Πανεπιστήμιο Πελοποννήσου
Σχολή Θετικών Επιστημών και Τεχνολογίας
Τμήμα Επιστήμης και Τεχνολογίας Υπολογιστών

Μεταπτυχιακή Εργασία

ΣΗΜΑΣΙΟΛΟΓΙΚΕΣ ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Χριστίνα Βακάλογλου
ΑΜ: 2010001

Επιβλέπων καθηγητής: Κωνσταντίνος Κούτρας

Τρίπολη, 2013

Περιεχόμενα

Ευρετήριο Εικόνων & Πινάκων

Περίληψη

Το σημασιολογικό διαδίκτυο, είναι το διαδίκτυο επόμενης γενιάς που αναπαριστάται από μια αρχιτεκτονική πολλών επιπέδων και παρέχει τις στρατηγικές για την υποστήριξη πολλών υπηρεσιών. Συντελεί έναν καθοριστικό ρόλο στην εξέλιξη της Ανάκτησης Πληροφορίας (IR) για την παροχή υπηρεσιών υψηλότερου επιπέδου, μέσω της χρήσης αυτοματοποιημένων πρακτόρων που επικοινωνούν και ανταλλάσσουν πληροφορίες. Η φιλοσοφία του σημασιολογικού ιστού είναι η δημιουργία ενός διαδικτύου που βασίζεται στις σχέσεις που υπάρχουν μεταξύ των πόρων, δηλαδή των αντικειμένων του πραγματικού κόσμου, σε αντίθεση με τον παγκόσμιο ιστό που είναι ένα διαδίκτυο βασισμένο σε κείμενα. Οι μηχανές αναζήτησης που λειτουργούν στο σημασιολογικό διαδίκτυο ονομάζονται Σημασιολογικές Μηχανές Αναζήτησης. Η θεμελιώδης διαφορά μεταξύ αυτών των νέων συστημάτων αναζήτησης έναντι των παραδοσιακών, είναι ότι στοχεύουν στην προσομοίωση της ανθρώπινης λογικής, ώστε να διαχειρίζονται τα αιτήματα των χρηστών τους σε εννοιολογικό επίπεδο, παρέχοντας ταυτόχρονα άμεση, έγκυρη και ενημερωμένη γνώση. Η χρήση της σημασιολογίας στην αναζήτηση είναι αποτελεσματική κατά την αναζήτηση διερεύνησης, όταν δηλαδή ο χρήστης αναζητά να ερευνήσει και να συλλέξει πληροφορίες για ένα αντικείμενο, χωρίς να γνωρίζει συγκεκριμένες πηγές. Αντίθετα παραμένει αντικείμενο της παραδοσιακής έρευνας και δεν απασχολεί την σημασιολογία, η πλοηγική αναζήτηση, κατά την οποία ο χρήστης ενδιαφέρεται να μεταβεί είτε σε συγκεκριμένες ιστοσελίδες, είτε να εντοπίσει σελίδες που περιέχουν μια φράση ή έναν συνδυασμό λέξεων. Το έργο των σημασιολογικών μηχανών αναζήτησης είναι σύνθετο και πολυεπίπεδο, καθώς επιδιώκουν την ικανοποίηση ερωτημάτων διατυπωμένων είτε με λέξεις κλειδιά είτε σε φυσική γλώσσα και ταυτόχρονα λαμβάνουν υπ' όψιν τους παράγοντες που σχετίζονται με την τοποθεσία, την επικαιρότητα, την πολυσημία των όρων, την αξιοπιστία των πηγών κ.α.

Για την πραγματοποίηση των στόχων της σημασιολογικής έρευνας απαιτείται η δόμηση της γνώσης και η ενσωμάτωση σημασιολογίας, ώστε τα δεδομένα να είναι αντιληπτά και επεξεργάσιμα σε επίπεδο μηχανής. Για την αναπαράσταση της γνώσης και την χρησιμοποίησή της από τα διάφορα συστήματα, το σημασιολογικό διαδίκτυο χρησιμοποιεί τις γλώσσες RDF και OWL που συντάσσονται στην XML. Η γλώσσα

RDF αποτελείται από την τριάδα υποκείμενο - ιδιότητα – αντικείμενο, που περιέχει πληροφορίες για την περιγραφή των πόρων και η γλώσσα OWL που χωρίζεται σε τρεις υπογλώσσες, χρησιμοποιείται για την αναπαράσταση των οντολογιών, δηλαδή των εννοιών που χρησιμοποιούνται για την περιγραφή μιας θεματικής περιοχής.

Αναφορικά με το ζήτημα της εύρεσης και μεταφοράς των δεδομένων, το σημασιολογικό διαδίκτυο χρησιμοποιεί ειδικά αυτοματοποιημένα προγράμματα, τους πράκτορες (agents) που στο σύνολό τους συνθέτουν ένα πολυπρακτορικό σύστημα στο οποίο συνεργάζονται και ενεργούν. Η σημασιολογική διαδικασία αρχίζει όταν οι πράκτορες λάβουν τα αιτήματα των χρηστών που στην συνέχεια θα επεξεργαστούν για να αναζητήσουν τα κατάλληλα αποτελέσματα και να τα προωθήσουν στους χρήστες. Καθ' όλη την διάρκεια της σημασιολογικής διαδικασίας, τα δεδομένα περνούν από τα διαφορετικά στάδια που απαρτίζεται μια σημασιολογική μηχανή αναζήτησης, καθένα από τα οποία αναλαμβάνει την επιτέλεση διαφορετικών λειτουργιών. Τα στάδια αυτά είναι η ανίχνευση του συστήματος για την ικανοποίηση της πληροφοριακής ανάγκης των χρηστών, η ευρετηρίαση των δεδομένων, ο συμπερασμός, η ανάκτηση και ταξινόμηση των αποτελεσμάτων.

Απόρροια της πολύχρονης σημασιολογικής έρευνας ήταν η δημιουργία της πρώτης σημασιολογικής μηχανής, της Swoogle, το 2004. Στη συνέχεια ακολούθησε ένας μεγάλος αριθμός σημασιολογικών συστημάτων, με διαφορετική στρατηγική κατεύθυνση και τεχνολογική εφαρμογή όπως π.χ. Watson, Sindice, Sig.Ma, Falcons, SWSE, PowerAqua, PowerMagpie κ.α. που κατάφεραν να πετύχουν μεγάλο βαθμό αποδοτικότητας και ποιότητας στις υπηρεσίες τους. Ωστόσο, το μέλλον των σημασιολογικών μηχανών αναζήτησης απαιτεί περισσότερη μελέτη από την επιστημονική κοινότητα, ώστε τα συστήματα αυτά να ανταποκριθούν στο έπακρο στους στόχους τους και να αντιμετωπίσουν επιτυχώς τις υπάρχουσες προκλήσεις.

Αναφορικά με την δομή των κεφαλαίων που ακολουθεί, το πρώτο κεφάλαιο περιλαμβάνει γενικές έννοιες των μηχανισμών ανάκτησης πληροφορίας, σύντομη ιστορική αναδρομή, έννοιες, στόχους και προκλήσεις των σημασιολογικών μηχανών αναζήτησης καθώς και την σύγκριση των χαρακτηριστικών και των στόχων των δύο συστημάτων, των παραδοσιακών και σημασιολογικών.

Στο δεύτερο κεφάλαιο περιγράφεται η αρχιτεκτονική του Σημασιολογικού διαδικτύου πάνω στο οποίο βασίζονται οι σημασιολογικές μηχανές αναζήτησης και αναλύονται τα επιμέρους επίπεδά της.

Το κεφάλαιο τρία, περιγράφει την έννοια του πράκτορα και την λειτουργία ενός πολυπρακτορικού συστήματος.

Στην συνέχεια στο κεφάλαιο τέσσερα, γίνεται αναφορά της αρχιτεκτονικής μιας σημασιολογικής μηχανής αναζήτησης καθώς και η περιγραφή της διαδικασία της σημασιολογικής έρευνας.

Ακολουθεί το κεφάλαιο πέντε που ασχολείται με την επεξήγηση της επεξεργασίας των σημασιολογικών δεδομένων και συγκεκριμένα, της επεξεργασίας των κειμένων, της ερμηνείας και κατανόησης του περιεχομένου των ερωτημάτων των χρηστών και της ταξινόμησης των αποτελεσμάτων από την μηχανή αναζήτησης.

Τέλος, στο έκτο κεφάλαιο περιγράφονται κάποια παραδείγματα σημασιολογικών συστημάτων που ήδη εφαρμόζονται και αναπτύσσεται η στρατηγική και ο τρόπος λειτουργίας τους.

Abstract

The semantic Web is the Web of the next generation, which is represented by a multilayer architecture and it provides the strategies to support many services. It has also an important role in Information Retrieval evolution for providing highest level services, through the usage of automated agents who communicate and swap information. Semantic Web's philosophy is the creation of a web which is based on relations between resources, which means real world objects, in contrast to World Wide Web which is based on documents. Search engine which operate on semantic web are called Semantic Search Engines. A fundamental difference between traditional and new search systems is that they aim to simulate human logic, so as to handle users' queries in semantic level by providing direct, valid and up-to-date knowledge. The usage of semantic in search is effectual during research search, which means when a user is looking for investigating and collecting information about a specific object and without to know where to search. In contrary, it stills remain a traditional search object and it does not part of semantic search, navigational search, during this the user is interested in being transformed whether in specific web pages, or to find pages which contain a phrase or a word combination. Semantic Search Engine's duty is difficult and multilevel, as they aimed at fulfilling queries posed with keywords or natural language and simultaneously they consider parameters which are related to location, polysemy, resources reliability etc.

For the realization of semantic search goals, it is necessary the knowledge to be structured and also a semantic incorporation to be performed, so as to data can be perceived and editable in machine level. Semantic web uses RDF and OWL languages which are expressed in XML for knowledge representation and usage from multiple systems. RDF consists of triples subject – property - object which contain information for resources description and OWL language which is divided in three sublanguages, it is used to represent ontologies, that is the meaning which are used for a domain description.

As regards the matter of discovering and data transfer, semantic web uses specified automated programs, agents who compose a multi-agent system in which they cooperate and react. Semantic process starts when agents get users queries which will edit, for searching best results and to present to users. During the semantic process,

the data pass through different stages of a semantic engine and every one undertake different functions. These stages are system crawling for satisfying users information needs, data indexing, inference, data retrieval and ranking.

A result of perennial semantic research was the creation of the first semantic search engine, Swoogle on 2004. It follows then a great range of semantic systems, with different strategy direction and technological application such as Watson, Sindice, Sig.Ma, Falcons, SWSE, PowerAqua, PowerMagpie etc. which they manage to achieve a great efficiency and quality degree in their quality. However semantic search engine future demands more research by the scientific community, so as to these systems can correspond to their goals to the fullest and to counteract successfully the existing challenges.

As concerned the structure's chapters which are included next, the first chapter includes general concepts about information retrieval, a brief historical description, concepts, goals and challenges of semantic web search engines, a comparison of characteristics and goals between these two systems, traditional and semantic.

In the second chapter we have a description of the semantic web architecture on which semantic search engines are based and every level is analyzed.

Chapter three describes the concept of an agent and the function of on multi-agent system.

Next, in chapter four, is a reference of semantic web engines architecture and the description of semantic search procedure.

It follows chapter five which is occupied with the explanation of semantic data process and specially, the meaning and understanding of user's queries content and result ranking by a semantic search engine.

Finally, in six chapter are mentioned some examples of some semantic systems which already exist, their strategies and their function way.

Εισαγωγή

Τα τελευταία έτη η χρήση του Διαδικτύου έχει αυξηθεί και συνεχίζει να αυξάνεται με ραγδαίο ρυθμό καθώς προσφέρει ένα μεγάλο σύνολο υπηρεσιών και δυνατοτήτων όπως ενημέρωση, ψυχαγωγία, κοινωνική δικτύωση, επικοινωνία, αγοραπωλησίες, εργασία από το σπίτι, εξ αποστάσεως εκπαίδευση. Τα παραπάνω μπορούν να μας προσφερθούν ευκολότερα δεδομένης της μείωσης του κόστους πρόσβασης στον παγκόσμιο ιστό και των ευρυζωνικών υπηρεσιών. Η πρόσβαση στην πληροφορία επιτυγχάνεται άμεσα και αποτελεσματικά μέσω των μηχανών αναζήτησης, δίνοντας τη δυνατότητα στους χρήστες να ανακτούν τις πληροφορίες που ζητούν, χωρίς να διαθέτουν εξειδικευμένες γνώσεις. Το διαδίκτυο χωρίς τις μηχανές αναζήτησης θα είχε εντελώς διαφορετική μορφή από τη σημερινή. Ο μεγάλος όγκος των πληροφοριών που διατίθεται στο διαδίκτυο θα ήταν δύσκολο να προσπελαστεί, να ανακτηθεί και να αξιοποιηθεί από τους χρήστες. Στατιστικές έρευνες που έχουν διεξαχθεί απεικονίζουν την σημαντικότητά τους. Στον παρακάτω πίνακα παρουσιάζεται ο ημερήσιος αριθμός αναζητήσεων της Google, της δημοφιλέστερης μηχανής αναζήτησης στο διαδίκτυο, κατά τα έτη 1998-2011. Η αύξηση των χρηστών είναι εντυπωσιακή. Το 2007 ο αριθμός αναζητήσεων είναι 1.200.000.000, ενώ το 2011 ο αριθμός αυτός ανέρχεται στα 4.717.000.000.

Year	Annual Number of Google Searches	Average Searches Per Day
2011	1,722,071,000,000	4,717,000,000
2010	1,324,670,000,000	3,627,000,000
2009	953,700,000,000	2,610,000,000
2008	637,200,000,000	1,745,000,000
2007	438,000,000,000	1,200,000,000
2000	22,000,000,000	60,000,000
1998	3,600,000	9,800

Πίνακας 1-Στατιστικά στοιχεία για την αύξηση των αναζητήσεων στην Google, τα έτη 1998 -2011 [1]

Η αύξηση αυτή, έχει επιφέρει στις εταιρίες υπηρεσιών αναζήτησης δημοτικότητα και τεράστια κερδοφορία, μέσω των στρατηγικών μάρκετινγκ που εφαρμόζουν, με αποτέλεσμα την αυξημένη ανταγωνιστικότητά τους, για την επικράτηση στο χώρο

του διαδικτύου. Συνέπεια όλων αυτών, είναι η αδιάκοπη προσπάθεια από μέρους τους για την βελτίωση της ποιότητας των υπηρεσιών αναζήτησης- που χαίρονται οι χρήστες ανά τον κόσμο- παράλληλα με την ανάγκη της εποχής, για ολοένα προηγμένες τεχνολογικές επιτεύξεις.

Ωστόσο ο προσανατολισμός των υπηρεσιών ανάκτησης πληροφορίας και η μελλοντική τάση που παρουσιάζεται, είναι η δημιουργία έξυπνων μηχανών αναζήτησης με συστήματα λειτουργίας που ανατρέπουν τον παραδοσιακό τρόπο ανάκτησης πληροφορίας. Οι μηχανές αυτές καλούνται σημασιολογικές μηχανές αναζήτησης – Semantic Search Engines και θα αποτελέσουν το αντικείμενο της εργασίας αυτής. Η ιδέα για την ύπαρξη τους προέρχεται από την προσπάθεια μετατροπής του Παγκόσμιου Ιστού σε Σημασιολογικό δίκτυο, αποτελούμενο από ένα σύνολο δομημένων δεδομένων με λογική, οργάνωση και σημασιολογία ώστε να είναι αντιληπτά σε επίπεδο μηχανής.

1. Εισαγωγή στις Σημασιολογικές Μηχανές Αναζήτησης (Semantic Search Engines)

1.1 Πρόλογος

Η μηχανή αναζήτησης είναι το εργαλείο που χρησιμοποιείται κατά την διαδικασία της Ανάκτησης Πληροφορίας (Information Retrieval) στον Παγκόσμιο Ιστό. Ακολουθώντας την τεχνολογική εξέλιξη, οι μηχανές αναζήτησης γίνονται πιο έξυπνες, αποδοτικές και ολοκληρωμένες συνδυάζοντας στην ήδη υπάρχουσα τεχνολογία τους, μεθοδολογίες λογικού συμπερασμού και εννοιολογικής σημασίας. Η δομή του κεφαλαίου περιέχει τις βασικές έννοιες που αφορούν τις μηχανές αναζήτησης, καθώς επίσης την ιστορία τους και την εξέλιξή τους, που θα οδηγήσει στην ανάπτυξη μιας νέας πραγματικότητας στην αναζήτηση της πληροφορίας, αυτής της Σημασιολογικής Αναζήτησης.

1.2 Τι είναι μια Μηχανή Αναζήτησης

Μια μηχανή αναζήτησης είναι μια εφαρμογή που εξυπηρετεί στην εύρεση πληροφοριών στο διαδίκτυο[2]. Διαθέτει μια Βάση Δεδομένων, στην οποία αποθηκεύονται τεράστιες συλλογές κειμένων και αρχείων διαφόρων ειδών, που έχουν δημοσιευθεί και διατίθενται στον ιστό. Η αναζήτηση πληροφορίας γίνεται με τη χρήση λέξεων- κλειδιών που θέτουν οι χρήστες ως ερωτήματα. Τα αποτελέσματα που επιστρέφονται, αποτελούνται από μια λίστα διευθύνσεων ιστοσελίδων με τις λέξεις του ερωτήματος, ταξινομημένη με βάση την πολιτική της εκάστοτε μηχανής αναζήτησης.

Μία μηχανή αναζήτησης αποτελείται από τρία μέρη: την Βάση Δεδομένων με τις ιστοσελίδες του διαδικτύου, την μηχανή αναζήτησης που λειτουργεί πάνω σε αυτή και ένα σύνολο προγραμμάτων που προσδιορίζει τον τρόπο με τον οποίο η μηχανή αναζήτησης θα παρουσιάσει τα αποτελέσματα.

1.3 Πρώιμες Μηχανές Αναζήτησης – Ιστορική Αναδρομή

Η τεχνολογία που διαθέτουν οι μηχανές αναζήτησης παρήλθε από πολλά στάδια μέχρι σήμερα. Η εξέλιξή αυτή, σημειώθηκε με αλματώδεις ρυθμούς χάρις στις

αυξανόμενες ανάγκες, την πολυμορφία των χρηστών και στην προσπάθεια ενσωμάτωσης καινοτόμων χαρακτηριστικών στον τρόπο ανάκτησης της πληροφορίας και εδραίωσης των εταιριών στο παγκόσμιο ιστό. Οι πρώτες αξιόλογες προσπάθειες στην ανάπτυξη μηχανών αναζήτησης, σημειώθηκαν το 1993, όταν εισήχθησαν μηχανές αναζήτησης με διαφορετική λειτουργία και χαρακτηριστικά [3].

