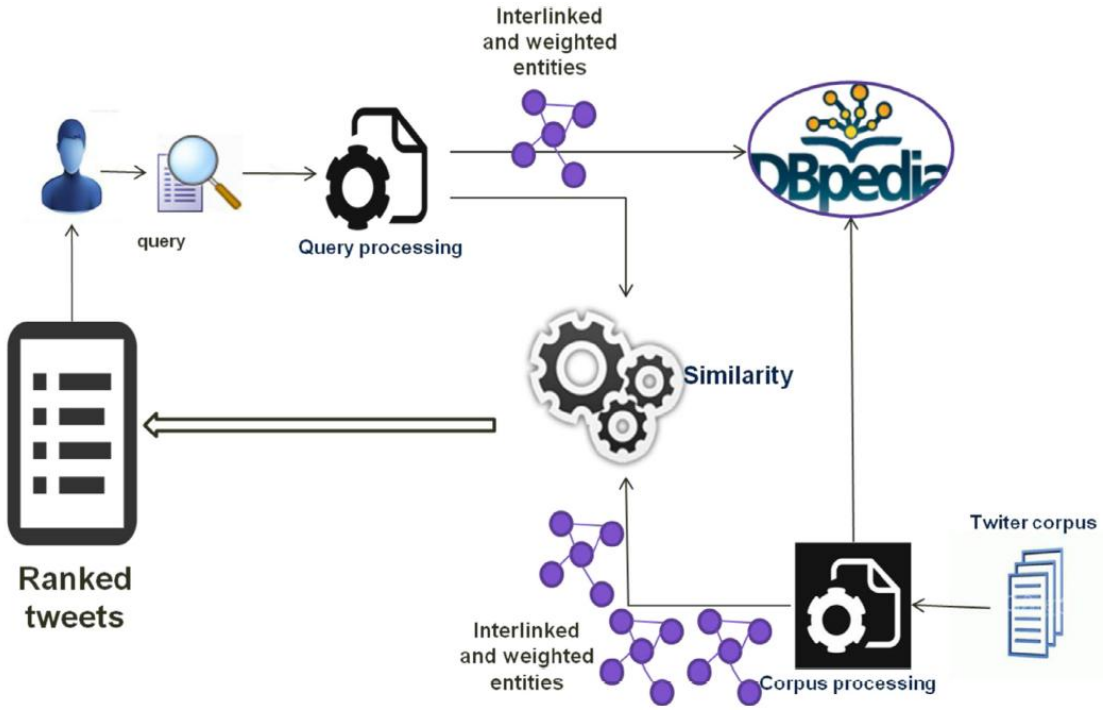
4.1 Graph-of-concepts Method for a Microblog Semantic Context Retrieval System

Οι F. Kalloubi, E. H Nfaoui και O. El Beqqali [27] προτείνουν μια μέθοδο γράφου εννοιών, που εξετάζει τις σχέσεις μεταξύ των εννοιών και των συναφών τους εννοιών, και περιγράφει κάθε στοιχείο στο γράφημα αξιοποιώντας τη συνδεδεμένη φύση της DBpedia ως βάση γνώσης και κεντρικότητα γράφων. Επιπλέον, προτείνουν ένα μέτρο ομοιότητας, που υπολογίζει την ομοιότητα μεταξύ δύο γραφημάτων. Αυτό το μέτρο λαμβάνει υπόψη την επικάλυψη μεταξύ ονομαστικών εννοιών, το οποίο έχει αποδειχτεί ότι επιτυγχάνει τα καλύτερα αποτελέσματα στην αναζήτηση μικρών κειμένων και τις σχέσεις μεταξύ των συναφών εννοιών στο γράφημα. Το μέτρο ομοιότητας δίνει, επιπρόσθετα, προτεραιότητα σε έννοιες που ταιριάζουν με τις ονομαζόμενες οντότητες στο κείμενο, επειδή έχουν μεγαλύτερο βάρος.