Η μηχανή αναζήτησης Jumpstation Search Machine, που συγκέντρωνε τις πληροφορίες από τους τίτλους των κειμένων των ιστοσελίδων και έδινε την ζητούμενη πληροφορία μέσω απλής γραμμικής αναζήτησης, δεν ήταν σε θέση να συμβαδίσει με την αύξηση του διαδικτύου και την ταχύτητα αύξησης των δεδομένων.

Μια άλλη μηχανή αναζήτησης, η World Wide Web Worm, δημιούργησε ευρετήρια με URL's και τίτλους ιστοσελίδων και παρέθετε τα αποτελέσματα με την σειρά που τα εντόπιζε.

Επίσης την ίδια χρονιά, δημιουργήθηκε η μηχανή αναζήτησης Repository-Based Software Engineering Spider (RSBE) αλλά σε αντίθεση με τις άλλες δύο μηχανές αναζήτησης,εφάρμοσε ένα σύστημα ταξινόμησης των αποτελεσμάτων αναζήτησης.

Το τρωτό σημείο και στις τρεις αυτές μηχανές αναζήτησης ήταν ότι ο χρήστης έπρεπε να γνωρίζει ακριβώς την ονομασία του κειμένου που ψάχνει.

Η ανάκτηση πληροφορίας έγινε αποδοτικότερη, με την δημιουργία της μηχανής Excite, η οποία εφάρμοζε μια στατιστική ανάλυση των σχέσεων μεταξύ των λέξεων. Το 1999 αγοράστηκε από την εταιρεία @Home, η οποία όμως μετά από δύο χρόνια χρεοκόπησε.

Το 1994 δημιουργήθηκε ο Web-Crawler, ο πρώτος ανιχνευτής που ευρετηριάζει ολόκληρες ιστοσελίδες. Αποτελεί αναπόσπαστο κομμάτι των μηχανών αναζήτησης μέχρι και τις μέρες μας.

Μια άλλη μηχανή αναζήτησης, η Lycos, παρείχε ένα μεγάλο κατάλογο από κείμενα και επιπρόσθετα με την λειτουργία ταξινομημένης ανάκτησης πληροφορίας, παρείχε ταίριασμα προθέματος και μπόνους εγγύτητας λέξης.

Η Infoseek δημιουργήθηκε επίσης το 1994. Υπήρξε ιδιαίτερα δημοφιλής από την επόμενη κιόλας χρονιά χωρίς να εφαρμόζει κάποια ιδιαίτερη καινοτομία στον τρόπο αναζήτησης, ενώ υπήρξε αρκετά ευάλωτη στις επιθέσεις spam.

Το έτος 1995, η Altavista, εισήλθε στο χώρο της ανάκτησης πληροφορίας, επιφέροντας προηγμένα χαρακτηριστικά αναζήτησης. Ένα από τα κυριότερα ήταν η δυνατότητα χρήσης φυσικής γλώσσας για την υποβολή ερωτημάτων. Το έτος 2003 η Altavista αγοράστηκε από την εταιρεία Overture.

Το 1996, δημιουργήθηκε η Inktomi, μια ευρέως διαδεδομένη μηχανή αναζήτησης, παρά την επιρρέπειά της στις επιθέσεις spam. Το 2003 πουλήθηκε στον διαδικτυακό κατάλογο Yahoo!, λόγω της αδυναμίας της να εδραιωθεί ως μια κερδοφόρα επιχείρηση.

Την επόμενη χρονιά, η μηχανή αναζήτησης Ask Jeeves έγινε από τις πιο δημοφιλής μηχανές αναζήτησης, χάριν στην δυνατότητα που παρείχε να θέτονται ερωτήματα στη φυσική γλώσσα, σε συνδυασμό με τη χρήση λέξεων- κλειδιών. Παρόλη την διάδοση της δεν κατάφερε να ορθοποδήσει οικονομικά και αγοράστηκε από την InterActiveCorp - IAC. Η Ask χρησιμοποιείται έως τις μέρες μας, χωρίς όμως να καταφέρνει να ανταγωνιστεί επιτυχώς άλλες εταιρείες, που κατέχουν την μερίδα του λέοντος στην αναζήτηση πληροφορίας.

Η AllTheWeb είναι μία ακόμα μηχανή αναζήτησης που λανσαρίστηκε το 1999 και παρείχε μια διεπαφή χρήστη εμπλουτισμένη με πολλά προηγμένα χαρακτηριστικά. Το 2003 αγοράστηκε από την Yahoo!.

Την περίοδο αυτή δημιουργούνται και οι πρώτες Μετα-μηχανές αναζήτησης. Μια μετα-μηχανή αναζήτησης συγκεντρώνει στην ιστοσελίδα της αναδιοργανωμένα αποτελέσματα από άλλες μηχανές αναζήτησης, απαλείφοντας τις επαναλαμβανόμενες εγγραφές. Η πρώτη μετα-μηχανή αναζήτησης ήταν η Meta-crawler, η οποία έκανε αναζήτηση στις μηχανές αναζήτησης Lycos, Altavista, Excite, Web-Crawler και Google και κατόπιν τα συνδύαζε και τα κατηγοριοποιούσε σε μια ιστοσελίδα. Οι μετα-μηχανές αναζήτησης ήταν χρήσιμες το διάστημα κατά το οποίο, οι μηχανές αναζήτησης περιείχαν διαφορετικό περιεχόμενο στη βάση δεδομένων τους, λόγω των μη τελειοποιημένων υπηρεσιών που παρείχαν για την ανίχνευση του διαδικτύου. Εντούτοις, δεν συμπεριλαμβάνουν στις αναζητήσεις τους, δυνατότητες που προσφέρονται από κάθε μηχανή ξεχωριστά[4].

Το 1997 δημιουργήθηκε η μηχανή αναζήτησης GoTo, που αργότερα μετονομάστηκε σε Overture και στη συνέχεια αγόρασε τις μηχανές AllTheWeb και Altavista. Τέλος το 2003, η Yahoo! αγοράζει την Overture υπό την ονομασία Yahoo Search. Με αυτή την συνεργασία και έχοντας στην κατοχή της τρεις μηχανές αναζήτησης, είναι η πρώτη εταιρεία που επωφελήθηκε από διάφορες μεθόδους μάρκετινγκ μέσω μηχανών αναζήτησης π.χ. το μοντέλο διαφήμισης Pay for click, με το οποίο ο διαφημιζόμενος καταβάλλει ένα αντίτιμο στην μηχανή αναζήτησης, για κάθε επίσκεψη στον ιστότοπό του[5].

Η Google, η κυρίαρχη μηχανή αναζήτησης, ξεκίνησε το 1998 και ήδη από το 2000 εδραιώθηκε στο χώρο του διαδικτύου. Η επιτυχία της οφείλεται κατά κύριο λόγο, στην καινοτομία να κατατάσσει τα αποτελέσματα με βάση τη δημοτικότητα των κειμένων, που υπολογίζεται μέσω του αλγόριθμου Page Rank. Η Google συνεχίζει να εξελίσσεται και να παρέχει στους χρήστες της ολοένα και πιο προηγμένες υπηρεσίες υψηλής ποιότητας.

Μια ακόμα δημοφιλής μηχανή αναζήτησης, η MSN Search της Microsoft, λειτούργησε την ίδια χρονιά. Εν συνεχεία, πήρε το όνομα Live Search, ενώ σήμερα είναι γνωστή με το όνομα Bing και προσφέρει προηγμένες υπηρεσίες.

1.4 Βασική Αρχιτεκτονική Μηχανών Αναζήτησης

Ο μηχανισμός ανάκτησης πληροφορίας μέσω μιας μηχανής αναζήτησης, επιτυγχάνεται μέσω ενός πράκτορα λογισμικού, του ανιχνευτή διαδικτύου (Web Crawler, ή αλλιώς Web Spider ή Web Robot). Ο ανιχνευτής μετακινείται στο διαδίκτυο και επισκέπτεται μία λίστα URL's, που του δίνεται από έναν URL server. Σκοπός του είναι να δημιουργήσει αντίγραφα όλων των ιστοσελίδων διαδικτύου, καθώς και ένα αρχείο με όλους τους υπερσυνδέσμους που βρίσκονται σε αυτές. Μετά το πέρας της διαδικασίας συλλογής, που καλείται ανίχνευση διαδικτύου (web crawling), γίνεται η εξαγωγή των λέξεων όλων των ιστοσελίδων μέσω του προγράμματος indexer, και καταγράφονται τα URL's τους. Με τον τρόπο αυτό επιτυγχάνεται η περαιτέρω επεξεργασία τους και η δημιουργία ευρετηρίου (index) με όλους τους όρους του διαδικτύου. Οι σελίδες που ανακτώνται κατά την διάρκεια της ανίχνευσης και της ευρετηρίασης διατηρούνται προσωρινά σε ένα αποθετήριο ιστοσελίδων ενώ διατηρούνται και οι σελίδες που χρησιμοποιήθηκαν πρόσφατα μετά από αναζήτηση, στη μνήμη cache για πιο εσπευσμένη μελλοντική χρήση.

Κατά την υποβολή ερωτημάτων από τους χρήστες, επιχειρείται το ταίριασμα των λέξεων κλειδιών και των λέξεων του ευρετηρίου και εν συνεχεία τα αποτελέσματα κατατάσσονται και παραθέτονται από ένα πρόγραμμα ταξινόμησης.

Κάθε μηχανή αναζήτησης, εφαρμόζει διαφορετικούς αλγορίθμους και μηχανισμούς ανάκτησης πληροφορίας που βασίζονται όμως σε αυτό το γενικό μοντέλο αρχιτεκτονικής.

Εικόνα 1.4- Βασική Αρχιτεκτονική Μηχανής Αναζήτησης [6]

1.5 Σημασιολογικές Μηχανές Αναζήτησης (Semantic Search Engines)

Η σημασιολογία (Semantics, από το ελληνικό επίθετο σημαντικός) ορίζεται ως η επιστημονική μελέτη της γλωσσικής σημασίας [7]. Ο κλάδος εξετάζει την πολλαπλότητα των σημασιών που είναι δυνατόν να χαρακτηρίζουν μια λέξη ή πρόταση και τις σημασιολογικές τους σχέσεις. Η σημασιολογία συνδέεται στενά με διάφορες περιοχές της επιστήμης των υπολογιστών όπως είναι ο προγραμματισμός, η λογική, τα δίκτυα κ.α. Τα τελευταία χρόνια άρχισε η εφαρμογή της σημασιολογικής έρευνας στην αναζήτηση όρων μέσω μηχανών αναζήτησης, με καινοτόμες προσεγγίσεις και με την φιλοδοξία να βελτιστοποιηθεί στο εγγύς μέλλον ο παραδοσιακός τρόπος αναζήτησης πληροφοριών.

Η **Σημασιολογική αναζήτηση** ορίζεται ως η αναζήτηση για πληροφορία που επιδιώκει να βελτιώσει την ακρίβεια των αποτελεσμάτων αναζήτησης, προσπαθώντας να αντιληφθεί την πραγματική πρόθεση του ερωτήματος του χρήστη. Η αναζήτηση δεν βασίζεται στην ερμηνεία του κάθε όρου ξεχωριστά, αλλά στην προσπάθεια ανάλυσης της σημασίας των λέξεων, ώστε να παραχθούν αποτελέσματα μεγαλύτερης συνάφειας με τα ερωτήματα.

Η ανάγκη για σημασιολογική αναζήτηση προέκυψε από την αδυναμία των μηχανών αναζήτησης να μιμηθούν τον τρόπο αντίληψης των ανθρώπων. Π.χ. ένα αμφίσημο ερώτημα γίνεται αντιληπτό από τον άνθρωπο από τα συμφραζόμενα μιας πρότασης,

πράγμα ανέφικτο από μια παραδοσιακή μηχανή αναζήτησης. Συνεπώς χρειάζεται να δημιουργηθούν έξυπνες μηχανές αναζήτησης που θα ενσωματώνουν εφαρμογές τεχνητής νοημοσύνης, ώστε να είναι σε θέση να αντιλαμβάνονται τις σχέσεις που υπάρχουν μεταξύ διαφορετικών οντοτήτων.

Συγκεκριμένα, οι σημασιολογικές μηχανές αναζήτησης θα πρέπει να είναι έξυπνες σε τέτοιο βαθμό ώστε να καταλαβαίνουν την πρόθεση του χρήστη, να έχουν την ικανότητα να διακρίνουν το είδος των όρων (ρήματα, ονόματα, έννοιες) αλλά και να συγκρίνουν ένα ερώτημα του χρήστη με άλλα σύνολα γνώσης. Εν συνεχεία, εφόσον γίνει κατανοητό το νόημα του ερωτήματος, στόχος είναι η παροχή άμεσης και συναφούς πληροφορίας, σε συνδυασμό με μια λίστα ιστοσελίδων αποτελούμενες από έγγραφα με τις λέξεις- κλειδιά του ερωτήματος.

Η κατανόηση της πρόθεσης του χρήστη, μπορεί να επιτευχθεί με την αποσαφήνιση των όρων του ερωτήματος. Από τις πολλαπλές έννοιες ενός όρου επιλέγονται οι πιο πιθανές, λαμβάνοντας υπόψη και τις σημασίες των άλλων λέξεων που υπάρχουν στο ερώτημα. Η αποσαφήνιση μιας έννοιας επηρεάζει και τον προσδιορισμό των εννοιών των άλλων όρων έως ότου καταλήξουμε σε μια πρόταση με την καλύτερη και πιο λογική απόδοση. Η γνώση που αποκτάται μέσω της διαδικασίας αποσαφήνισης πηγάζει από την ύπαρξη ενός οργανωμένου σημασιολογικού δικτύου [8].

1.6 Σημασιολογικά Μοντέλα (Semantic Models)

Η σημασιολογία γενικά ασχολείται με την έννοια των πραγμάτων. Η έννοια δομείται μέσω ενός σημασιολογικού μοντέλου, που αναπαριστά τις σχέσεις και τις σημασίες μεταξύ διαφόρων στοιχείων. Τα σημασιολογικά μοντέλα που έχουν προταθεί και χρησιμοποιούνται στην σημασιολογική έρευνα είναι το **Γλωσσολογικό (Linguistic)** και το **Εννοιολογικό Μοντέλο (Conceptual Model)** [9]. Το γλωσσολογικό μοντέλο αναπαριστά τη σχέση που υπάρχει μεταξύ των όρων π.χ. συνωνυμία, παράγωγα κτλ, δηλαδή ασχολείται με την σημασία σε επίπεδο λέξεων και το εννοιολογικό περιγράφει τις σχέσεις που υπάρχουν μεταξύ των οντοτήτων, δηλαδή ασχολείται με την σημασία στο επίπεδο των οντοτήτων του πραγματικού κόσμου που υποδηλώνουν οι λέξεις. Παράδειγμα γλωσσολογικού μοντέλου είναι οι θησαυροί ενώ του εννοιολογικού τα διαγράμματα οντοτήτων συσχετίσεων.

Τα μοντέλα αυτά λειτουργούν σε ανθρώπινο εννοιολογικό επίπεδο και παράλληλα παρέχουν ορισμούς των ίδιων εννοιών χρήσιμων και σε υπολογιστές. Η δόμηση της

γνώσης γίνεται με τρόπο ώστε να παρέχεται μια κοινή γλώσσα που επιτρέπει πιο αποδοτική επικοινωνία και επίλυση προβλημάτων.

1.7 Είδη Αναζήτησης

Η αναζήτηση πληροφορίας κατατάσσεται σε δύο είδη, στην πλοηγική αναζήτηση και στην αναζήτηση διερεύνησης[10], [11].

Κατά την **Πλοηγική Αναζήτηση (Navigational Search)**, ο χρήστης εισάγει λέξεις-κλειδιά με σκοπό η μηχανή αναζήτησης να εμφανίσει αποτελέσματα με όλα τα έγγραφα που περιέχουν τις λέξεις αυτές και εν συνεχεία, μέσω των αποτελεσμάτων να μεταφερθεί στον ιστότοπο που επιθυμεί. Η προσέγγιση της μηχανής ανάκτησης πληροφορίας στην πλοηγική αναζήτηση, είναι το ακριβές ταίριασμα όλων ή κάποιων λέξεων του ερωτήματος του χρήστη ή απλά το ταίριασμα του προθέματος της λέξης με το περιεχόμενο των ιστοσελίδων. Τα παραγόμενα αποτελέσματα προέρχονται από διάφορες περιοχές και είναι διαφόρων μορφών (αρχεία, βίντεο, εικόνες, ιστοσελίδες). Στην πλοηγική αναζήτηση, ο χρήστης δεν αναζητά την απόδοση της έννοιας κάποιου όρου αλλά ένα μέσο πρόσβασης σε συγκεκριμένες πηγές πληροφορίας. Το είδος αυτό της αναζήτησης, δεν αποτελεί αντικείμενο της σημασιολογικής έρευνας αλλά συντελείται από τις παραδοσιακές μηχανές αναζήτησης σε άριστο βαθμό.

Κατά την **Αναζήτηση Διερεύνησης (Research search)**, ο χρήστης θέτει ένα ερώτημα στη μηχανή αναζήτησης, με σκοπό να ενημερωθεί και να συλλέξει πληροφορίες για κάποιο αντικείμενο του ενδιαφέροντός του. Μέσα από την λίστα αποτελεσμάτων που του παρέχεται, δεν αναζητά κάποιο συγκεκριμένο έγγραφο αλλά διερευνά τα αποτελέσματα για τον εντοπισμό των πιο σχετικών. Η αναζήτηση του χρήστη μπορεί να απαιτεί την επίσκεψη διαφορετικών ιστοσελίδων για την συγκέντρωση στοιχείων από διαφορετικές πηγές ενημέρωσης, ώστε να αποκτήσει μία πληρέστερη αντίληψη του θέματος.

Η αναζήτηση διερεύνησης είναι το είδος αναζήτησης που εστιάζει και επικεντρώνει τις προσπάθειες της η σημασιολογική έρευνα, για την διευκόλυνση του χρήστη στην προσέγγιση της γνώσης.

1.8 Σημαντικοί Παράγοντες της Σημασιολογικής Αναζήτησης

Η σημασιολογική αναζήτηση δεν επικεντρώνεται μονάχα στην εύρεση του νοήματος ενός ερωτήματος και στην πρόθεση του χρήστη ή στην απόδοση ακριβούς και σχετικής πληροφορίας. Μια έξυπνη μηχανή αναζήτησης, θα πρέπει να λαμβάνει υπόψη της αρκετούς παράγοντες, για να παρέχει πιο σχετική και χρήσιμη πληροφορία[12]. Οι παράγοντες αυτοί περιλαμβάνουν:

- Τρέχουσα επικαιρότητα. Τα συστήματα σημασιολογικής αναζήτησης, θα πρέπει να είναι σε θέση να αντιλαμβάνονται ποια ερωτήματα αναφέρονται στην τρέχουσα επικαιρότητα και να επιστρέφουν αποτελέσματα ανάλογα με τις τρέχουσες εξελίξεις. Π.χ. την επομένη της τελετής των Oscar, στο αίτημα του χρήστη "Καλύτερη ταινία της χρονιάς", η σημασιολογική μηχανή θα πρέπει να εμφανίσει την άμεση απάντηση με την καλύτερη ταινία του συγκεκριμένου έτους.
- Τοποθεσία. Στα ερωτήματα που αφορούν τοπικές πληροφορίες (π.χ. καιρός, ώρα) η μηχανή αναζήτησης, θα πρέπει να παρέχει αποτελέσματα βασισμένα στην τοποθεσία του χρήστη. Π.χ. "Τι θερμοκρασία έχει;", η άμεση απάντηση που θα λάβει ο χρήστης θα πρέπει να αφορά τον τόπο κατοικίας του.
- Παραλλαγές όρων. Τα αποτελέσματα που παίρνουν οι χρήστες θα πρέπει να περιέχουν τους όρους αλλά και όλες τις σημασιολογικές παραλλαγές τους (πτώσεις, χρόνοι).
- Συνώνυμα Λέξεων. Η σημασιολογική μηχανή αναζήτησης θα πρέπει να κατανοεί τα συνώνυμα λέξεων και στην αναζήτηση ενός όρου να παρέχονται και τα αποτελέσματα και της συνώνυμης λέξης.
- Γενικευμένα ερωτήματα. Επιπλέον η σημασιολογική μηχανή αναζήτησης, θα πρέπει να είναι ικανή να θέτει συσχετισμούς μεταξύ γενικευμένων και ειδικευμένων ερωτημάτων και να παρέχει κατάλληλα και σχετικά αποτελέσματα. Π.χ. Στο ερώτημα "Εγκυμοσύνη", οι απαντήσεις θα πρέπει να περιλαμβάνουν και σελίδες με τον όρο "Τοκετό".
- Συσχέτιση εννοιών. Θα πρέπει επίσης, να αντιλαμβάνεται την ευρεία έννοια του ερωτήματος και να επιστρέφει σχετικά αποτελέσματα. Π.χ. Στο ερώτημα "Εφημερεύοντα Νοσοκομεία", θα πρέπει να επιστρέφονται αποτελέσματα και για εφημερεύοντα Φαρμακεία.

- Ερωτήματα σε Φυσική Γλώσσα. Λόγω της συχνής χρήσης της φυσικής γλώσσας κατά τον σχηματισμό ερωτημάτων από τους χρήστες, η σημασιολογική μηχανή αναζήτησης θα πρέπει να κατανοεί και να επιστρέφει την κατάλληλη απάντηση σε αυτά τα ερωτήματα.
- Απόδοση εννοιών μέσα από μια ομάδα λέξεων. Συνδυάζοντας διαφορετικές λέξεις που περιλαμβάνονται σε ένα ερώτημα, η πραγματική του έννοια μπορεί να διαφοροποιηθεί. Η σημασιολογική μηχανή αναζήτησης, θα πρέπει να είναι ικανή να διακρίνει τέτοιες διαφοροποιήσεις. Π.χ. Το ερώτημα "Windows paint", έχει αμφίσημη σημασία καθώς μπορεί να αναφέρεται στο βάψιμο ενός παραθύρου ή στο πρόγραμμα "Paint"των Windows.

1.9 Σύγκριση Παραδοσιακής και σημασιολογικής αναζήτησης

Η αναζήτηση και η εύρεση πληροφορίας είναι μια διαδικασία που ενίοτε απαιτεί εκτενή και εντατική προσπάθεια. Στόχος της σημασιολογικής έρευνας είναι να παράγει πιο άμεσα και συνεπή αποτελέσματα από αυτά της παραδοσιακής αναζήτησης, εστιάζοντας στην σημασία των όρων και όχι απλά στον εντοπισμό τους. Για το λόγο αυτό, οι δύο αυτοί μηχανισμοί αναζήτησης έχουν διαφορετικό τρόπο προσέγγισης στην ανεύρεση αποτελεσμάτων [2], [10], [13].

Στην σημασιολογική έρευνα, οι μηχανές αναζήτησης δεν εστιάζουν στα κείμενα αλλά στην αναζήτηση οντοτήτων και τα δεδομένα που υπάρχουν στο σημασιολογικό δίκτυο είναι ευέλικτα καταχωρημένα, αναπαριστώντας δεδομένα των οντοτήτων του πραγματικού κόσμου. Ο τρόπος αυτός αναπαράστασης των δεδομένων, κάνει εφικτή την αντίληψη της έννοιάς τους από τα υπολογιστικά συστήματα, χωρίς να παρακάμπτει την αμφισημία τους ή την ύπαρξη συνώνυμών τους. Η υποβολή των ερωτημάτων υποστηρίζεται από τη φυσική γλώσσα, ενώ οι απαντήσεις που παράγονται είναι ακριβείς και εμπεριστατωμένες και προέρχονται μετά από την εννοιολογική επεξεργασία των όρων του ερωτήματος ή συνάγονται από υπάρχοντα δεδομένα μέσω της λογικής.

Οι σημασιολογικές μηχανές αναζήτησης, έχοντας πρόσβαση σε δομημένα δεδομένα, παρέχουν στον χρήστη σύνθετες πληροφορίες που προέρχονται από πολλαπλά κείμενα, δίνοντας του παράλληλα τη δυνατότητα διασταύρωσης των πηγών προέλευσής των αποτελεσμάτων. Συγκριτικά με μία παραδοσιακή μηχανή αναζήτησης, η αξιοπιστία των αποτελεσμάτων είναι πιο ισχυρή καθώς στην

παραδοσιακή αναζήτηση παρέχονται και αποτελέσματα που στηρίζονται σε λιγότερο έγκυρες πηγές πληροφόρησης (ιστολόγια, χώροι συζήτησης).

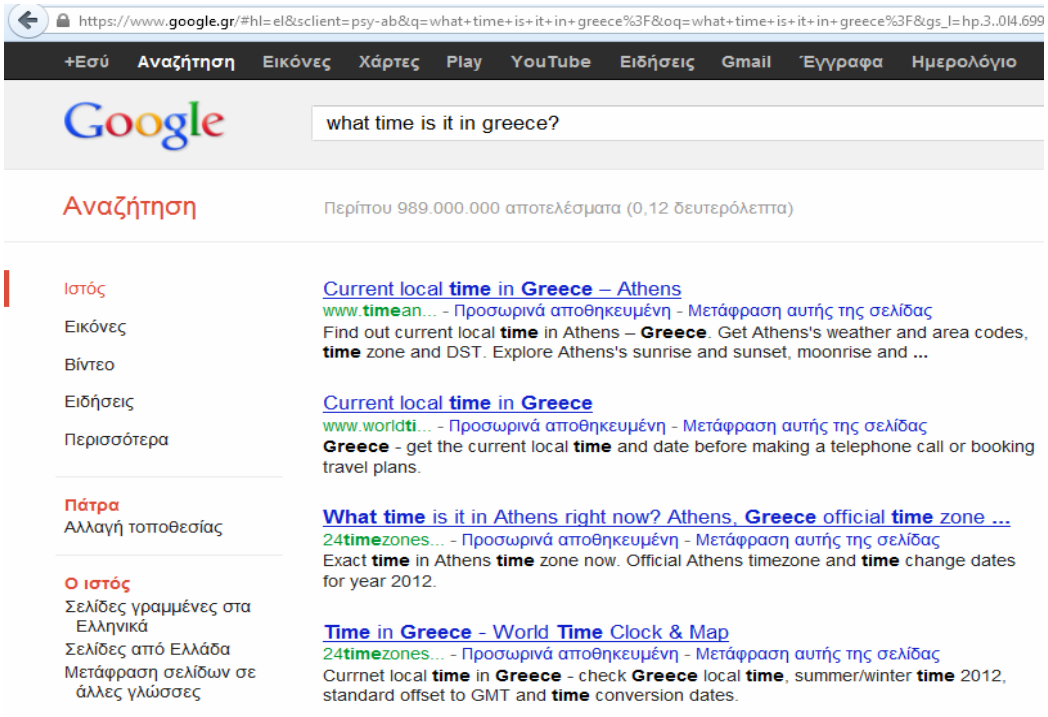
Όλες αυτές οι λειτουργίες στοχεύουν στην διευκόλυνση του χρήστη και την ελάττωση της προσπάθειας που απαιτείται για την ανάκτηση πληροφοριών, συνεπάγονται όμως μεγάλες ερευνητικές προκλήσεις της τεχνολογίας ανάκτησης πληροφορίας, όσον αφορά το σχεδιασμό του συστήματος, την αρχιτεκτονική, τους αλγορίθμους, την υλοποίηση και τη διεπαφή χρήστη.

Συνοψίζοντας, οι δύο αυτές τεχνολογίες αναζήτησης, επιτελούν τον ίδιο σκοπό με διαφορετικές όμως προσεγγίσεις και η αποτελεσματικότητα τους έγκειται στο είδος της πληροφορίας που γίνεται αντικείμενο αναζήτησης. Η παραδοσιακή αναζήτηση αποδίδει πολύ καλά κατά την πλοηγική αναζήτηση, όταν π.χ. ο χρήστης αναζητά συγκεκριμένα θέματα όπως ένα αρχείο, κείμενο, άλμπουμ, βίντεο, ιστοσελίδα κτλ. Εντούτοις, όταν πρόκειται για πληροφορίες που αφορούν ημερομηνία, ώρα, τόπο, πρόσωπα, γεγονότα κ.τ.λ., ή απαιτούν συνδυασμό και συσχέτιση γνώσης, η σημασιολογική αναζήτηση είναι καταλληλότερη και αποδίδει στο έπακρο.

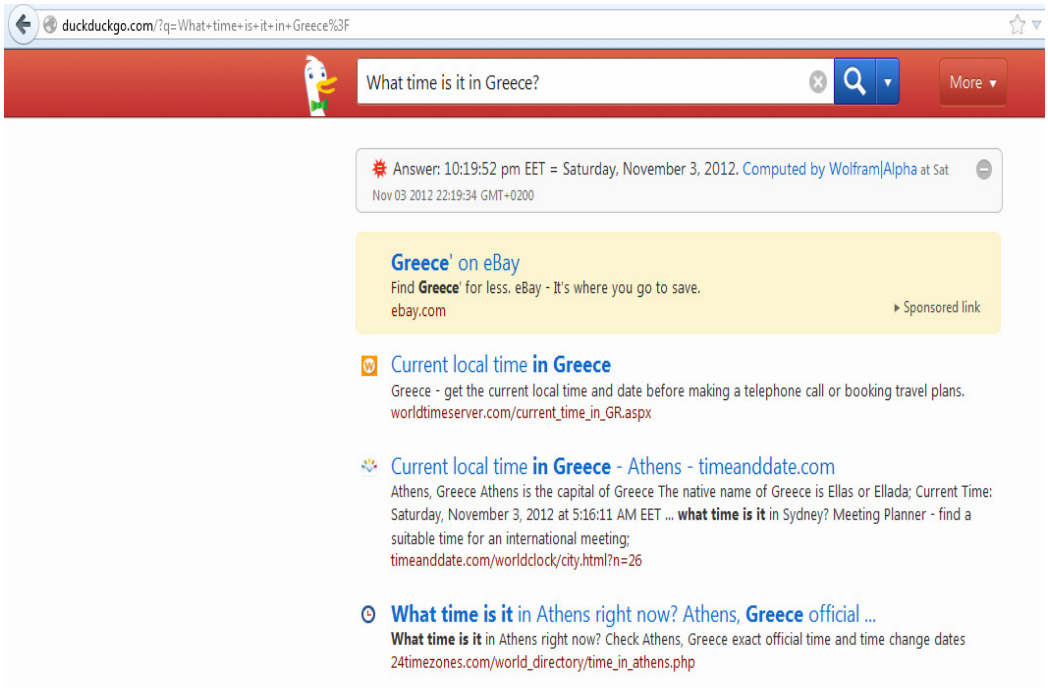
1.10 Παράδειγμα

Στη συνέχεια βλέπουμε ένα παράδειγμα για το πως διαχειρίζονται το ίδιο ερώτημα μία παραδοσιακή και μία σημασιολογική μηχανή αναζήτησης. Το ερώτημα «Τι ώρα είναι στην Ελλάδα;» είναι διατυπωμένο στη φυσική γλώσσα και η απάντηση που το ικανοποιεί είναι συγκεκριμένη. Στη μηχανή αναζήτησης της Google, τα αποτελέσματα που δίνονται αφορούν τους όρους του ερωτήματος δηλαδή «ώρα» και «Ελλάδα». Ο χρήστης για να ικανοποιήσει το αίτημα του για πληροφόρηση, θα πρέπει να περιηγηθεί στις σελίδες με τους υπερσυνδέσμους που παραθέτει η μηχανή. Στη σημασιολογική μηχανή αναζήτησης Duck Duck Go, τα αποτελέσματα όπως παρατηρούμε και στην εικόνα είναι διαφορετικά. Στο πάνω μέρος των αποτελεσμάτων βρίσκεται η απάντηση που καλύπτει άμεσα το ερώτημα του χρήστη, δηλαδή την τρέχουσα ημερομηνία, όπως αυτή θα αντιλαμβανόταν και θα δινόταν από τον άνθρωπο. Στη συνέχεια, ομοίως με την Google, η σημασιολογική μηχανή αναζήτησης παραθέτει αποτελέσματα βασισμένα στις λέξεις-κλειδιά του ερωτήματος. Θα πρέπει να σημειωθεί ότι η μηχανή αναζήτησης Google, έχει αρχίσει το τελευταίο διάστημα να υιοθετεί και να ενσωματώνει σημασιολογικές προσεγγίσεις στην τεχνολογία της, με σκοπό να προσφέρει ευρύτερες υπηρεσίες, απαντώντας σε

πολύπλοκα ερωτήματα και να εξακολουθήσει να είναι η δημοφιλέστερη μηχανή αναζήτησης του διαδικτύου.



Εικόνα 1.9.1 - Αποτελέσματα αναζήτησης της Google στο ερώτημα "What time is it in Greece" [14]



Εικόνα 1.9.2 - Αποτελέσματα αναζήτησης της σημασιολογικής μηχανής Duck Duck Go[15]

2. Σημασιολογικό Διαδίκτυο (Semantic Web) και Σημασιολογικές Μηχανές Αναζήτησης

2.1 Πρόλογος

Οι σημασιολογικές μηχανές αναζήτησης εφαρμόζονται στο Σημασιολογικό Διαδίκτυο που αποτελεί προέκταση του Παγκόσμιου Ιστού και αναπτύσσεται για να προσφέρει εξελιγμένες δυνατότητες. Το σημασιολογικό διαδίκτυο αποτελείται από τα μετα-δεδομένα, δεδομένα δηλαδή που αφορούν τα δεδομένα και τις οντολογίες που προσδιορίζουν τον τρόπο που αποδίδεται η σημασιολογία στους πόρους του ιστού. Στο κεφάλαιο αυτό θα περιγραφούν οι όροι που απαρτίζουν, συσχετίζονται και συνεπικουρούν στην υπόστασή του.

2.2 Σημασιολογικές Μηχανές και Οντολογίες

Η σημασιολογική αναζήτηση για να επιτελέσει τον σκοπό της, δηλαδή την εύρεση της πραγματικής πρόθεσης του ερωτήματος του χρήστη, ενσωματώνει στην τεχνολογία της υλοποιήσεις Τεχνητής Νοημοσύνης (Artificial Intelligence). Η επιστήμη της Τεχνητής Νοημοσύνης, είναι θεμελιώδους σημασίας στην ανάκτηση πληροφορίας και στην δημιουργία έξυπνων συστημάτων, καθώς προσπαθεί να επιτύχει την κατάκτηση γνώσης μέσω της μίμησης της λειτουργίας του ανθρώπινου εγκεφάλου. Η αντίληψη της πληροφορίας από τον άνθρωπο γίνεται μετατρέποντας τις λέξεις σε οντότητες και κατανοώντας τις σχέσεις που υπάρχουν μεταξύ των ιδιοτήτων τους. Η Τεχνητή Νοημοσύνη προσομοιώνει την λειτουργία αυτή μέσω των οντολογιών. Με τον όρο οντολογία εννοούμε ένα σύνολο δομημένων και ιεραρχημένων όρων που περιγράφουν ένα πεδίο ενδιαφέροντος το οποίο μπορεί να χρησιμοποιηθεί ως θεμέλιο σε μία βάση γνώσης. Η οντολογία δηλαδή περιγράφει πράγματα και έννοιες και ορίζει τις σημασιολογικές σχέσεις που υπάρχουν μεταξύ τους και ουσιαστικά είναι μια δήλωση μιας λογικής θεωρίας. Οι οντολογίες υποστηρίζουν πολλές εργασίες όπως κατάκτηση, επεξεργασία, επαναχρησιμοποίηση και μετάδοση γνώσης. Μία ακόμα βασική τους λειτουργία είναι η ταξινόμηση, η κατηγοριοποίηση δηλαδή των εννοιών με ιεραρχική μορφή. Συνήθως αναπαριστάται με την μορφή δέντρου, ώστε να εκφράσει μια σχέση υπαγωγής μεταξύ των εννοιών.

Όλη αυτή η πληροφορία που παράγεται μέσω των οντολογιών χρησιμοποιείται από τους ευφυείς πράκτορες (intelligent agents), τις αυτόνομες εκείνες οντότητες που εκτελούν ένα σύνολο λειτουργιών για την εξόρυξη και την ανταλλαγή γνώσης, με σκοπό να εξυπηρετηθούν οι ανάγκες των χρηστών.