Η χρήση της μεθόδου επικεντρώνεται σε πλατφόρμες μικρο-ιστολογίων, microblogs, όπως το Twitter, όπου η επικοινωνία γίνεται με την ανταλλαγή πολύ μικρών κειμένων περιορισμένου αριθμού γραμμάτων, και οι οποίες αποτελούν την κύρια πλατφόρμα των Μέσων Κοινωνικής Δικτύωσης για την ανταλλαγή μηνυμάτων και περιεχομένου μεταξύ των χρηστών.

4.1.1 Μέθοδος

Δεδομένης της φύσης των αιτημάτων αναζήτησης στις πλατφόρμες microblogging, οι συγγραφείς καθορίζουν μια νέα μέθοδο για την εξαγωγή σχετικών οντοτήτων από τα κείμενα και για την προσθήκη σημασιολογίας στις δημοσιεύσεις αντιστοιχίζοντας επιφανειακές φόρμες, που απεικονίζουν το πλαίσιο στο οποίο λαμβάνει χώρα η δημοσίευση.



Σχ. 13: Η διαδικασία της μεθόδου Graph-of-concepts[27]

Χρησιμοποιούν λέξεις και φράσεις κλειδιά που συνήθως είναι και ονομαζόμενες οντολογίες στις πλατφόρμες. Αρχικά, αναζητείται η λέξη/φράση κλειδί που θα επιλεχθεί. Εάν υπάρχει ήδη φόρμα για αυτήν, τότε προστίθεται στη λίστα με τις ήδη εντοπισμένες. Σε περίπτωση που δεν υπάρχει, τότε εκτελείται διαχωρισμός των φράσεων/ουσιαστικών και γίνεται αναζήτησή της στη βάση δεδομένων μέχρι να βρεθεί αντιστοίχιση σε φόρμα. Για την υλοποίηση της αναζήτησης αυτής χρησιμοποιούνται πολλοί τύποι πηγών από τη DBpedia, όπως label, name κ.α..

Οι επιφανειακές φόρμες χρησιμοποιούνται στη συνέχεια για την κατασκευή του γράφου των διασυνδεδεμένων οντοτήτων που απεικονίζει το περιεχόμενο του κειμένου. Μετά την κατασκευή του γράφου οντοτήτων και σχέσεων, εντοπίζεται πόσο κεντρικός είναι ένας κόμβος χρησιμοποιώντας την κεντρικότητα του γράφου που βασίζεται στις σχέσεις μεταξύ των οντοτήτων. Για καλύτερες επιδόσεις αγνοούνται οι απομονωμένοι κόμβοι, μιας και τέτοιου τύπου κόμβοι δεν δίνουν επαρκείς πληροφορίες για το περιεχόμενό τους. Το μαθηματικό μοντέλο που χρησιμοποιείται για τον υπολογισμό της κεντρικότητας περιγράφεται αναλυτικά στην μελέτη[27].

Μετά το βήμα αυτό, έχει προκύψει ένα σύνολο διασυνδεδεμένων και σταθμισμένων εννοιών στο γράφο, που αντιπροσωπεύουν κάθε μικρό κείμενο. Αυτό δίνει τη

δυνατότητα να καθορίσουν τον αλγόριθμο ομοιότητας σημασιολογικού περιεχομένου. Ο αλγόριθμος ελέγχει το αίτημα του χρήστη και του συνόλου των γράφων διασυνδεδεμένων εννοιών για ομοιότητες. Χρησιμοποιεί δύο κριτήρια ομοιότητας: το τοπικό, που προσδιορίζει την ομοιότητα μεταξύ δύο εννοιών σε επίπεδο περιεχομένου στο οποίο παρουσιάζονται, και το καθολικό, που προσδιορίζει την ομοιότητα μεταξύ του αιτήματος του χρήστη και του κειμένου που αντιπροσωπεύεται από τους γράφους οντοτήτων.