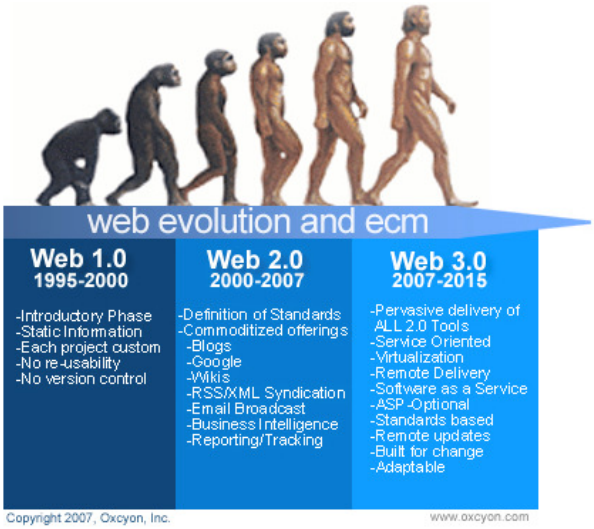
2.3 Τι είναι το Σημασιολογικό Διαδίκτυο (Semantic Web)

Οι οντολογίες, όπως προαναφέρθηκε, είναι ένα βασικό στοιχείο των σημασιολογικών μηχανών αναζήτησης και μέσω αυτών είναι εφικτή η αναπαράσταση και η ανάκτηση της πληροφορίας. Εντούτοις, οι σημασιολογικές μηχανές αναζήτησης για να υλοποιηθούν χρειάζονται έναν ιστό ίδιας φιλοσοφίας και τεχνολογίας, που υποστηρίζει τη σημασιολογική συσχέτιση των όρων. Συνεπώς, δεν μπορούν να εφαρμοσθούν στον παγκόσμιο ιστό με την προϋπάρχουσα δομή και σύστασή του. Το νέο διαδίκτυο που δημιουργήθηκε και καλείται Σημασιολογικό (Semantic Web), προέκυψε από αυτή ακριβώς την ανάγκη, να μετατραπεί ο Παγκόσμιος ιστός σε ένα διαδίκτυο όπου οι υπολογιστές μπορούν να επεξεργαστούν την πληροφορία, να την ερμηνεύσουν και να την συσχετίσουν, ώστε οι άνθρωποι να αποκτήσουν την απαιτούμενη γνώση. Το σημασιολογικό διαδίκτυο είναι ένα τεράστιο κατανεμημένο σύστημα ιστοσελίδων στο οποίο η πληροφορία έχει δομή, οργάνωση και σημασιολογία και υποστηρίζεται η αποδοτική αναζήτηση, επεξεργασία και ενοποίηση των δεδομένων. Η ιδέα για την ύπαρξη του σημασιολογικού διαδικτύου ξεκίνησε από τον δημιουργό του παγκόσμιου ιστού Tim Berners-Lee και η υλοποίηση του τελείται υπό την καθοδήγηση της Κοινοπραξίας του Παγκόσμιου Ιστού -W3C (World Wide Web Consortium) που υποστηρίζει τις τεχνολογίες που απαιτούνται για την ανάπτυξη του[16], [17].

2.3.1 Το Διαδίκτυο πριν τον Σημασιολογικό Ιστό

Το σημασιολογικό διαδίκτυο ή αλλιώς Web 3.0, είναι το μεταγενέστερο διαδίκτυο των δύο προϋπαρχόντων γενιών διαδικτύου, του Web 1.0 και του Web 2.0 [18]. Στο Web 1.0, το διαδίκτυο βρισκόταν σε ένα αναδυόμενο στάδιο και παρείχε έναν μονόδρομο τρόπο επικοινωνίας. Με την πλειοψηφία των ιστοσελίδων να αφορούν οργανισμούς και επιχειρήσεις, οι παροχές προς τους χρήστες περιορίζονταν στη δυνατότητα χρησιμοποίησης του διαδικτύου για την επιτέλεση διάφορων εργασιών

όπως ενημέρωση, αγορά προϊόντων ή υπηρεσιών, προβολή εικόνων και βίντεο, χρήση λεξικών κ.α.



Το Web 2.0 είναι το σημερινό μοντέλο του διαδικτύου, που δίνει επιπλέον δυνατότητες στους χρήστες για να αλληλεπιδρούν, να επικοινωνούν, να συνεργάζονται, να ανταλλάσσουν αρχεία και πληροφορίες. Εκτός από τις επιχειρήσεις, πολλοί ιδιώτες δημοσιεύουν πληροφορίες και μπορούν να παρέμβουν στο περιεχόμενο των ιστοσελίδων. Σε αυτή τη γενιά του διαδικτύου, οι δυνατότητες που παρέχονται σε απλούς χρήστες είναι απεριόριστες. Έτσι λοιπόν χρήστες χωρίς εξειδικευμένες γνώσεις μπορούν, για παράδειγμα, να εκθέσουν και να ανταλλάξουν τις απόψεις τους μέσω ιστολογιών (blog), να μοιραστούν μουσικά αρχεία ή ταινίες μέσω ομότιμων δικτύων (peer-to peer network) όπως e-Mule, Napster, ή να επικοινωνήσουν μέσω κοινωνικών δικτύων όπως Facebook, Twitter κ.α.

2.3.2 Σημασιολογικός Ιστός και Συνδεδεμένα Δεδομένα – (Linked Data)

Η μετάβαση από το Web 2.0 στο σημασιολογικό ιστό, προϋποθέτει την αλλαγή της διαχείρισης του περιεχομένου του διαδικτύου και της εφαρμογής ενός νέου τρόπου οργάνωσής του. Το διαδίκτυο δηλαδή θα πρέπει να μετατραπεί από ένα διαδίκτυο κειμένων σε ένα διαδίκτυο διασυνδεόμενων δεδομένων. Ο λόγος είναι ότι το σημασιολογικό διαδίκτυο δεν χρειάζεται απλώς πρόσβαση στα δεδομένα άλλα χρειάζεται να γνωρίζει και τις σημασιολογικές σχέσεις που τα διέπουν, ώστε να ερμηνεύονται και να χρησιμοποιούνται από τους ανθρώπους και τους υπολογιστές. Συνεπώς, τα διασυνδεδεμένα στοιχεία δεν θα είναι πια κείμενα, αρχεία ή κάποιο άλλο είδος πολυμεσικού περιεχομένου, αλλά τα ίδια τα δεδομένα μεταξύ τους. Βασική προϋπόθεση για την επίτευξη του σκοπού αυτού, είναι η ύπαρξη μιας βάσης δεδομένων ή αλλιώς ενός συνόλου δεδομένων για την συλλογή των δεδομένων και την οργάνωση τους με τέτοιο τρόπο ώστε να έχουν περιεχόμενο, συνάφεια και σκοπό. Το σύνολο των πρακτικών που εφαρμόζονται για την δημοσίευση και την διασύνδεση των δεδομένων είναι γνωστό ως Συνδεδεμένα Δεδομένα (Linked Data).

Σύμφωνα με τον Tim Berners-Lee, τα συνδεδεμένα δεδομένα θα πρέπει να ακολουθούν βασικές αρχές[19]:

1. Χρήση URI's για την ονομασία των πόρων του διαδικτύου.

   Κάθε δεδομένο ή πόρος του διαδικτύου έχει ένα μοναδικό αναγνωριστικό που το διακρίνει από τα άλλα δεδομένα που είναι δημοσιευμένα στο διαδίκτυο. Το αναγνωριστικό αυτό είναι το URI- Uniform Resource Identifier και μέσω αυτού θα ονομάζονται οι πόροι του διαδικτύου.

2. Χρήση HTTP URI's ώστε οι χρήστες να αναζητούν αυτές τις ονομασίες των πόρων μέσω του πρωτοκόλλου HTTP.

   Τα URI's θα πρέπει να παραπέμπουν στο σημείο του διαδικτύου που θα έχει πρόσβαση ο χρήστης μέσω του πρωτοκόλλου μεταφοράς HTTP. Όταν δηλαδή ο χρήστης ανατρέξει σε έναν πόρο, το πρωτόκολλο HTTP χρησιμοποιεί το URI για να αποκτήσει πρόσβαση στο περιεχόμενο που οδηγεί ο πόρος αυτός.

3. Κατά την αναζήτηση ενός URI, οι πληροφορίες παρέχονται με τη χρήση των προτύπων (RDF και SPARQL).

   Η σημασιολογία των δεδομένων και η σχέση τους εκφράζεται με τη χρήση αυτών των προτύπων.

4. Χρήση συνδέσμων προς άλλα URIs, έτσι ώστε να μπορούν να ανακαλυφθούν περισσότερες πληροφορίες.

   Για να αξιοποιηθεί πλήρως το δυναμικό του διαδικτύου, τα δεδομένα μιας βάσης δεδομένων θα πρέπει να συνδέονται με άλλα εξωτερικά δεδομένα μιας άλλης βάσης δεδομένων.

2.3.3 Συνδεδεμένα Ανοικτά Δεδομένα (Linked Open Data)

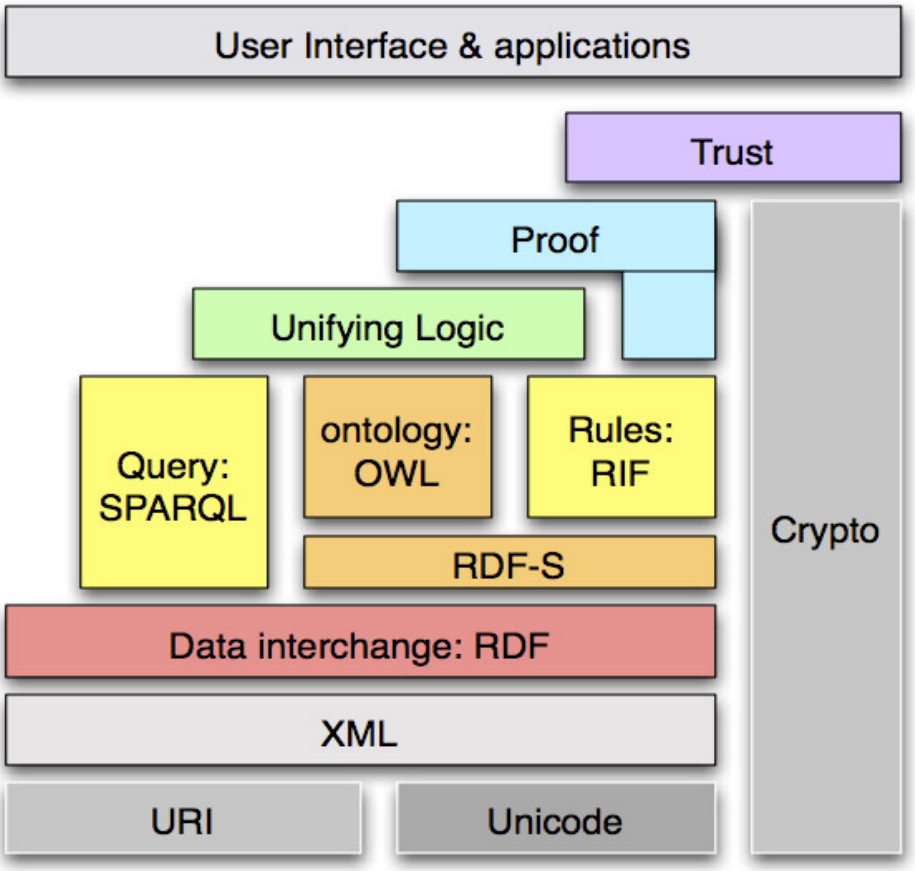
Τα Συνδεδεμένα Ανοικτά Δεδομένα (Linked Open Data), ακολουθούν την πρακτική των συνδεδεμένων δεδομένων (Linked Data) βελτιώνοντας την δυναμική του διαδικτύου. Τα ανοικτά δεδομένα βασίζονται σε μεγάλο βαθμό στην συμμετοχή αναρίθμητων ιδιωτών και επιχειρήσεων, που έχουν την ελευθερία να επαναχρησιμοποιούν, εμπλουτίζουν, μοιράζονται και προσπελαύνουν τα δημοσιευμένα δεδομένα, χωρίς τον περιορισμό πνευματικών δικαιωμάτων, ευρυσιτεχνιών ή άλλων μηχανισμών ελέγχου. Η συγκεκριμένη πρωτοβουλία, που ξεκίνησε το 2007, απέκτησε σύντομα μεγάλη απήχηση και πολλοί μεγάλοι οργανισμοί όπως το BBC και η βιβλιοθήκη του Κογκρέσου συμμετείχαν

προσθέτοντας πληροφορίες στο διαδίκτυο. Ένα τυπικό παράδειγμα συνδεδεμένων Ανοικτών δεδομένων είναι η DBpedia, που ουσιαστικά περιλαμβάνει δεδομένα της Wikipedia καθώς και συνδέσμους για άλλα σύνολα δεδομένων π.χ. Geonames.

Το διαδίκτυο δεδομένων αποτελεί μία δυναμική οντότητα, καθώς νέα δεδομένα και σχέσεις με σημασιολογική έννοια προστίθενται συνεχώς. Ωστόσο, η αύξηση των όρων, η επέκταση ή η διαφοροποίηση των ορισμών και η ελεύθερη δημοσίευση και συσχέτιση δεδομένων είναι παράγοντες που επισείουν κινδύνους για την αξιοπιστία και την ποιότητα του περιεχομένου του διαδικτύου και θα πρέπει να ληφθούν υπόψη [20].

2.4 Αρχιτεκτονική Σημασιολογικού Διαδικτύου

Η αρχιτεκτονική του σημασιολογικού διαδικτύου που σχεδίασε η κοινοπραξία W3C, απαρτίζεται από διάφορα επίπεδα που απεικονίζονται στο σχήμα και αναπτύσσονται στη συνέχεια [21], [22], [23].



Εικόνα 2.4 - Βασική Αρχιτεκτονική Σημασιολογικού Ιστού

2.4.1 URI (Uniform Resource Identifier)

Η τεχνολογία URI – Uniform Resource Identifier, εφαρμόζεται στο φυσικό επίπεδο της αρχιτεκτονικής του σημασιολογικού ιστού, όμοια με το φυσικό επίπεδο της αρχιτεκτονικής του παγκόσμιου ιστού. Το πρότυπο URI είναι υπεύθυνο για την διευθυνσιοδότηση των εγγράφων στο διαδίκτυο. Τα δύο υποσύνολα του URI είναι το URL – Uniform Resource Locator, που περιλαμβάνει τους μηχανισμούς πρόσβασης και την τοποθεσία ενός κειμένου και το υποσύνολο URN – Uniform Resource Name που επιτρέπει την αναγνώριση των πόρων.

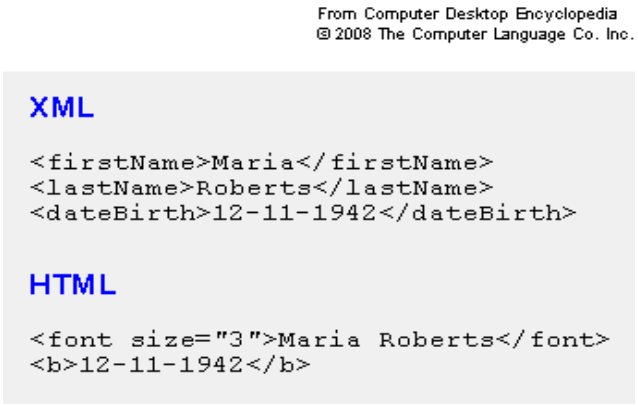
2.4.2 Unicode

Το πρότυπο Unicode είναι ένα διεθνές πρότυπο, που παρέχει ένα κωδικοποιημένο σύνολο χαρακτήρων και χρησιμοποιείται για την αναπαράσταση, κωδικοποίηση και διαχείριση των γλωσσών όλων των συστημάτων γραφής, με τελικό σκοπό την χρήση του από τα υπολογιστικά συστήματα. Το πρότυπο Unicode δημοσιεύτηκε για πρώτη φορά το 1991 και γρήγορα καθιερώθηκε ως το πληρέστερο πρότυπο κωδικοποίησης. Η τελευταία του έκδοση συμπεριλαμβάνει περισσότερους από 110.000 χαρακτήρες, επιστημονικά σύμβολα, σύμβολα μουσικής, κανόνες, ιδιότητες χαρακτήρων κ.α.
Στο σημασιολογικό διαδίκτυο, όπως και στον παγκόσμιο ιστό, το πρότυπο αυτό εφαρμόζεται στο φυσικό επίπεδο [24].

2.4.3 XML

Η γλώσσα XML – Extensible Markup Language, είναι μια γλώσσα σήμανσης που σχεδιάστηκε για την δόμηση, αποθήκευση και διανομή δεδομένων. Η απλοποιημένη της μορφή την καθιστά ένα εργαλείο ανεξάρτητο του υλικού και του λογισμικού μέρους και της επιτρέπει την μετάδοση πληροφορίας μεταξύ ασύμβατων συστημάτων και διαφορετικών εφαρμογών.
Η XML διαφέρει από την HTML που είναι και αυτή μία γλώσσα σήμανσης, καθώς παρέχει περισσότερες δυνατότητες. Για την σύνταξη της χρησιμοποιεί ετικέτες (tags), που εμφωλιάζουν άλλες ετικέτες, όπως χρησιμοποιεί και η HTML. Αν και η μορφή των δύο γλωσσών έχει κατανοητό περιεχόμενο από τους ανθρώπους, η γλώσσα XML

καταφέρνει να γίνεται αντιληπτή και από τις μηχανές, καθώς δεν περιγράφει απλά την παρουσίαση μιας ιστοσελίδας όπως η HTML, αλλά το περιεχόμενο της.
Στην εικόνα 2.4, βλέπουμε ένα παράδειγμα σύνταξης των δύο γλωσσών. Η XML περιγράφει τη λογική και φυσική δομή των δεδομένων, ενώ η HTML περιορίζεται στην παρουσίαση και τον τύπο γραφής των δεδομένων.



Εικόνα 2.4. 3 - Σύγκριση γλωσσών XMLκαι HTML [25]

Βασικά στοιχεία της γλώσσας είναι τα Namespaces και το σχήμα XML. Η ονοματολογία (Namespaces) χρησιμοποιείται για να αποφευχθεί η σύγχυση κατά την χρήση κοινών ονομάτων στις ετικέτες των στοιχείων και το σχήμα XML (XML-schema) χρησιμοποιείται για να ορίσει την δομή ενός XML εγγράφου.