Τέλος, τα αποτελέσματα του αλγορίθμου ταξινομούνται και επιστρέφονται στο χρήστη.

4.1.2 Αξιολόγηση

Τα αποτελέσματα της έρευνας [27] δείχνουν ότι:

- ✓ Η μέθοδος επιτυγχάνει καλά αποτελέσματα από άποψη ακρίβειας, ανάκλησης και τυπικών μέτρων αξιολόγησης της ανάκτησης πληροφοριών
- ✓ Επιτυγχάνει καλύτερα αποτελέσματα σε μακροσκελή ερωτήματα σε σύγκριση με σύντομα.
- ✓ Επιπλέον, για να αποδείξουν την αποτελεσματικότητα της προτεινόμενης μεταβλητής κεντρικότητας, συγκρίθηκαν οι επιδόσεις της με τον κλασσικό συντελεστή κεντρικότητας και τα πειραματικά αποτελέσματα δείχνουν ότι η προτεινόμενη αποφέρει πιο σημαντικά αποτελέσματα.

Ωστόσο, το σύστημά τους δεν είναι σε θέση να ικανοποιήσει με ακρίβεια τις ανάγκες των χρηστών, αν το ερώτημα δεν περιέχει ονομαζόμενες οντότητες. Όσο λιγότερες οντότητες υπάρχουν στο κείμενο, τόσο λιγότερο η μέθοδος αυτή μπορεί να επιφέρει τα επιθυμητά αποτελέσματα. Ο συνδυασμός της λεξικής ομοιότητας και της σημασιολογικής ομοιότητας μπορεί να λύσει αυτό το ζήτημα. Η ενσωμάτωση επιπλέον βάσεων γνώσης θα μπορούσε να βελτιώσει συνολικά το προτεινόμενο σύστημα.

4.2 Building a relatedness graph from Linked Open Data

Στην Τεχνολογία της Πληροφορικής είναι πολύ σημαντικό να προσδιορίζουμε εάν και κατά πόσο δύο οντότητες έχουν σχέση μεταξύ τους. Ο υπολογισμός της σχετικότητας μεταξύ ενός ζεύγους εννοιών μπορεί να είναι πολύ χρονοβόρος, καθώς συχνά απαιτεί όχι μόνο αμιγή πληροφορία, όπως περιγραφή κειμένου, αλλά και δεδομένα που προέρχονται από εξωτερικές πηγές, όπως ο αριθμός των συνδέσεων μιας ιστοσελίδας. Αυτός είναι ο λόγος που ο υπολογισμός της σχετικότητας γίνεται συνήθως εκτός σύνδεσης και τα αποτελέσματα αποθηκεύονται σε συγκεκριμένες δομές δεδομένων, τα γραφήματα γνώσης. Εδώ πρέπει να τονιστεί η διαφορά της σχετικότητας με την ομοιότητα, καθώς δύο έννοιες μπορεί να είναι σχετικές χωρίς να είναι όμοιες μεταξύ τους.

Οι T. Di Noia, V. C. Ostuni, J. Rosati, P. Tomeo, E. Di Sciascio, R. Mirizzi, C. Bartolini [28], δημιούργησαν ένα σημασιολογικό γράφο σχετικότητας, που δίνει έμφαση στο IT τομέα, και χρησιμοποιεί γνώση που προέρχεται από τη DBpedia τόσο για τον προσδιορισμό εννοιών με μοναδικό τρόπο, αλλά και για να υπολογιστούν οι σχετικοί μεταξύ τους κόμβοι, βασιζόμενοι σε μια σημασιολογική εξερεύνηση των γειτονικών κόμβων.

4.2.1 Μέθοδος

Διατυπώνουν το μέτρο συγγένειας ως ένα πρόβλημα κατάταξης. Με βάση το αίτημα του χρήστη, κατατάσσουν τα δεδομένα ανάλογα με τη σχετικότητά τους σε σχέση με αυτό. Χρησιμοποιούν διαφορετικά χαρακτηριστικά κατάταξης [28], τα οποία προέρχονται από την κατάταξη γράφων της DBpedia, την κατάταξη κειμένου της Wikipedia και την κατάταξη βάσει Ιστού. Στόχος τους είναι να αποκτήσουν ένα ισχυρό μέτρο σχετικότητας, που χρησιμοποιεί τα πλεονεκτήματα και απορρίπτει τα μειονεκτήματα και των τριών κατατάξεων:

- Η κατάταξη που βασίζεται σε γράφους της DBpedia τους επιτρέπει να εκμεταλλεύονται σαφείς σημασιολογικές συνδέσεις μεταξύ πόρων / δεξιοτήτων, αλλά δεν θα μπορούσαν να είναι αρκετές για να εκτιμήσουν τη συγγένεια μεταξύ των πόρων.
- Η προσέγγιση που υιοθετείται στην ταξινόμηση βάσει κειμένων της Wikipedia, επιτρέπει στο σύστημα να ανακαλύπτει συνδέσεις μεταξύ πόρων

με βάση κοινές λέξεις-κλειδιά, αλλά η έννοια πίσω από αυτές τις λέξεις-κλειδιά δεν αξιοποιείται.

- Η κατάταξη βάσει Ιστού παρέχει μια στατιστική υπόδειξη για τη σχετικότητα μεταξύ των πόρων, που βασίζεται στη δημοτικότητά τους και στην συνυπάρχουσα παρουσία τους στον Ιστό, αλλά στερείται σαφούς σημασιολογίας.

Το μαθηματικό μοντέλο περιγράφεται αναλυτικά στη μελέτη τους [28]. Συνοπτικά, τα βήματά του είναι τα εξής:

i. Εντοπίζονται και συλλέγονται όλοι οι πόροι από τη DBpedia που έχουν σχέση με τον τομέα που θα επικεντρωθεί η αναζήτηση, μετά από την διερεύνηση των RDF γράφων.
ii. Μετά την συλλογή των πόρων, λαμβάνεται το Λεξιλόγιο των πόρων αυτών που απαιτείται για τη κατασκευή του γράφου σχετικότητας, και αρχίζει ο υπολογισμός της σχετικότητας ανά ζεύγος.
iii. Στη συνέχεια, και αφού έχει δημιουργηθεί ο γράφος σχετικότητας, υπολογίζουν τα χαρακτηριστικά για ταξινόμηση των αποτελεσμάτων βάσει διάφορων χαρακτηριστικών.
iv. Μόλις υπολογίσουν τα διαφορετικά χαρακτηριστικά κατάταξης, τα συγκεντρώνουν για να αποκτήσουν μια αποτελεσματική και ακριβή κατάταξη των πόρων, χρησιμοποιώντας τόσο μια *supervised* προσέγγιση για να μάθουν μια λειτουργία κατάταξης, αλλά και μια *unsupervised* μέθοδο βασισμένη σε μια τεχνική ψηφοφορίας.
v. Τέλος, τα τελικά κατανεμημένα αποτελέσματα επιστρέφονται στο χρήστη.

4.2.2 Αξιολόγηση

Για την αξιολόγηση του μοντέλου τους, οι συγγραφείς [28] εκτέλεσαν εκτεταμένη έρευνα και σύγκριση με άλλες τεχνικές που να αξιολογούν αυτόματα τη σημασιολογική συγγένεια μεταξύ λέξεων, κειμένων ή εννοιών, με τρόπο που αντιστοιχεί στενά με αυτό των ανθρώπων. Οι περισσότερες παραδοσιακές μέθοδοι εκμεταλλεύονται συγκεκριμένες λεξικές πηγές όπως τα λεξικά ή καλά δομημένες ταξινομίες για τον υπολογισμό των σημασιολογικών μέτρων.