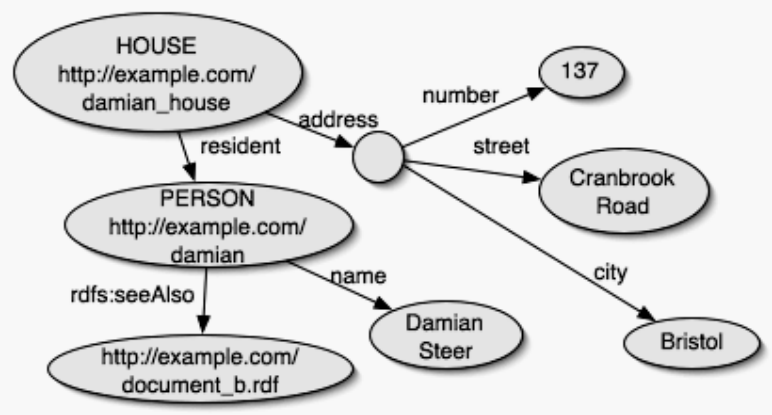
2.4.4 Μοντέλο RDF (RDF Model) και RDFS (RDF Schema)

Το Πλαίσιο Περιγραφής Πόρων – RDF (Resource Description Framework) είναι ένα πολύ χρήσιμο εργαλείο για την αναπαράσταση δηλώσεων που αφορούν μεταδεδομένα. Η φιλοσοφία του πλαισίου επιτρέπει την σταδιακή οικοδόμηση της γνώσης, τον διαμοιρασμό της και την επαναχρησιμοποίησή της. Το RDF έχει σχεδιαστεί για να είναι κατανοητό από τους υπολογιστές και η σύνταξη του έχει βασιστεί στην XML που καλείται RDF/XML γλώσσα. Βασικά στοιχεία του RDF είναι το RDF Model και RDF Schema.
Το μοντέλο RDF – RDF Model, είναι ένα γράφημα που κόμβοι του είναι οι όροι των δεδομένων και ακμές οι σχέσεις μεταξύ των όρων. Οι κόμβοι και οι σχέσεις, αποτελούν τους πόρους του πλαισίου και προσδιορίζονται από ένα δικό τους URI που τους προσδίδει μία διαφορετική ιδιότητα. Κάθε ακμή του γραφήματος συνθέτεται από

τρία βασικά στοιχεία του RDF πλαισίου, τον κόμβο προέλευσης που είναι το υποκείμενο, την ιδιότητα που αποδίδουμε στο υποκείμενο και τον κόμβο προορισμού που είναι το αντικείμενο. Το τρίπτυχο αυτό υποκείμενο – ιδιότητα - αντικείμενο αναπαριστά μία δήλωση (statement) για την σχέση των οντοτήτων που συνδέονται στους κόμβους. Ένας γράφος συνθέτεται από πολλές αλληλουχίες τριάδων, μία για κάθε ακμή. Στην κάθε τριάδα ορίζεται ένα διαφορετικό αναγνωριστικό. Οι δηλώσεις μπορούν να αντιστοιχήσουν δεδομένα διαφορετικού τύπου, π.χ. ένας πίνακας με σχέσεις μπορεί να μετατραπεί σε ένα σύνολο από τριάδες.

Στη συνέχεια παραθέτεται ένα παράδειγμα ενός RDF γράφου και η σύνταξη του (RDF Syntax) στην RDF/XML [26], [28].

```
<House rdf:resource="http://example.com/damian_house">
    <address parseType="resource">
        <number>137</number>
        <street>Cranbook Road</street>
        <city>Bristol</city>
    </address>
    <resident>
        <Person rdf:resource="http://example.com/damian">
            <name>Damian Steer</name>
            <mailbox rdf:resource="mailto:damian@example.com"/>
            <rdfs:seeAlso
rdf:resource="http://example.com/document_b.rdf"/>
        </Person>
    </resident>
</House>
```



Εικόνα 2.3. 1 - RDF Model

Μία δήλωση που εξάγεται από το παραπάνω παράδειγμα είναι: "Στο σπίτι κατοικεί ένα άτομο" και αποτελείται από το εξής τρίπτυχο:

Υποκείμενο: http://example.com/damian_house

Ιδιότητα: Resident

Αντικείμενο: http://example.com/damian

Το σχήμα RDF –RDF Schema (RDFS) είναι μία επέκταση του RDF πλαισίου. Το πλαίσιο RDF χρησιμοποιείται για να περιγράψει ιδιότητες και τιμές ενώ το RDFS για να περιγράψει και να ταξινομήσει κλάσεις και ιδιότητες συγκεκριμένων εφαρμογών. Το RDFS έχει παρόμοια δομή με αυτή του Αντικειμενοστραφή Προγραμματισμού και κάποια από τα στοιχεία που περιέχει είναι κλάσεις, τύποι, υποκλάσεις, ιδιότητες, στοιχεία.

RDFa

Το RDFa (Resource Description Framework in Attributes) [39] είναι ένα πρότυπο που ενσωματώνει RDF δεδομένα σε κείμενα HTML. Παρέχει ένα σύνολο χαρακτηριστικών (attributes) που αυξάνουν την πληροφορία του διαδικτύου με υποδείξεις αναγνώσιμες από τις μηχανές και την περιγραφή για το πως θα εξαχθούν δεδομένα από τα χαρακτηριστικά αυτά.

2.4.5 Γλώσσα Ερωτημάτων SPARQL

Η γλώσσα ερωτημάτων του πλαισίου RDF είναι η γλώσσα SPARQL (Simple Protocol And RDF Query Language), που χρησιμοποιείται για την πρόσβαση στα δεδομένα RDF [27]. Οι περισσότεροι τύποι των ερωτημάτων είναι τρίπτυχα υποκείμενο – ιδιότητα - αντικείμενο, όπως αυτά του RDF πλαισίου, με την διαφορά ότι οι τιμές τους μπορεί να είναι μεταβλητές. Η γλώσσα SPARQL μέσω της διενέργειας ερωτημάτων, καταφέρνει την εξαγωγή τιμών από δομημένα και ημιδομημένα δεδομένα, την εύρεση δεδομένων από μη συσχετισμένες οντότητες και την εφαρμογή συνδέσεων σε ανόμοιες βάσεις δεδομένων.

Ένα ερώτημα SPARQL συμπεριλαμβάνει τα εξής:

- Prefix Declarations, δηλαδή δηλώσεις που χρησιμοποιούνται για την συντόμευση των URI's
- Dataset Definition, που ορίζει το RDF Graph στο οποίο τίθεται το ερώτημα
- Result Clause, που διαπιστώνει την πληροφορία που επιστρέφεται ως απάντηση
- Query Pattern, δηλαδή το μοτίβο του ερωτήματος, μέσω του οποίου προσδιορίζεται το ερώτημα που θα γίνει στο σύνολο δεδομένων και τα

- Query Modifiers, που ταξινομούν τα αποτελέσματα.

Επιπλέον η γλώσσα SPARQL έχει τέσσερις διαφορετικούς τύπους:

- SELECT: Το ερώτημα αυτό επιστρέφει έναν πίνακα με τις τιμές των μεταβλητών που απαντούν ερώτημα.
- CONSTRUCT: Το αποτέλεσμα του ερωτήματος είναι ένας γράφος RDF που δημιουργείται υποκαθιστώντας τις μεταβλητές στο μοτίβο του ερωτήματος.
- DESCRIBE: Το ερώτημα επιστρέφει έναν γράφο RDF που περιγράφει τους πόρους που ζητήθηκαν στο ερώτημα
- ASK: Η απάντηση ενός ερωτήματος ASK είναι μία Boolean τιμή που εκφράζει αν υπάρχει λύση ή όχι στο μοτίβο του ερωτήματος

Επί παραδείγματι, ακολουθεί ένα ερώτημα SELECT στο οποίο ζητείται από το dataset http://dig.csail.mit.edu/2008/webdav/timbl/foaf.rdf να εξαχθούν όλα τα ονόματα που βρίσκονται στο αρχείο foaf, -ένα αρχείο το οποίο περιλαμβάνει πληροφορίες για πρόσωπα.

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

Τα αποτελέσματα που αναμένεται να ληφθούν μετά την υποβολή του ερωτήματος είναι ένας πίνακας όπως αυτός που φαίνεται παρακάτω:

name
"mc schraefel"
"John Klensin"
"Libby Miller"
"Henrik Nielsen"
"John Markoff"

Πίνακας 2.4.5 Αποτελέσματα ερωτήματος SPARQL

2.4.6 Οντολογίες και γλώσσα OWL (Web Ontology Language)

Οι οντολογίες, όπως αναφέρθηκε σε προηγούμενη ενότητα, δίνουν την ακριβή περιγραφή των πραγμάτων και των μεταξύ τους σχέσεων. Η υλοποίησή τους γίνεται μέσω της γλώσσας OWL (Web Ontology Language) που χρησιμοποιείται για την επεξεργασία της πληροφορίας [28]. Η γλώσσα OWL και το πλαίσιο RDF έχουν παρόμοια λειτουργία, με τη διαφορά ότι η OWL έχει μεγαλύτερη δυνατότητα επεξήγησης και έτσι συμπληρώνει τις αδυναμίες του RDF Schema στην περιγραφή περίπλοκων σχέσεων των αντικειμένων ή την εξαγωγή συμπερασμάτων. Η OWL συντάσσεται στην XML για να μπορεί να χρησιμοποιηθεί μεταξύ διαφορετικών τύπων υπολογιστών με διαφορετικά λειτουργικά συστήματα.

Η σύνταξη της OWL περιλαμβάνει μια αλληλουχία από σχόλια, αξιώματα και γεγονότα. Τα σχόλια χρησιμοποιούνται για να περιγράψουν πληροφορίες που σχετίζονται με την οντολογία, αναφορές για άλλες οντολογίες ή πληροφορίες σύνταξης. Το κύριο περιεχόμενο της οντολογίας περιλαμβάνεται στα αξιώματα και στα γεγονότα, που παρέχουν πληροφορίες για κλάσεις και ιδιότητες. Επίσης οι οντολογίες έχουν ονόματα που σχετίζονται με πληροφορίες που είναι χρήσιμες για την δημοσίευση της οντολογίας στο διαδίκτυο.

Η γλώσσα OWL χωρίζεται σε τρεις υπογλώσσες ανάλογα με τις λειτουργίες που παρέχει. Οι γλώσσες αυτές είναι οι εξής [38]:

- OWL Lite: Είναι η γλώσσα με τις λιγότερες δυνατότητες έκφρασης και για αυτό χρησιμοποιείται για απλή δόμηση και ταξινόμηση των οντολογιών. Υπερτερεί όμως από τις άλλες δύο γλώσσες στο ότι είναι πιο εύκολα επεξεργάσιμη από τους υπολογιστές και πιο κατανοητή από τους ανθρώπους.
- OWL DL(Description Logic): Η γλώσσα αυτή υποστηρίζει σε ένα ικανοποιητικό επίπεδο την λογική περιγραφή και την εξαγωγή συμπερασμάτων.
- OWL Full: Συνδυάζει τις αρχές όλων των γλωσσών OWL, δηλαδή της Lite και της DL και για αυτό προσφέρει την μέγιστη περιγραφικότητα των οντολογιών. Είναι πλήρως συμβατή συντακτικά και σημασιολογικά με το RDF και το RDFS, σε αντίθεση με τις άλλες δύο υπογλώσσες.

2.4.7 Κανόνες RIF (Rule Interchange Format)

Ο σχεδιασμός αυτού του επιπέδου βασίζεται στο γεγονός ότι υπάρχει μια πληθώρα κανόνων που ανταλλάσσονται μεταξύ των συστημάτων του ιστού. Οι κανόνες RIF συνιστούν μια ομάδα που ονομάζονται διάλεκτοι[63]. Οι πιο σημαντικοί κανόνες είναι:

RIF-BLD (Basic Logic Dialect): Η διάλεκτος βασικής λογικής αντιστοιχεί στην λογική Horn με κάποιες συντακτικές και σημασιολογικές επεκτάσεις.

RIF-PRD (Production Rule Dialect): Η διάλεκτος παραγωγής κανόνων, στοχεύει στην καταγραφή των κύριων πλευρών των συστημάτων παραγωγής κανόνων. Η παραγωγή κανόνων προσδιορίζεται με υπολογιστικούς μηχανισμούς ad hoc.

RIF-Core: Η διάλεκτος Core είναι ένα υποσύνολο των RIF-BLD και RIF-PRD και έτσι περιορίζει την ανταλλαγή των κανόνων μεταξύ των λογικών κανόνων και των κανόνων παραγωγής.

RIF-FLD (Framework for Logic Dialects): Το πλαίσιο αυτό εισήχθηκε με σκοπό να μειώσει το ποσοστό των προσπαθειών που χρειάζονται ώστε να οριστούν και να επαληθευθούν νέοι λογικοί διάλεκτοι που επεκτείνουν τις δυνατότητες των κανόνων RIF-BLD.

2.4.8 Επίπεδο Λογικής (Logic)

Μέσω του επιπέδου Λογικής, εφαρμόζονται στο σημασιολογικό διαδίκτυο στοιχεία του συστήματος λογικής πρώτου βαθμού, ώστε π.χ. να είναι εφικτή η κατανόηση εννοιών, ο συσχετισμός δηλώσεων, η αναπαράσταση της κατηγορηματικής λογικής κ.α. [28]

2.4.9 Επίπεδο Αποδείξεων (Proof)

Όλη αυτή η γνώση που αποκτήθηκε στα προηγούμενα επίπεδα σε συνδυασμό με τους κανόνες, χρησιμοποιείται από το επίπεδο αποδείξεων μέσω μηχανισμών συμπερασμάτων, για την παροχή εξηγήσεων στους χρήστες. Επίσης τους παρέχεται η

δυνατότητα ανάλυσης της διαδικασίας εξαγωγής συμπερασμάτων. Η ανάλυση μπορεί να επιτευχθεί με τη χρήση κατανοητών μέσων (δέντρα αποδείξεων, λογικές προτάσεις, χρήση τεχνικών παραγωγής εξηγήσεων σε φυσική γλώσσα).

2.4.10 Αξιοπιστία (Trust)

Για την παροχή ενός επιπέδου αξιοπιστίας εφαρμόζονται εργαλεία, όπως είναι η κρυπτογραφία με σκοπό να μην είναι εφικτή η αλλοίωση της πληροφορίας και η πρόσβαση της από μη εξουσιοδοτημένες πηγές και οι ψηφιακές υπογραφές για την πιστοποίηση της προέλευσης των πηγών των πληροφοριών ή της ταυτότητας των χρηστών.

2.5 Τύποι Σημασιολογικών Δεδομένων

Κάθε είδους πληροφοριακός πόρος που διατίθεται για την βελτίωση της διαδικασίας αναζήτησης διαχειρίζεται και φυλάσσεται στο σύστημα ως δεδομένο. Τα δεδομένα χωρίζονται σε σημασιολογικά και ανεπεξέργαστα δεδομένα.

Τα σημασιολογικά δεδομένα περιγράφουν αντικείμενα του πραγματικού κόσμου σε όρους οντοτήτων καθώς και τις μεταξύ τους σχέσεις. Κατά κύριο λόγο περιγράφουν σύνολα δεδομένων RDF, δεδομένα που περιέχονται σε οντολογίες και αναπαριστώνται από γλώσσες του σημασιολογικού ιστού όπως είναι η γλώσσα OWL και το πρότυπο για την αναπαράσταση και ανταλλαγή κανόνων RIF. Τα σημασιολογικά δεδομένα μπορούν να θεωρηθούν ως γράφοι, όπου οι κόμβοι αναπαριστούν τις οντότητες και τις τιμές των χαρακτηριστικών τους και οι ακμές αντιπροσωπεύουν τα χαρακτηριστικά των οντοτήτων ή των σχέσεων μεταξύ των οντοτήτων.

Διακρίνονται τρεις τύποι σημασιολογικών δεδομένων που χρησιμοποιούνται στα σημασιολογικά συστήματα[32]. Τα σημασιολογικά αυτά δεδομένα είναι:

1. Περιγραφές οντοτήτων που περιέχονται σε δομημένες συλλογές εγγράφων:

Αυτή η κατηγορία περιλαμβάνει οντότητες που περιγράφονται εκτενώς μέσα σε ένα κείμενο. Η Wikipedia είναι από τα πιο δημοφιλή παραδείγματα, καθώς είναι μία συλλογή από κείμενα όπου κάθε σελίδα της αντιστοιχεί σε μια οντότητα και συνδέεται με άλλες σελίδες.

2. Περιγραφές δομημένων οντοτήτων σε RDF: Τα δεδομένα αυτά βρίσκονται στον ιστό, καταγράφοντας τις πληροφορίες σχετικά με τα αντικείμενα του πραγματικού κόσμου όπως είναι οι περιγραφές RDF πόρων. Η αναζήτηση Yahoo, Facebook Like, μία μεγάλη ποσότητα περιγραφών RDF πόρων, όπως εστιατόρια, ταινίες έχουν ενσωματωθεί στις σελίδες του ιστού ως RDFa. Ολοένα και περισσότερα δεδομένα διαθέτονται ελεύθερα στο διαδίκτυο και συνδέονται με άλλα σύνολα δεδομένων.
3. Περιγραφές οντοτήτων βασισμένες στη λογική, προσδιορισμένες από γλώσσες αναπαράστασης γνώσης. Τα δεδομένα αυτά περιγράφονται από περισσότερο εκφραστικές γλώσσες αναπαράστασης γνώσης όπως η OWL και η F-Logic, κάνοντας χρήση μοντέλων σημασιολογίας και συλλογιστικής. Ωστόσο οι περισσότερες προσεγγίσεις που υπάρχουν δεν βασίζονται τόσο στην συλλογιστική, η οποία είναι ωφέλιμη αλλά όχι ακόμα απαραίτητη, όσο στην σημασιολογία για την σωστή κατανόηση των διφορούμενων ερωτημάτων και δεδομένων.

3.Ευφυείς Πράκτορες (Intelligent Agents)

3.1 Πρόλογος

Η εύρεση και μεταφορά των δεδομένων στον σημασιολογικό ιστό πραγματοποιείται μέσω ειδικών προγραμμάτων, των πρακτόρων, που δέχονται ερεθίσματα από το περιβάλλον τους και δρουν ανάλογα. Η διαχείριση της σημασιολογικής πληροφορίας απαιτεί την ύπαρξη πολλών πρακτόρων που συνεργάζονται σε μία ενιαία πλατφόρμα. Παρακάτω αναλύεται η έννοια και η λειτουργία ενός πράκτορα και ενός συστήματος πολλών πρακτόρων καθώς και η συμβολή τους στον σημασιολογικό ιστό και την αναζήτηση πληροφορίας.

3.2 Τι είναι οι Ευφυείς Πράκτορες (Intelligent Agents)

Ένας ευφυής πράκτορας (intelligent agent) περιγράφεται ως μία οντότητα που διαθέτει ένα επίπεδο ευφυΐας και αυτονομίας, ώστε να μπορεί να εκτελεί χωρίς την επίβλεψη του χρήστη και με βάση αυτά που αντιλαμβάνεται ένα σύνολο από λειτουργίες.

Ο ευφυής πράκτορας αντιπροσωπεύει μία ξεχωριστή κατηγορία λογισμικού που έχει την ικανότητα να αποκτά γνώση, να δρα αυτόνομα και να πραγματοποιεί τους στόχους του. Για να αντιληφθεί το περιβάλλον στο οποίο βρίσκεται χρησιμοποιεί τους αισθητήρες (sensors), ενώ δρα σε αυτό μέσω δρώντων αντικειμένων (actuators)[29], [30].

Τα βασικά χαρακτηριστικά που θα πρέπει να διαθέτει ένας ευφυής πράκτορας είναι:

- Αυτονομία: Ένας πράκτορας θα πρέπει να είναι σε θέση μόνος του να επιλέγει και να πράττει τις ενέργειες που απαιτούνται για την ικανοποίηση των στόχων που αναλαμβάνει.
- Επικοινωνία με το χρήστη: Η αυτονομία του πράκτορα αφορά τις αποφάσεις και τις ενέργειες του, παρόλα ταύτα, η επικοινωνία του με τον χρήστη είναι απαραίτητη ώστε να πληροφορείται για τις όποιες ανάγκες ή επιθυμίες έχει.
- Αντιδραστικότητα: Η αντιδραστικότητα αναφέρεται στην ικανότητα του να αντιλαμβάνεται το περιβάλλον του και να δρα κατάλληλα.