Τα πλεονεκτήματα του σημασιολογικού γράφου σχετικότητας που δημιουργήθηκε από τους συγγραφείς, σε σχέση με τις υπόλοιπες μεθόδους είναι τα ακόλουθα:

- ✓ Η χρήση της DBpedia καθιστά τη μέθοδο εφαρμόσιμη και σε άλλους τομείς. Μπορούν, επίσης, να χρησιμοποιηθεί και από domain τα οποία δεν είναι πλήρως ή πλούσια δομημένα, μιας και η προσέγγιση δεν βασίζεται μόνο σε οντολογικές πληροφορίες για τον υπολογισμό της σχετικότητας.
- ✓ Χρησιμοποιούν περισσότερα από ένα μέτρα και εργαλεία ώστε να μπορούν να αξιοποιούν τα πλεονεκτήματα όλων (κατάταξη γράφων, κειμένου, Ιστού).
- ✓ Δεν είναι απαραίτητο να γνωρίζουν όλα τα στοιχεία-κόμβους του γράφου, αλλά αρκεί ένα μικρό αντιπροσωπευτικό σετ κόμβων για να ανακαλύψουν όλους του υπόλοιπους συγγενικούς πόρους.
- ✓ Τέλος, δεν απαιτείται ένα αρχικό σετ δεδομένο που να διαθέτει ήδη μέτρα σχετικότητας για την εκπαίδευση του μοντέλου.

4.3 Content-based Recommender System using Semantic Web Knowledge Graph Embeddings

Στη Διδακτορική του διατριβή [20], ο P. Ristoski παρουσιάζει μια εκτεταμένη έρευνα πάνω στις προσεγγίσεις που χρησιμοποιούν τους γράφους γνώσης του Σημασιολογικού Ιστού στη Διερεύνηση Δεδομένων. Επικεντρώνεται κυρίως στα συστήματα συστάσεων, όπου οι γράφοι αποτελούν πολύτιμη πηγή ανάκτησης βασικής γνώσης.

Παρά τα δεδομένα από διάφορους τομείς που δημοσιεύονται ως LOD, οι πηγές που συνδυάζουν γνώση από πολλαπλούς τομείς, όπως η DBpedia, χρησιμοποιούνται κυρίως σε συστήματα συστάσεων. Δεδομένου ότι τα προς σύσταση στοιχεία-αντικείμενα είναι συνδεδεμένα σε ένα σύνολο LOD, η πληροφορία από το σύνολο αυτό μπορεί να χρησιμοποιηθεί για να προσδιοριστεί ποιο από τα αντικείμενα είναι πιο όμοιο σε αυτά που ο χρήστης έχει αναζητήσει/χρησιμοποιήσει στο παρελθόν.

Συνήθως, Τα επιλεγμένα δεδομένα εξάγονται από τη DBpedia και μετασχηματίζονται σε μια προτεινόμενη μορφή, δηλαδή κάθε κόμβος του γραφήματος αντιπροσωπεύεται από ένα επίπεδο διάνυσμα δυαδικών και / ή αριθμητικών χαρακτηριστικών. Ωστόσο,

η DBpedia περιέχει περισσότερες πληροφορίες από αυτές που εκφράζονται σε αυτές τις προτεινόμενες μορφές. Πιο συγκεκριμένα, οι σημασιολογικές διαδρομές μεταξύ των οντοτήτων είναι μια πολύ χρήσιμη πληροφορία για την σύσταση συστημάτων συσχέτισης μεταξύ τομέων.

Ένας από τους κύριους περιορισμούς των παραδοσιακών προσεγγίσεων των συστημάτων συστάσεων βάσει περιεχομένου είναι ότι οι πληροφορίες στις οποίες βασίζονται είναι γενικά ανεπαρκείς για να προκαλέσουν το ενδιαφέρον των χρηστών και να χαρακτηρίσουν όλες τις πτυχές της αλληλεπίδρασής τους με το σύστημα. Η εισαγωγή στοιχείων από το Σημασιολογικό Ιστό και ιδιαίτερα των γράφων γνώσης προτείνεται για την κάλυψη των συγκεκριμένων περιορισμών.

4.3.1 Μέθοδος

Στην προσέγγιση που παρουσιάζεται στην εν λόγω διατριβή [20], χρησιμοποιείται μια τεχνική ενσωμάτωσης γράφων RDF2vec [31] για τη δημιουργία χαρακτηριστικών, στο πλαίσιο των συστημάτων σύστασης με βάση το περιεχόμενο. Η προσέγγισή του βασίζεται σε έναν απλό αλγόριθμο συστάσεως.

Η μέθοδος αξιολογεί την εγγύτητα των αντικειμένων μέσω της ομοιότητας μεταξύ των αντίστοιχων χαρακτηριστικών διανυσμάτων, και στη συνέχεια επιλέγει ένα υποσύνολο αυτών των αντικειμένων για κάθε στοιχείο. Έτσι ο αλγόριθμος μπορεί να υπολογίσει τη βαθμολογία ενός νέου στοιχείου από το χρήστη με βάση τα αντικείμενα που έχουν ήδη αξιολογηθεί και την ομοιότητα που έχει υπολογιστεί. Ανάλογα με τη βαθμολογία του στοιχείου, το σύστημα καταλήγει στη σύσταση ή μη του αντικειμένου.

4.3.2 Αξιολόγηση

Για την αξιολόγηση της μεθόδου του, ο P. Ristoski αξιολογεί διαφορετικές παραλλαγές της προσέγγισής του σε τρία σύνολα δεδομένων, και τα συγκρίνει με τις κοινές προσεγγίσεις για τη δημιουργία συστημάτων που βασίζονται σε περιεχόμενο από τα LOD, καθώς και με τις καινοτόμες συνεργατικές και υβριδικές προσεγγίσεις. Επιπλέον, διερευνά τη χρήση δύο διαφορετικών γραφημάτων γνώσης, της DBpedia και της Wikidata. Τα αποτελέσματά της αξιολόγησής παρουσιάζονται αναλυτικά και γραφικά στη διατριβή του [20].

Αρχικά αποδεικνύει ότι ένα σύστημα συστάσεων περιεχομένου που βασίζεται στην ομοιότητα μεταξύ αντικειμένων που υπολογίζονται σύμφωνα με λανθάνοντα χαρακτηριστικά διανύσματα, φέρει καλύτερα αποτελέσματα από τον ίδιο τύπο συστήματος χρησιμοποιώντας ρητά χαρακτηριστικά (π.χ. τύπους, κατηγορίες κ.λπ.), τόσο σχετικά με την ακρίβεια όσο και με τη συνολική ποικιλομορφία των αποτελεσμάτων.

Σχετικά με τα σύνολα δεδομένων των LOD, συμπεραίνει ότι η DBpedia αποδίδει καλύτερα από τη Wikidata σε κάθε μέτρο που χρησιμοποιείται, και ότι η χρήση μεθόδων kernel για την εξαγωγή χαρακτηριστικών δεν προσδίδουν κάποιο πλεονέκτημα στον αλγόριθμο σύστασης.

Τέλος καταλήγει στο συμπέρασμα ότι η προσέγγισή του φέρει καλύτερα αποτελέσματα σε σχέση με τις καινοτόμες συνεργατικές και υβριδικές προσεγγίσεις.

5. Συμπεράσματα

Στην παρούσα εργασία παρουσιάστηκαν κάποιες μέθοδοι που χρησιμοποιούν Σημασιολογικά στοιχεία και Γράφους για τη βελτιστοποίηση της διαδικασίας Διερεύνησης Δεδομένων.