- Προσαρμοστικότητα: Ένας ευφυής πράκτορας θα πρέπει να είναι σε θέση να προσαρμόζεται σε όποια αλλαγή συμβαίνει στο περιβάλλον του χωρίς να παρεκκλίνει από τον στόχο του.
- Δυνατότητα μάθησης: Η δυνατότητα να μαθαίνει παρέχει στον πράκτορα τη γνώση που χρειάζεται, για να είναι σε θέση να διεκπεραιώσει τα αιτήματα που θέτονται από τους χρήστες.

3.3 Κατηγορίες Ευφυών Πρακτόρων

Οι ευφυείς πράκτορες διακρίνονται σε πέντε κατηγορίες με βάση το επίπεδο ικανοτήτων και αντίληψης τους [29].

1. Απλοί Αντανακλαστικοί Πράκτορες (Simple Reflex Agents). Οι πράκτορες αυτοί ενεργούν βασισμένοι την τρέχουσα αντίληψη τους, αγνοώντας οποιαδήποτε άλλη προηγούμενη. Ο τρόπος δράσης τους υπακούει στον κανόνα της μορφής εάν–τότε: IF <συνθήκη> Then <δράση>. Οι αντανακλαστικοί πράκτορες αποδίδουν όταν αντιλαμβάνονται πλήρως το περιβάλλον δράσης τους. Σε μερικώς παρατηρήσιμα περιβάλλοντα, ελλοχεύει ο κίνδυνος ο πράκτορας να παγιδευτεί σε έναν ατέρμονα βρόχο. Εντούτοις, υπάρχουν κάποιοι αντανακλαστικοί πράκτορες με την ικανότητα να αντιληφθούν τις ενέργειες που έχουν κάνει και να αποφύγουν την επανάληψή τους.
2. Αντανακλαστικοί Πράκτορες Βασισμένοι σε Μοντέλα (Model-based reflex agents). Οι πράκτορες αυτοί, έχουν την ικανότητα να ενεργούν σε μερικώς παρατηρήσιμα περιβάλλοντα. Στο εσωτερικό του πράκτορα είναι αποθηκευμένη η τρέχουσα κατάσταση που αντιλαμβάνεται καθώς και ένα μοντέλο που περιγράφει την κατάσταση του περιβάλλοντος που δεν του είναι αντιληπτή. Το μοντέλο αυτό διατηρεί επίσης και γνώση από παρελθοντική αντίληψη του πράκτορα.
3. Πράκτορες Βασισμένοι σε Στόχους (Goal-based agents). Η αντίληψη των πρακτόρων αυτής της κατηγορίας προέρχεται από το αποθηκευμένο μοντέλο και από τους στόχους του. Με αυτόν τον τρόπο λαμβάνει υπόψιν του πληροφορίες για τις επιθυμητές καταστάσεις, εντοπίζει τις πιθανές επιλογές που ικανοποιούν τους στόχους του και καταλήγει στον τρόπο ενέργειας.

4. Πράκτορες Βασισμένοι στην Ωφέλεια (Utility-based agents). Μέσω μιας συνάρτησης ωφέλειας (utility function), ο πράκτορας αντιλαμβάνεται πόσο χρήσιμη είναι μία κατάσταση και σε ποιο βαθμό θα τον βοηθήσει να ικανοποιήσει το στόχο του. Στη συνέχεια από τα αποτελέσματα που θα λάβει θα πρέπει να επιλέξει να ενεργήσει με βάση την κατάσταση που μεγιστοποιεί την ικανοποίησή του.
5. Πράκτορες Μάθησης (Learning agents). Η ικανότητα που έχει ένας πράκτορας να μαθαίνει, του επιτρέπει να λειτουργεί καλύτερα σε άγνωστα περιβάλλοντα. Τα βασικά τους στοιχεία είναι το στοιχείο μάθησης (learning element), που συγκεντρώνει τη γνώση που αποκτήθηκε και αποφασίζει τι θα τροποποιηθεί στον τρόπο δράσης του πράκτορα, το στοιχείο απόδοσης (performance element) που δρα βασισμένο την τροποποίηση του στοιχείου μάθησης και τέλος το στοιχείο problem generator, που είναι υπεύθυνο για την απόκτηση νέας γνώσης μέσω νέων δράσεων.

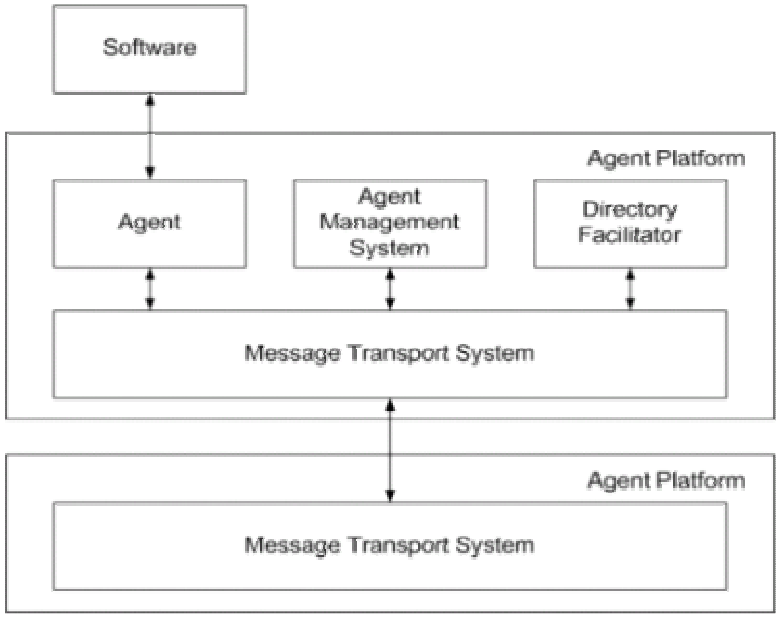
3.4 Συστήματα πολλών πρακτόρων (Multi-agents system)

Στο σημασιολογικό ιστό, η επικοινωνία μεταξύ πρακτόρων είναι απαραίτητη για την ανταλλαγή και τον διαμοιρασμό της πληροφορίας. Με τον τρόπο αυτό διαμορφώνεται ένα σύστημα πολλών πρακτόρων (multi-agent system) [31], [32], [33]. Σε ένα τέτοιο σύστημα κάθε πράκτορας θα πρέπει να διαθέτει κάποιες συγκεκριμένες επικοινωνιακές ικανότητες ώστε να είναι σε θέση να επιτελεί τις λειτουργίες της παράδοσης, της λήψης, της ανάλυσης και της κατανόησης των μηνυμάτων.

Η περιγραφή και η λήψη των μηνυμάτων αφορά το φυσικό επίπεδο, η ανάλυση το συντακτικό και η κατανόηση των μηνυμάτων το σημασιολογικό επίπεδο. Τα επίπεδα αυτά ορίζονται από τον FIPA (Foundation for Intelligent Physical Agents), τον οργανισμό προτυποποίησης για τα συστήματα πολλών πρακτόρων. Στο σημασιολογικό επίπεδο το πρότυπο περιγράφει το περιεχόμενο της γλώσσας και την χρήση των οντολογιών που χρησιμοποιούνται για την αποκωδικοποίηση της πληροφορίας που περιγράφεται στο μήνυμα.

3.4.1 Δομή συστήματος πολλών πρακτόρων

Σύμφωνα με τον οργανισμό FIPA, η δομή που έχει ένα σύστημα πολλών πρακτόρων απεικονίζεται στην εικόνα 3.4.1[28].



Εικόνα 3.4.1 - Δομή Συστήματος Πολλών Πρακτόρων

Με βάση αυτή τη δομή, μία πλατφόρμα πρακτόρων AP (agent platform), αποτελεί τη φυσική υποδομή που δρουν οι πράκτορες και απαρτίζεται από τους πράκτορες, το λογισμικό υποστήριξής τους, το λειτουργικό σύστημα, τη μηχανή και τα στοιχεία διαχείρισης των πρακτόρων. Τα στοιχεία διαχείρισης αποτελούνται:

- Το σύστημα διαχείρισης πράκτορα AMS (Agent Management System), το οποίο ελέγχει την πρόσβαση και την χρήση της πλατφόρμας από τους πράκτορες,
- Η υπηρεσία μεταφοράς μηνυμάτων MTS (Message Transport Service), που υποστηρίζει την μεταφορά των μηνυμάτων μεταξύ των πρακτόρων,
- και έναν προαιρετικό κατάλογο DF (Directory Facilitator), στον οποίο καταγράφονται οι υπηρεσίες που παρέχουν οι πράκτορες.

Οι πράκτορες επικοινωνούν μεταξύ τους με την γλώσσα ACL (Agent Communication Language), που περιγράφει συνήθως παραμέτρους όπως ο αποστολέας και ο παραλήπτης, ενώ η γλώσσα που χρησιμοποιείται συνήθως για την περιγραφή του περιεχομένου του μηνύματος είναι η SL (Semantic Language).

Η υλοποίηση ενός συστήματος πολλών πρακτόρων συνήθως πραγματοποιείται με το πλαίσιο JADE (Java Agent DEvelopment Framework) που είναι συμμορφωμένο με τις προδιαγραφές που ορίζει ο FIPA.

3.5 Πράκτορες και οντολογίες

Το περιεχόμενο των μηνυμάτων που ανταλλάσσεται μεταξύ των πρακτόρων προέρχεται από την γνώση που αποκτήθηκε μέσω των οντολογιών. Συνεπώς η περιγραφή ενός γεγονότος από μια οντολογία θα πρέπει να είναι ακριβής και εύκολα προσβάσιμη. Για τον σκοπό αυτό, την διευκόλυνση δηλαδή της επικοινωνίας σε ένα σύστημα πολλών πρακτόρων, ο FIPA έχει προτείνει την ύπαρξη ενός πράκτορα οντολογίας (ontology agent- OA) στην πλατφόρμα πράκτορα, που θα προσφέρει υπηρεσίες όπως για παράδειγμα, εύρεση και πρόσβαση στις οντολογίες, επιλογή της κατάλληλης οντολογίας που χρειάζεται για την επικοινωνία των πρακτόρων, απόκριση σε ερωτήματα που αφορούν τις σχέσεις μεταξύ των όρων ή των οντολογιών, μετάφραση εκφράσεων μεταξύ διαφορετικών οντολογιών ή μεταξύ διαφορετικών γλωσσών. Ένας πράκτορας οντολογίας μπορεί να προσφέρει κάποιες ή και όλες από αυτές τις υπηρεσίες, χρησιμοποιώντας την ορισμένη υπηρεσία Ontology service ontology.
Η ύπαρξη ενός πράκτορα οντολογίας δεν είναι απαραίτητη σε ένα σύστημα πολλών πρακτόρων, αλλά αν αυτή αποφασιστεί τότε θα πρέπει να είναι συμβατή με τις προδιαγραφές του FIPA όπως ορίζεται στην προδιαγραφή για την υπηρεσία της οντολογίας (Ontology Service Specification).
Επίσης αναγκαία για την κατανόηση της γνώσης και της οντολογίας από τους πράκτορες είναι η ύπαρξη ενός μοντέλου γνώσης και ένα πρότυπο οντολογίας, που θα περιγράφει τις αρχές των εννοιών, των χαρακτηριστικών και των σχέσεων. Το πρότυπο που προτείνεται από τον FIPA και έχει οριστεί για την περιγραφή και την διαχείριση των οντολογιών είναι το μοντέλο γνώσης βασισμένο στο πλαίσιο OKBC (frame-based OKBC Knowledge Model).

3.6 Σημασιολογική Αναζήτηση και Πράκτορες

Οι πράκτορες τεχνητής νοημοσύνης συντελούν σε ένα νέο τρόπο πλοήγησης και ανάκτησης πληροφορίας. Τα βασικά χαρακτηριστικά τους, η ικανότητα αντίληψης

και η αυτονομία στην εκτέλεση λειτουργιών, επιτρέπουν την συλλογή πληροφορίας από διαφορετικές πηγές, την κατανόηση της πληροφορίας, την επεξεργασία της αναζήτησης και τον διαμοιρασμό της γνώσης.

Στις παραδοσιακές μηχανές αναζήτησης, ο ρόλος ενός πράκτορα ήταν η αναζήτηση πληροφοριών στις ιστοσελίδες με την χρήση ευριστικών κανόνων ή η διαμεσολάβηση τους για την ανταλλαγή πληροφοριών σε ετερογενείς πηγές πληροφόρησης που είχαν όμως αυστηρή σύνταξη και σημασιολογία.

Στις σημασιολογικές μηχανές αναζήτησης, κατά την διαδικασία εύρεσης πληροφορίας οι πράκτορες λαμβάνουν τα αιτήματα εξυπηρέτησης των χρηστών, τα κατανοούν και τα επεξεργάζονται και στη συνέχεια αναζητούν τις σχετικές πληροφορίες από τις πηγές του διαδικτύου. Ταυτόχρονα επικοινωνούν με τους άλλους πράκτορες, συγκρίνουν τις πληροφορίες και καταλήγουν στα αποτελέσματα που δίνονται στον χρήστη.

Η αναζήτηση της πληροφορίας μέσω των ευφυών πρακτόρων, βελτιώνει την απόδοση και την ποιότητα της έρευνας στον ιστό και κατά αυτόν τον τρόπο επιτυγχάνεται ο στόχος της σημασιολογικής αναζήτησης. Μέσω των αυτοματοποιημένων συσχετισμών των ευφυών πρακτόρων, οι απαντήσεις που δίνονται στον χρήστη υπερβαίνουν το πρόβλημα της χαοτικής διάστασης του διαδικτύου και αποκτούν δομημένη και σημασιολογική οργάνωση.

4. Βασικές Λειτουργίες μιας Σημασιολογικής Μηχανής Αναζήτησης.

4.1 Εισαγωγή

Σε αυτή την ενότητα θα αναπτυχθεί η αρχιτεκτονική μιας σημασιολογικής μηχανής ενημέρωσης, δηλαδή ο γενικός σχεδιασμός και η βασική δομή που εξασφαλίζουν την αποδοτική διαχείριση των ερωτημάτων των χρηστών. Επιπλέον, μελετάται η μεθοδολογία της αρχιτεκτονικής αυτής, η διαδικασία της σημασιολογικής έρευνας όπως ολοκληρώνεται μέσα από τα επιμέρους στάδια της σημασιολογικής μηχανής και ο μηχανισμός αποθήκευσης. Πριν από τον σχεδιασμό όμως, θα πρέπει να περιγραφούν οι παράγοντες που λαμβάνει υπ' όψιν της η σημασιολογική έρευνα και οι προκλήσεις που θα πρέπει να αντιμετωπίσει, ώστε μία σημασιολογική μηχανή να έχει την καλύτερη δυνατή απόδοση.

4.2 Προκλήσεις Σημασιολογικής Έρευνας

Τα τελευταία χρόνια έχει αυξηθεί σημαντικά το ποσοστό των δεδομένων του σημασιολογικού διαδικτύου που είναι διαθέσιμο και αξιοποιήσιμο. Εξαιτίας της αύξησης αυτής, οι σημασιολογικές μηχανές αναζήτησης έχουν να αντιμετωπίσουν σημαντικές προκλήσεις όσον αφορά την προσβασιμότητα των δεδομένων. Οι προκλήσεις αυτές είναι [34], [54]:

Ανομοιογένεια: Η πρώτη πρόκληση με την οποία έρχεται αντιμέτωπη η σημασιολογική έρευνα, σχετίζεται με την ανομοιογένεια των εφαρμογών και του διαδικτύου. Παρά την προσπάθεια για την τυποποίηση των τεχνολογιών, εξακολουθεί να υπάρχει η διαφορετικότητα σε πολλές διαστάσεις της πληροφορίας, όπως π.χ. στην ποιότητα των οντολογιών, την πολυπλοκότητα και την μοντελοποίηση. Για τον λόγο αυτό, είναι απαραίτητη η δημιουργία ομοιογενών μηχανισμών πρόσβασης που θα επιτρέπουν όμοιο τρόπο διαχείρισης της γνώσης σε ετερογενή συστήματα.

Επεκτασιμότητα: Λόγω του μεγάλου πλήθους των κειμένων και των τρίπτυχων που υπάρχουν, το σημασιολογικό διαδίκτυο έχει ξεπεράσει το μέγεθος οποιασδήποτε υπάρχουσας βάσης γνώσης, σε οποιαδήποτε υπάρχουσα σημασιολογική εφαρμογή. Για το λόγο αυτό η επέκταση των μηχανισμών πρόσβασης απαιτεί την δυνατότητα

εκμετάλλευσης όλων των αυξανόμενων σημασιολογικών πληροφοριών καθώς και τη διαχείριση των πληροφοριών με αδόμητο περιεχόμενο.

Ποιότητα: Η ποιότητα των δεδομένων είναι ένας κρίσιμος παράγοντας σε όλα τα συστήματα λήψης αποφάσεων και επεξεργασίας δεδομένων. Στο σημασιολογικό διαδίκτυο η πληροφορία προέρχεται από πολλές διαφορετικές πηγές και για αυτό το επίπεδο της ποιότητας που σχετίζεται με την ορθότητα, ακρίβεια ή αξιοπιστία παρέχεται σε διαφορετικό βαθμό.

Τις προκλήσεις που περιγράψαμε, προσπαθούν να αντιμετωπίσουν οι σημασιολογικές μηχανές αναζήτησης που έχουν δημιουργηθεί. Στόχος τους η ανάπτυξη μιας υποδομής που επιτυγχάνει την αποτελεσματική πρόσβαση στα σημασιολογικά δεδομένα, με κοινά χαρακτηριστικά. Τα χαρακτηριστικά αυτά είναι: η παροχή ενός ευρετηρίου για όλα τα σημασιολογικά δεδομένα που είναι δημοσιευμένα, ώστε να φιλτράρουν το σχετικό σύνολο δεδομένων που χρησιμοποιείται για τις απαντήσεις των ερωτημάτων, η παροχή υπηρεσιών ταξινόμησης και η παροχή άλλων προηγμένων υπηρεσιών στους χρήστες, για την υποστήριξη της αυτοματοποιημένης επεξεργασίας των δεδομένων και της διεξαγωγής έξυπνου φιλτραρίσματος και εργασίες ολοκλήρωσης.