Η βιβλιογραφία που υπάρχει γενικά σε αυτό τον τομέα εμπλουτίζεται συνεχώς, αφού ο τομέας της Διερεύνησης Δεδομένων έχει πολλές εφαρμογές και μπορεί να αποφέρει τεράστια πλεονεκτήματα στους χρήστες.

Στο σύνολο των μεθόδων που περιγράφονται στην παρούσα εργασία, παρατηρούμε ότι η χρήση Σημασιολογικών Στοιχείων και Γράφων Γνώσης βελτιώνει σε κάθε περίπτωση τη διαδικασία Διερεύνησης. Αυξημένη ταχύτητα προσπέλασης δεδομένων, καλύτερο φιλτράρισμα δεδομένων, βέλτιστη απόδοση όσον αφορά την ποιότητα των αποτελεσμάτων και αυξημένη ικανοποίηση των χρηστών είναι μερικά από τα πλεονεκτήματα που συναντώνται κατά τη μελέτη των μεθόδων αυτών.

Ωστόσο, οι βασικές μέθοδοι Διερεύνησης έχουν καλυφθεί ικανοποιητικά από τις υφιστάμενες μελέτες έως σήμερα. Η ερευνητική κοινότητα στρέφεται στην βελτίωση των αποτελεσμάτων της Διερεύνησης μέσα από το συνδυασμό διάφορων συνιστωσών, όπως χρήση περισσότερων από μιας βάσης δεδομένων ή χρήση υβριδικών μεθόδων διερεύνησης, στην ταξινόμηση και αναπροσαρμογή των αποτελεσμάτων ώστε η παρουσίασή τους στους χρήστες να είναι πιο κατανοητή και προσπελάσιμη, και τέλος στην βελτίωση της εξατομίκευσης των μεθόδων εμβαθύνοντας στα κριτήρια δημιουργίας και διερεύνησης, όπως είναι η ομοιότητα, η συσχέτιση κ.α..

Νέα εργαλεία αναπτύσσονται συνεχώς για την ικανοποίηση των στόχων αυτών. Μέχρι τώρα δεν έχει αναπτυχθεί κάποιο το οποίο να είναι ικανό να καλύψει τις απαιτήσεις για ακρίβεια, συνέπεια, και ταχύτητα κατά τη Διερεύνηση Δεδομένων.

Τέλος, αξίζει να σημειωθεί ότι μια κατεύθυνση που αξίζει να ερευνηθεί είναι η επεκτασιμότητα και η δυνατότητα εύκολης μετάβασης των μεθόδων από τον ένα τομέα στον άλλον και από τη μια βάση δεδομένων σε άλλη. Τα έως τώρα μοντέλα συνήθως καταπιάνονται με ένα μόνο τομέα και μια βάση δεδομένων και επικεντρώνονται σε αυτά. Το Διαδίκτυο όμως προσφέρει αναρίθμητες πληροφορίες

και δεδομένα προς Διερεύνηση, τα οποία όμως δεν είναι πάντα δομημένα ή εκφρασμένα με τον επιθυμητό τρόπο.

6. Βιβλιογραφία

[1] A. Lausch, A. Schmidt, L. Tischendorf, "Data mining and linked open data – New perspectives for data analysis in environmental research", Ecological Modelling 295, 2015, p. 5-17

[2] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "Advances in knowledge discovery and data mining", American Association for Artificial Intelligence Menlo Park, CA, USA, 1996, p. 1-34

[3] L.H. Witten, F. Eibe, M.A. Hall, "Data mining practical machine learning tools and techniques", Morgan Kaufmann Publishers, 2011

[4] W.W. Chu, "Data mining and knowledge discovery for Big Data. Methodologies, challenge and opportunities", Springer Verlag Berlin Heidelberg, 2014

[5] RapidMiner: http://rapidminer.com

[6] P. Ritoski, C. Bizer, H. Paulheim, "Mining the Web of Linked Data with RapidMiner", Web Semantics: Science, Services and Agents on the World Wide Web 35, 2015, p.142-151

[7] R: http://www.r-project.org

[8] Orange: https://orange.biolab.si/

[9] WEKA: https://www.cs.waikato.ac.nz/~ml/weka/index.html

[10] KNIME: https://www.knime.com/

[11] https://data-flair.training/blogs/data-mining-tools-techniques/

[12] S. Irfan, B.V. Babu, "Information Retrieval in Big Data using Evolutionary Computation: A Survey", International Conference on Computing, Communication and Automation (ICCCA), 2016

[13] https://www.techentice.com/the-data-veracity-big-data/

[14] I. Lee, "Big data: Dimensions, evolutions, impacts, and challenges", Business Horizons 60, 2017, p. 293-303

[15] P. Ristoski, H. Paulheim, "Semantic Web in data mining and knowledge discovery: A comprehensive survey", Web Semantics: Science, Services and Agents on the World Wide Web 36, 2016, p. 1-22

[16] C. Bizer, T. Heath, T. Berners-Lee, "Linked Data – The story so far", Special Issue on Linked Data, International Journal on Semantic Web and Information Systems

[17] S. Kabir, S. Ripon, M. Rahman, T.Rahman, "Knowledge-based Data Mining using Semantic Web", IERI Procedia 7, 2014, p. 113-119

[18] Q.K. Quboa, M. Saraee, "A State-of-the-Art Survey on Semantic Web Mining", Intelligent Information Management 5, 2013, p. 10-17

[19] K. Sridevi, R. Umarani, "A Survey of Semantic based Solutions to Web Mining", International Journal of Emerging Trends & Technology in Computer Science vol.1 iss.2, 2012

[20] P. Ristoski, "Exploiting Semantic Web Knowledge Graphs in Data Mining", Διδακτορική Διατριβή, Πανεπιστήμιο του Manheim, 2017

[21] DBPedia: https://wiki.dbpedia.org/

[22] YAGO: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

[23] Wikidata: https://www.wikidata.org

[24] L. Ehrlinger, W. Wöß, “Towards a Definition of Knowledge Graphs”, Semantics 2016: Posters and Demos Track, Leipzig, Germany, 2016

[25] D2RQ tool: http://d2rq.org/

[26] M. Sah, V. Wade, “Personalized concept-based search on the Linked Open Data”, Web Semantics: Science, Services and Agents on the World Wide Web 36, 2016, p. 32-57

[27] F. Kalloubi, E. H Nfaoui, O. El Beqqali, “Micro blog semantic context retrieval system based on linked open data and graph-based theory”, Expert Systems With applications 53, 2016, p. 138-148

[28] T. Di Noia, V. C. Ostuni, J. Rosati, P. Tomeo, E. Di Sciascio, R. Mirizzi, C. Bartolini, “Building a relatedness graph from Linked Open Data: A case study in the IT domain”, Expert Systems With Applications 44, 2016, p. 354-366

[29] UMBEL: http://umbel.org/

[30] R. Delbru, S. Campinas, G. Tummarello, “Searching web data: an entity retrieval and high-performance indexing model”, Journal of Web Semantics 10, 2012, p. 33-58

[31] P. Ristoski, J. Rosati, T. Di Noia, R. de Leone, H. Paulheim, “RDF2Vec: RDF Graph Embeddings and Their Applications”, The Semantic Web Journal, 2016

[32] D. Moussallem, M. Wauer, A. Ngonga Ngomo, “Machine Translation using Semantic Web Technologies: A Survey”, Web Semantics: Science, Services and Agents on the World Wide Web 51, 2018

[33] J. Oliveira, C. Delgado, A.C. Assaife, “A recommendation approach for consuming linked open data”, Expert Systems With Applications 72, 2018, p. 407-420

[34] D. Karagiannis, R. Buchmann, “Linked Open Models: Extending Linked Open Data with conceptual model information”, Information Systems 56, 2016, p. 174-197