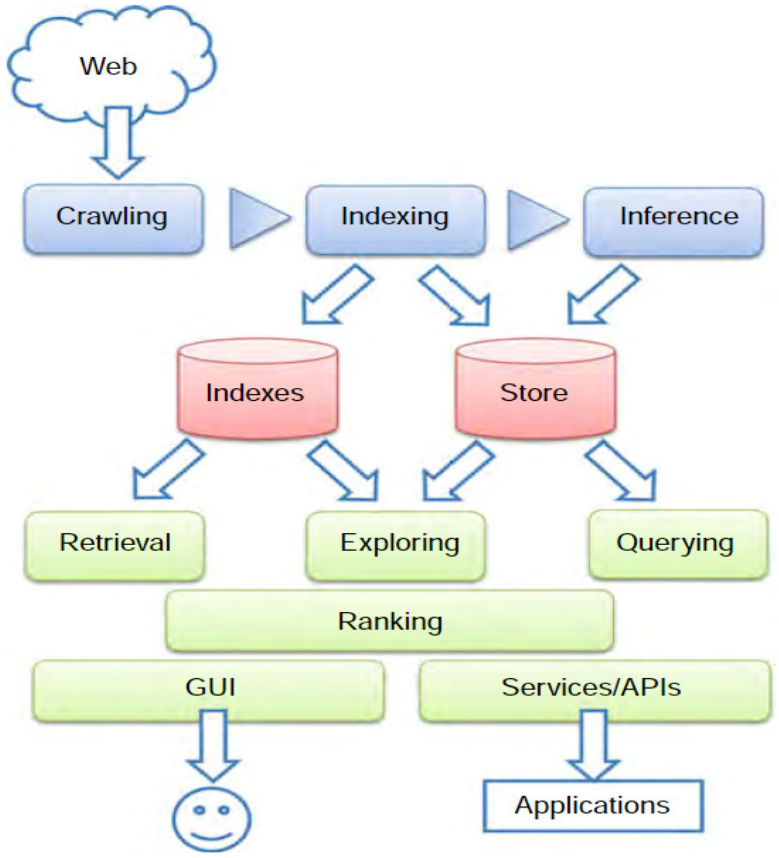
4.3 Βασική αρχιτεκτονική σημασιολογικών μηχανών αναζήτησης

Η ανάγκη για την δημιουργία αποδοτικών, ισχυρών και επεκτάσιμων σημασιολογικών μηχανών αναζήτησης, συγκέντρωσε ιδιαίτερο ενδιαφέρον από την σημασιολογική κοινότητα. Τα συστήματα που αναπτύχθηκαν, παρόλο που λαμβάνουν υπόψη τους διαφορετικές οπτικές της σημασιολογικής έρευνας, εστιάζουν σε διαφορετικά σημεία και βασίζονται σε διαφορετικές παραδοχές, λειτουργούν πάνω σε μία κοινή βάση [34].

Όπως έχουμε αναφέρει, η σημασιολογική μηχανή αναζήτησης είναι ένα σύστημα που συγκεντρώνει, ευρετηριάζει και αναλύει τα κείμενα του σημασιολογικού ιστού για να παρέχει μηχανισμούς αναζήτησης και υποβολής ερωτημάτων. Τα σημασιολογικά κείμενα περιέχουν πληροφορίες κωδικοποιημένες με την χρήση τυποποιημένων γλωσσών του σημασιολογικού διαδικτύου όπως είναι η RDF, RDFS και OWL.

Στο παρακάτω σχεδιάγραμμα (Εικόνα 4.1), παρουσιάζεται ένα γενικό πλαίσιο των κοινών δραστηριοτήτων που πραγματοποιούνται από μία σημασιολογική μηχανή αναζήτησης. Κάποια από αυτά τα μέρη δεν υπάρχουν σε όλα τα συστήματα π.χ.

μπορεί να μην χρησιμοποιούν crawler ή ο τρόπος που τον εφαρμόζουν να διαφοροποιείται. Στις επόμενες υποενότητες, θα αναπτυχθούν και θα περιγραφούν αναλυτικότερα τα στάδια αυτού του γενικού πλαισίου, που συνιστούν την αρχιτεκτονική μιας σημασιολογικής μηχανής αναζήτησης. Η αρχιτεκτονική αυτή, όπως παρατηρούμε, έχει σημαντικές ομοιότητες με την αρχιτεκτονική μιας παραδοσιακής μηχανής αναζήτησης, όσον αφορά την δομή και τον σχεδιασμό της.



Εικόνα 4.1 Βασική αρχιτεκτονική σημασιολογικής μηχανής αναζήτησης

4.3.1 Ανίχνευση Σημασιολογικού Ιστού (Crawling)

Κατά την διαδικασία εξερεύνησης του σημασιολογικού ιστού, διενεργούνται οι εργασίες της ανίχνευσης, ευρετηρίασης, εξαγωγής και κατηγοριοποίησης δεδομένων. Όσον αφορά την ανίχνευση (crawling) του ιστού για την εύρεση κειμένων του σημασιολογικού ιστού (SWDs), εφαρμόζεται με διαφορετικές τεχνικές συγκριτικά με την ανίχνευση όπως αυτή εφαρμόζεται στις παραδοσιακές μηχανές αναζήτησης. Στις

παραδοσιακές μηχανές αναζήτησης, οι HTML ανιχνευτές εξάγουν συνδέσμους από HTML σελίδες με σκοπό να βρουν επιπλέον πηγές για ανίχνευση. Ο μηχανισμός αυτός δεν λειτουργεί αποδοτικά για πηγές δομημένων δεδομένων, καθώς συχνά οι υπερσύνδεσμοι δεν έχουν μεταξύ τους εννοιολογική σχέση. Στην σημασιολογική ανίχνευση αναζητούνται κείμενα SWDs που είναι αποθηκευμένα σε διάφορες μορφές RDF, OWL, FOAF, RSS κτλ. και η εύρεση τους επιτυγχάνεται με έναν RDF crawler. Η διαφορά της εύρεσης RDF από HTML δεδομένων, είναι ότι το RDF διαθέτει έναν μηχανισμό που ενσωματώνει πολλά RDF μοντέλα και συντελεί στην δημιουργία ενός ενοποιημένου μοντέλου. Με αυτόν τον τρόπο, αντί για μία βάση δεδομένων αποτελούμενη από λέξεις-κλειδιά και τους συνδέσμους με τις τοποθεσίες που βρίσκονται τα HTML κείμενα, δημιουργείται ένα μοντέλο που περικλείει όλη την πληροφορία που έχει εντοπιστεί.

Ωστόσο, τα κείμενα με σημασιολογικές πληροφορίες που διατίθενται στο διαδίκτυο έχουν αυξηθεί όπως και οι σχέσεις που τα συνδέουν μεταξύ τους. Επίσης πολλά κείμενα SWDs δείχνουν σε άλλα που δεν περιέχουν σημασιολογική πληροφορία. Εξαιτίας των παραγόντων αυτών και για τον περιορισμό των υποψήφιων συνδέσμων ακολουθούνται ευριστικοί κανόνες. Ο ανιχνευτής του σημασιολογικού ιστού, μπορεί να χρησιμοποιεί συμβατικές μηχανές αναζήτησης για να ανακαλύψει αρχικούς σπόρους SWDs. Όσον αφορά θέματα, όπως πόσο συχνά γίνεται η επανεπισκεψιμότητα των κειμένων για την παρακολούθηση αλλαγών, ισχύουν οι ίδιοι κανόνες για το συμβατικό και το σημασιολογικό διαδίκτυο. Παρόλα αυτά, η τροποποίηση ενός SWD μπορεί να επιφέρει περισσότερο εκτεταμένες παρά τοπικές επιδράσεις και μπορεί να πυροδοτήσει σημαντική δουλειά για μια σημασιολογική μηχανή αναζήτησης.

4.3.2 Ευρετηρίαση (Indexing)

Κατά την διαδικασία της ευρετηρίασης μπορεί να χρησιμοποιηθεί ο κλασσικός μηχανισμός ευρετηρίασης, για να συσχετίσει τα σημασιολογικά έγγραφα με ένα σύνολο όρων, αλλά τα περισσότερα από τα υπάρχοντα συστήματα ενισχύουν την χρησιμότητα αυτών των ευρετηρίων για την αναζήτηση πλήρων κειμένων με πρόσθετες πληροφορίες, όπως στοιχεία μεταδεδομένων που σχετίζονται με κάθε έγγραφο ή ευρετήρια με τα περιεχόμενα των εγγράφων, που περιγράφουν τις σχέσεις μεταξύ των οντοτήτων.

4.3.3 Συμπερασμός (Inference)

Ο συμπερασμός χρησιμοποιείται για να βελτιώσει τα σύνολα δεδομένων που έχουν συγκεντρωθεί και να συμπεριλάβει την πληροφορία που συνάγεται. Οι διαδικασίες του συλλογισμού μπορούν να χρησιμοποιηθούν κατά την διάρκεια του ευρετηριασμού ή κατά την διάρκεια της υποβολής ερωτημάτων.

4.3.4 Κατάταξη (Ranking)

Η διαδικασία της ταξινόμησης των αποτελεσμάτων, έχει ως στόχο της να διευκολύνει την επιλογή της πιο σχετικής πληροφορίας και είναι ήδη γνωστή από τις παραδοσιακές μηχανές αναζήτησης. Η επικρατέστερη μέθοδος στις μηχανές του παγκόσμιου ιστού, ήταν ο αλγόριθμος PageRank της Google, που ταξινομούσε τα αποτελέσματα με βάση την δημοτικότητά τους. Ο αλγόριθμος αυτός, δεν είναι το ίδιο αποδοτικός στον σημασιολογικό ιστό, καθώς οι σύνδεσμοι μπορεί να οδηγούν σε διαφορετικού είδους δεδομένα και για αυτό κάθε σύνδεσμος πρέπει να διαχειρίζεται με διαφορετικό τρόπο. Επίσης η έννοια της συνάφειας στον σημασιολογικό ιστό, μπορεί να είναι πιο ασαφής και εξαρτημένη από το περιεχόμενο. Κατά συνέπεια, τα συστήματα του σημασιολογικού διαδικτύου που έχουν αναπτυχθεί, υιοθετούν διαφορετικές προσεγγίσεις από την χρήση απλών μετρικών της ανάκτησης πληροφορίας μέχρι πιο εξειδικευμένες μετρικές και προσαρμοσμένες κατατάξεις.

4.3.5 Ανάκτηση (Retrieval)

Οι δυνατότητες ανάκτησης δεδομένων ποικίλλουν στα διάφορα συστήματα. Οι περιοχές εισόδου μπορούν να είναι λέξεις-κλειδιά ή ερωτήματα και τα αποτελέσματα URIs σημασιολογικών εγγράφων, όροι του σημασιολογικού ιστού (κλάσεις, ιδιότητες) ή αντικείμενα. Τα αποτελέσματα μπορούν επίσης να παρουσιάζονται με πρόσθετα σημασιολογικά μεταδεδομένα.

4.3.6 Επερωτήσεις (Querying)

Η λειτουργία της αναζήτησης βασίζεται γενικά στην έρευνα μέσω λέξεων κλειδιών, παρόλα αυτά μερικά συστήματα παρέχουν πιο επίσημους τρόπους για την υποβολή

ερωτήσεων στη συλλογή των κειμένων που περιέχουν. Ένα παράδειγμα είναι η χρήση της γλώσσας SPARQL που επιτρέπει στους χρήστες και στις εφαρμογές, να έχουν άμεση πρόσβαση στο περιεχόμενο των κειμένων, επιτρέποντας έτσι την εκμετάλλευσή τους.

4.3.7 Περιήγηση (Exploring)

Οι σημασιολογικές μηχανές αναζήτησης επιτρέπουν στον χρήστη να περιηγηθεί στα κείμενα που έχουν εντοπιστεί, να ελέγξει τις πληροφορίες που συνδέονται με τα έγγραφα ή να βελτιώσει το ερώτημα μέσω των μηχανισμών επέκτασης.

4.3.8 Διεπαφή αναζήτησης (Search interface)

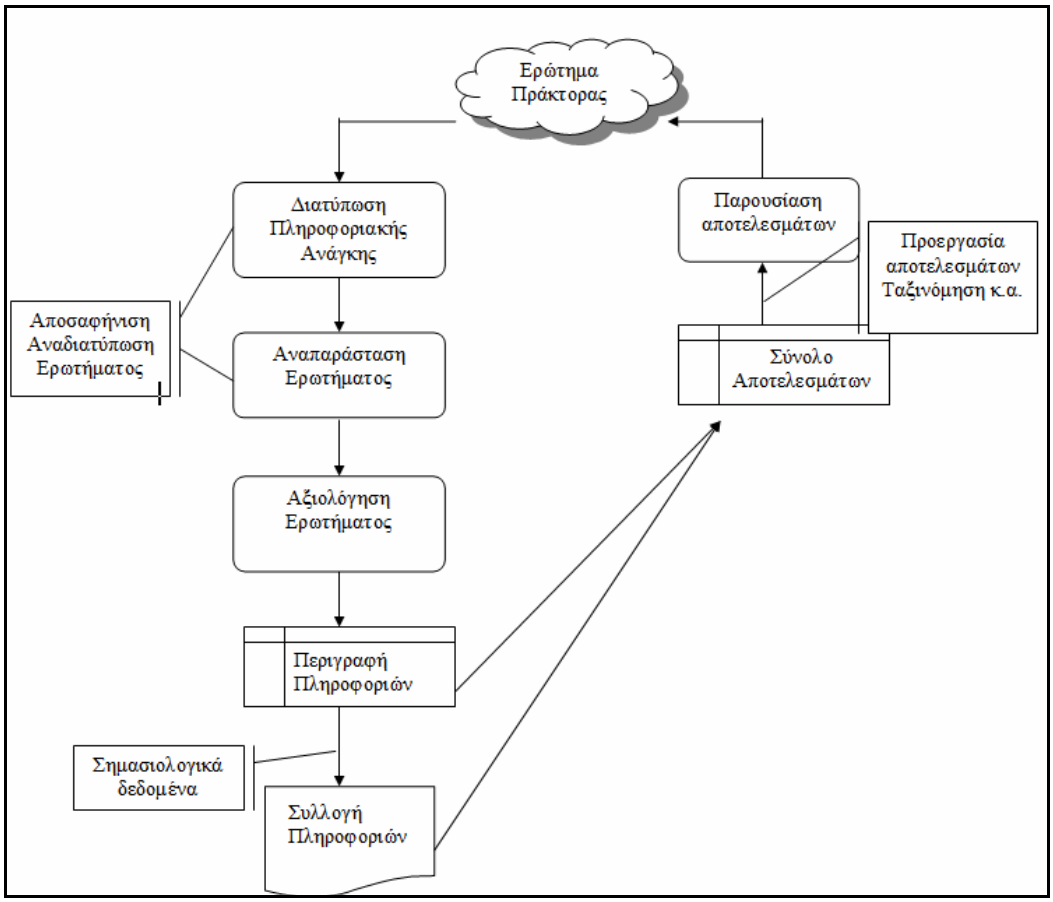
Τα περισσότερα συστήματα αναθέτουν σε πράκτορες την διαδικασία αναζήτησης, επιτρέποντας τους να αποκτούν απευθείας πρόσβαση στα μεταδεδομένα και τα αποτελέσματα ενώ η επικοινωνία με τους χρήστες γίνεται μέσω γραφικών διεπαφών χρήστη.

4.4 Διαδικασία σημασιολογικής έρευνας

Η διαδικασία της σημασιολογικής έρευνας αρχίζει με τον πράκτορα να γίνεται γνώστης της πραγματικής πληροφοριακής ανάγκης του χρήστη [55]. Η ανάγκη αυτή εκφράζεται από τον πράκτορα είτε σε μορφή όρων, φυσικού ερωτήματος, κειμένου, ερωτήματος SPARQL κτλ. Αν υπάρχει διαφοροποίηση στη σύνταξη της αναπαράστασης του ερωτήματος και στην διατύπωση της πληροφοριακής ανάγκης, τότε θα πρέπει να γίνει η μετάφρασή του ερωτήματος. Στην περίπτωση που η σύνταξη είναι ακριβής, αλλά διαφέρει η εννοιολογική περιγραφή των στοιχείων, ή είναι ασαφής, συγκεκριμένη, ή ασυνεπής, τότε θα πρέπει να βελτιωθεί ή να επαναδιατυπωθεί η διαδικασία της μετάφρασης. Κατά την διαδικασία αυτή, η χρήση των οντολογιών μπορεί να βοηθήσει να ξεπεραστούν οι ασάφειες της φυσικής γλώσσας, μέσω της κωδικοποιημένης γνώσης για τη σχέση μεταξύ των εκφράσεων της φυσικής γλώσσας και των εννοιών τους. Αυτό το επιτυγχάνει παρέχοντας ένα καλά ορισμένο λεξιλόγιο για την διαμόρφωση ασαφών, πολύ συγκεκριμένων και πολύπλοκων ερωτημάτων.

Στη συνέχεια, για την αναζήτηση μιας λύσης, το υποσύστημα αξιολόγησης ερωτήματος ταιριάζει την επίσημη αναπαράσταση της πληροφοριακής ανάγκης με τις επίσημες αναπαραστάσεις του πληροφοριακού περιεχομένου του αντικειμένου με σκοπό να αναγνωρίσει τα πληροφοριακά αντικείμενα, όπως κείμενα, εγγραφές βάσεις δεδομένων κτλ. Τα αντικείμενα αυτά είναι περισσότερο πιθανό να ικανοποιούν την ανάγκη για πληροφορία. Για τον λόγο αυτό εφαρμόζονται διάφορες προσεγγίσεις ανάκτησης πληροφορίας, όπως το Vector Space model, το μοντέλο Boolean, διάφορες προσεγγίσεις βασισμένες στη λογική, και διάφορες υβριδικές προσεγγίσεις που συνδυάζουν ανάκτηση πληροφορίας βασισμένη στο κείμενο και σε μεταδεδομένα ή ανάκτηση βασισμένη στην ομοιότητα.

Με την χρήση των σημασιολογικών μεταδεδομένων για την περιγραφή των πληροφοριακών αντικειμένων, υπάρχουν πολλά πλεονεκτήματα. Τα μεταδεδομένα επιτρέπουν την περιγραφή πλευρών του πληροφοριακού αντικειμένου, που δεν περιέχονται στο αντικείμενο αυτό, την ακριβή αναφορά σε τμήματα της γνώσης και όχι σε ολόκληρο το αντικείμενο πληροφορίας ή αντίθετα την γενική αναφορά σε ένα τμήμα γνώσης. Επίσης μέσω τυποποιημένων ή ομοιογενών μεταδεδομένων, πληροφοριακά στοιχεία από ετερογενείς πηγές μπορούν να συγκεντρωθούν και να συσχετιστούν. Μετά το πέρας της επεξεργασίας των αποτελεσμάτων αναζήτησης ένας αριθμός πιθανών απαντήσεων έχει ανακτηθεί, που μπορεί να υποστηριχθεί από τις οντολογίες. Τα αποτελέσματα κατατάσσονται και ταξινομούνται σύμφωνα με δηλωτικούς κανόνες. Κατά την ανάκτηση κειμένων, μπορούν να εφαρμοστούν αλγόριθμοι εξαγωγής πληροφορίας για την απόκτηση της πραγματικής γνώσης από τα έγγραφα. Η γνώση αυτή, μπορεί να αποθηκευτεί σε οντολογικές δομές δεδομένων και να υποστεί περαιτέρω επεξεργασία.



Εικόνα 4.4 - Διαδικασία σημασιολογικής αναζήτησης

Ειδικά για την περίπτωση μεγάλων συνόλων απαντήσεων ή σύνθετων περιοχών πληροφορίας, ορισμένες πτυχές του συνόλου των απαντήσεων αφήνονται για περαιτέρω επιθεώρηση και περιήγηση από ερευνητές. Μετά την επεξεργασία, οι απαντήσεις μπορεί να υπόκεινται σε εξατομίκευση ή στην προσαρμογή του περιεχομένου τους. Στην εικόνα 4.4 περιγράφονται σχηματικά τα βήματα της διαδικασίας της σημασιολογικής έρευνας που περιγράφτηκε προηγουμένως.

4.5 Μεθοδολογίες σημασιολογικής έρευνας

Στο πεδίο της σημασιολογικής έρευνας, υπάρχουν ορισμένες κοινές μεθοδολογίες που χρησιμοποιούνται στις εφαρμογές για την αναζήτηση σημασιολογικών πληροφοριών. Εν συνεχεία, αναλύονται πέντε από τις πιο διαδεδομένες, των οποίων η γνώση, η κατανόηση και ο τρόπος χρήσης τους, θα χρησιμεύσει και μελλοντικά στις επόμενες προσεγγίσεις [35].

- **Διάσχιση RDF Γράφου**

Το μοντέλο δεδομένων RDF αποτελεί ένα δίκτυο, του οποίου τα μονοπάτια και οι ακμές κωδικοποιούν τις πληροφορίες. Στη σημασιολογική έρευνα υπάρχουν διάφοροι τρόποι της διάσχισης του δικτύου. Π.χ. από ένα δοσμένο αρχικό στιγμιότυπο του δικτύου, τοποθετούνται στιγμιότυπα με πρόσθετες σχετικές πληροφορίες. Άλλος τρόπος είναι η πλοήγηση στο επίπεδο των οντολογικών πληροφοριών μίας θεματικής περιοχής, με περιορισμό του ερωτήματος μέσω της επιλογής κλάσεων και σχέσεων, που θα χρησιμοποιηθούν στην πραγματική αναζήτηση των στιγμιότυπων. Η απλή διάσχιση του γράφου είναι επίσης ένας συνήθης τρόπος που χρησιμοποιείται για την συγκέντρωση όλων των πληροφοριών ένα ερωτήματος.

- **Αντιστοίχιση εννοιών**

Στην σημασιολογική έρευνα, η αντιστοίχιση των εννοιών με λέξεις κλειδιά είναι η επικρατέστερη μεθοδολογία για την εύρεση πληροφορίας. Ο σημαντικότερος λόγος της επικράτησης αυτής είναι ότι όλος ο όγκος της πληροφορίας που είναι προς αναζήτηση, δυνατό να κωδικοποιηθεί όλη η γνώση που υπάρχει . Επιπρόσθετα, πολλές εφαρμογές υποστηρίζουν τη μέθοδο αυτή με τη χρήση της φυσικής γλώσσας κατά την διατύπωση των ερωτημάτων, που είναι προσφιλέστερη στους χρήστες.

- **Μοτίβα γράφων**

Τα μοτίβα γράφων των λογικών γλωσσών, είναι ιδιαίτερα χρήσιμα στην σημασιολογική αναζήτηση και χρησιμοποιούνται σε πολλές διαφορετικές λειτουργίες. Συχνά χρησιμοποιούνται για να διαμορφώνουν και να κωδικοποιούν σύνθετα ερωτήματα και προσδιορίζουν τον εντοπισμό υπογράφων στο δίκτυο RDF. Σε κάποια συστήματα συνδέουν πόρους μεταξύ τους ή διαμορφώνουν σχήματα για τον εντοπισμό μονοπατιών που συνδέονται μεταξύ τους με βάση τις ονομασίες των πόρων. Επίσης, στην παρουσίαση των αποτελεσμάτων οι παράμετροι σχετικά με το που θα μεταφερθεί η πληροφορία, δίνεται με απλά μοτίβα γράφων.

- **Λογική**

Η λογική και ο συμπερασμός συνδέονται με το μεγαλύτερο όραμα του σημασιολογικού ιστού. Ωστόσο, λίγες εφαρμογές κάνουν πλήρη χρήση του συμπερασμού ή κάποιου άλλου συστήματος κανόνων, για δύο κυρίως λόγους. Το σημασιολογικό διαδίκτυο σχεδιάστηκε για να λειτουργεί σε μία ανοιχτή παραδοχή του κόσμου, ενώ η λογική λειτουργεί πολύ καλά σε έναν κλειστό κόσμο. Επίσης ο σημασιολογικός ιστός προϋποθέτει μια μεγάλη ποσότητα δεδομένων, που αποτελεί πρόβλημα για τους περισσότερους τρέχοντες αλγορίθμους συμπερασμού.

- **Fuzzy Logic**

Η χρήση της Fuzzy Logic στο σημασιολογικό διαδίκτυο έχει στόχο την αναπαράσταση και αιτιολόγηση των γλωσσών του Σημασιολογικού Ιστού (SWL), οι οποίες χρησιμοποιούνται για να παρέχουν την περιγραφή των εννοιών, των όρων και των σχέσεων για τον σχηματισμό των μεταδεδομένων. Οι σημασιολογικές γλώσσες που έχουν ήδη περιγραφεί, δηλαδή οι RDF, RDFS, OWL και RIF χρησιμοποιούν για την σημασιολογική περιγραφή διάφορους τομείς της λογικής θεωρίας π.χ. τον λογικό προγραμματισμό και την περιγραφική λογική.

4.6 Αποθήκευση Δεδομένων

Το ζήτημα της αποθήκευσης των μεταδεδομένων στον σημασιολογικό ιστό είναι μια πολύπλοκη δραστηριότητα που έχει απασχολήσει ένα σημαντικό και ευρύ τομέα της έρευνας τα τελευταία χρόνια, με αποτέλεσμα την δημιουργία ενός αριθμού μεθοδολογιών, τεχνολογιών και εργαλείων.

Οι προσεγγίσεις για την αποθήκευση των δεδομένων RDF μπορεί να χωριστεί σε δύο κατηγορίες: η πρώτη βασίζεται σε ένα εσωτερικό σύστημα αποθήκευσης (native storage system) ενώ η δεύτερη στη χρήση σχεσιακών βάσεων δεδομένων (database-based storage).

Τα εσωτερικά συστήματα της πρώτης κατηγορίας χρησιμοποιούν τις δικές τους βάσεις δεδομένων, χωρίς να κάνουν χρήση των λειτουργιών αποθήκευσης και ανάκτησης άλλων συστημάτων διαχείρισης βάσεων δεδομένων. Συγκρίνοντας τις δύο κατηγορίες, τα εσωτερικά συστήματα είναι αποδοτικότερα από τα συστήματα βάσεων δεδομένων όσον αφορά τον χρόνο ενημέρωσης και φόρτωσης, αλλά μειονεκτούν στην διαδικασία διενέργειας επερωτήσεων καθώς διαθέτουν λιγότερα χαρακτηριστικά διαχείρισης των ερωτημάτων[60],[62].

4.6.1 Εσωτερικά συστήματα αποθήκευσης RDF

- **AllegroGraph**

Πρόκειται για μία εμπορική βάση δεδομένων RDF γράφων και ένα πλαίσιο εφαρμογής για την αποθήκευση και επερώτηση RDF δεδομένων. Εφαρμόζεται ως αυτόνομος server και προσφέρει διεπαφές για απομακρυσμένη πρόσβαση, όταν η επικοινωνία μεταξύ του server και του client πραγματοποιείται μέσω του HTTP. Οι

δυνατότητες του συστήματος περιλαμβάνουν την αποθήκευση γράφων που αναπαριστώνται από κόμβους, ακμές και δεδομένα, αλλά και πρόσθετα χαρακτηριστικά όπως χωρικό και χρονικό συλλογισμό και ανάλυση κοινωνικών δικτύων.

- **OWLIM**

Το σύστημα OWLIM παρέχεται από την Ontotext σε δύο εκδόσεις: την SwiftOWLIM που έχει σχεδιαστεί για μεσαίου όγκου δεδομένα και εκτελεί στην κύρια μνήμη τον συλλογισμό και την αξιολόγηση του ερωτήματος και την BigOWLIM που είναι σχεδιασμένη για μεγάλου όγκου δεδομένα και μέσω δεικτών που χρησιμοποιεί μπορεί να διαχειρίζεται δισεκατομμύρια τρίπτυχα RDF. Και οι δύο εκδόσεις δεν παρέχουν ειδικό μηχανισμό επεκτασιμότητας, αλλά επιτρέπουν τον ορισμό κανόνων για την εξαγωγή λογικών συμπερασμάτων. Η έκδοση BigOWLIM, χρησιμοποιείται σε μεγάλο αριθμό εφαρμογών του σημασιολογικού ιστού και συνδεδεμένων δεδομένων και γενικά η OWLIM διατίθεται για το σύστημα Sesame ως επίπεδο SAIL (Storage and Inference Layer).

- **4Store & 5Store**

Το σύστημα 4Store είναι μία RDF βάση δεδομένων που έχει σχεδιαστεί από την Garlik Inc για συστήματα όπως είναι τα Unix, Linux, Mac OS. Το 5Store είναι το τελευταίο RDF σύστημα αποθήκευσης που ανέπτυξε η Garlik για εμπορική λειτουργία και διαθέτει παρόμοια χαρακτηριστικά με το 4Store, αλλά με βελτιωμένες δυνατότητες επεκτασιμότητας και απόδοσης.

4.6.2 Αποθηκευτικά συστήματα Σχεσιακών Βάσεων Δεδομένων

Τα συστήματα Jena και Oracle είναι από τα πιο γνωστά σχεσιακά συστήματα, που εστιάζουν στην φυσική οργάνωση των δηλώσεων RDF σχετικά με τους πόρους του διαδικτύου. Όπως ξέρουμε, οι δηλώσεις που χρησιμοποιούν τα μοντέλα RDF, έχουν την μορφή των τρίπτυχων {υποκείμενο, κατηγόρημα, αντικείμενο} και ως εκ τούτου μπορούν εύκολα να υλοποιηθούν ως σχεσιακοί πίνακες τριών στηλών.

- **JENA**

Το σύστημα Jena είναι ένα εργαλείο σημασιολογικού ιστού της Java, που για την διαχείριση των τρίπτυχων παρέχει μία εφαρμογή API και για την μόνιμη αποθήκευση τους διάφορα συστήματα όπως Sleepycat/Berkeley, MySQL, Oracle, Interbase και άλλα. Η εφαρμογή Jena έχει τη δυνατότητα αποθήκευσης στη μνήμη μεγάλου

αριθμού τρίπτυχων και οι πολλές επιλογές αποθήκευσης που διαθέτει τις επιτρέπει να εφαρμόζεται σε πολλά συστήματα.

- **Oracle**

Το σύστημα Oracle χρησιμοποιεί ένα νέο τύπο αντικειμένου, που έχει δημιουργηθεί πάνω στην κορυφή του δικτύου μοντέλου δεδομένων NDM (Network Data Model) του Oracle Spatial, για την αποθήκευση, διαχείριση και ανάλυση δικτύων ή γράφων της βάσης δεδομένων. Σε αυτή την προσέγγιση τα τρίπτυχα αναλύονται και φυλάσσονται στο δίκτυο RDF μία φορά και επαναχρησιμοποιούνται όταν χρειαστεί, με αποτέλεσμα να ελαχιστοποιείται ο όγκος της αποθήκευσης. Ωστόσο, το μειονέκτημα της προσέγγισης οφείλεται στην πολυπλοκότητα του μοντέλου στην διαχείριση της RDF πληροφορίας καθώς ο χρήστης θα πρέπει να εκτελεί όλες τις λειτουργίες των χαμηλότερων επιπέδων.

4.6.3 Υβριδικά συστήματα αποθήκευσης

Τα συστήματα που υποστηρίζουν και τις δύο προηγούμενες αρχιτεκτονικές μεθόδους των εσωτερικών και σχεσιακών συστημάτων ονομάζονται υβριδικά συστήματα. Παραδείγματα της προσέγγισης αυτής είναι τα συστήματα Sesame και Virtuoso.

- **SESAME**

Το σύστημα Sesame είναι μια RDF βάση δομένων ανοιχτού κώδικα για την υποστήριξη της εξαγωγής συμπερασμάτων από τα σχήματα RDF Schema και την αναζήτηση πληροφοριών. Υποστηρίζει δύο προσεγγίσεις αποθήκευσης, το αυστηρά σχεσιακό σχήμα και το σχεσιακό σχήμα αντικειμένου. Στο αυστηρά σχεσιακό σχήμα οι βασικές ιδιότητες που περιγράφονται από το RDF μετατρέπονται σε πίνακες της βάσης δεδομένων. Κάθε πόρος και λεκτική σταθερά κωδικοποιείται με ένα id, με σκοπό την μείωση χώρου στην βάση δεδομένων. Επίσης στον πίνακα υπάρχει η προσθήκη της στήλης “is_derived”, με την πληροφορία για την προέλευση της δήλωσης, εφόσον αυτή προέκυψε από τις πληροφορίες του σχήματος. Η προσέγγιση αυτή είναι αποδοτική όταν το σχήμα RDF τροποποιείται συχνά, γιατί διατηρεί ένα σταθερό σχήμα των database πινάκων.

Η δεύτερη προσέγγιση, είναι το σχεσιακό σχήμα αντικειμένου η οποία προκαλεί χαμηλή απόδοση του συστήματος, όταν το RDF schema αλλάζει συχνά. Η αιτία είναι ότι σε κάθε προσθήκη κάποιας νέας κλάσης ή ιδιότητας το σύστημα βάσης δεδομένων, δημιουργεί εκ νέου έναν πίνακα.

- **Virtuoso**

Το υβριδικό σύστημα OpenLink Virtuoso Universal Server, αποτελεί μία αποθηκευτική επιλογή για ένα ευρύ φάσμα μοντέλων δεδομένων, συμπεριλαμβανομένων των αδόμητων δεδομένων και των σχεσιακών δεδομένων RDF και XML. Μέσω της ενοποιημένης αποθήκευσης, μπορεί να χρησιμεύσει ως ένα εργαλείο ολοκλήρωσης για δεδομένα που προέρχονται από ετερογενείς πηγές. Είναι ένα σύστημα ανοιχτού κώδικα που συγκεντρώνει σημαντικό ενδιαφέρον καθώς χρησιμοποιείται για να φιλοξενήσει σημαντικά σύνολα συνδεδεμένων δεδομένων, όπως π.χ της DBpedia.

5. Επεξεργασία Κειμένων, Ερωτημάτων και Αποτελεσμάτων

5.1 Πρόλογος

Κατά την διαδικασία της σημασιολογικής έρευνας και εξερεύνησης, ένα στάδιο νευραλγικής σημασίας είναι η επεξεργασία των σημασιολογικών δεδομένων. Η αντιμετώπιση του ζητήματος στο κεφάλαιο αυτό, περιλαμβάνει ως επιμέρους θέματα την εξερεύνηση του περιεχομένου των κειμένων, την ερμηνεία και την κατανόηση του περιεχομένου των αιτημάτων των χρηστών καθώς και την παρουσίαση και οργάνωση των αποτελεσμάτων από την μηχανή αναζήτησης.

5.2 Εξερεύνηση Κειμένων

Σε αντίθεση με το κλασσικό μοντέλο ανάκτησης πληροφορίας που χρησιμοποιούνται μη δομημένα δεδομένα, ο στόχος στο σημασιολογικό μοντέλο, όπως ήδη έχουμε πει, είναι η σύλληψη του περιεχομένου των κειμένων από την πλευρά των οντοτήτων, των εννοιών και των συσχετίσεων. Συγκεκριμένα, η εξαγωγή της γνώσης αφορά την αναγνώριση και ονοματολογία των οντοτήτων, την επιλογή των ετικετών και την εξαγωγή σχέσεων. Στην συνέχεια περιγράφονται οι διαδικασίες της ανίχνευσης εννοιών, οντοτήτων και σχέσεων στα κείμενα [36].

Ανίχνευση εννοιών στα κείμενα.

Τα κείμενα αναλύονται για την εξαγωγή των εννοιών που περιέχουν και την χρησιμοποίησή τους στην μοντελοποίηση των κειμένων ως εννοιολογικών γράφων. Για την μετάφραση και την αναπαράσταση κειμένων είναι ιδιαίτερα συχνή η χρήση εννοιών που προέρχονται από θησαυρούς όπως είναι π.χ. Wordnet και UMLS. Η χρήση τους έγκειται στην αναγνώριση όρων με πολλές έννοιες και στην αντιστοίχιση των συνώνυμων λέξεων με μια έννοια. Π.χ. στο σύστημα C-Search [56], όπου τα κείμενα αναπαριστώνται ως έννοιες ανάλογες του WordNet, οι συντάκτες για να ονομάσουν τις λέξεις και να συμπεράνουν την λογική αμφισημία τους, χρησιμοποιούν τις πληροφορίες για τα μέρη του λόγου (part of speech - POS), το WordNet, ως λεξιλογική βάση και τις σχέσεις που υπάρχουν σε επίπεδο λέξεων. Ενώ οι λέξεις αντιστοιχίζονται σε ατομικές έννοιες, οι φράσεις αναπαριστώνται ως σύνθετες έννοιες χρησιμοποιώντας φόρμουλες λογικής περιγραφής.

Ανίχνευση οντοτήτων και σχέσεων στα κείμενα.

Για την ανίχνευση οντοτήτων και σχέσεων απαιτείται η συντακτική ανάλυση των κειμένων που περιλαμβάνει αναγνώριση λέξεων (tokenization), εφαρμογή ετικετών POS tagging, ή ακόμα παραγωγή ολόκληρων δέντρων ανάλυσης (parse trees) από διάφορα συστήματα εξαγωγής δεδομένων. Εκτός από τις πληροφορίες που παράγονται από την συντακτική ανάλυση, χρησιμοποιούνται και διάφοροι σημασιολογικοί πόροι. Συγκεκριμένα, χρησιμοποιούνται συχνά τα λεξιλογικά μοντέλα γεωγραφικής πληροφόρησης Gazetteers. Η ανίχνευση των οντοτήτων γίνεται με ταίριασμα των λέξεων με τις καταχωρήσεις των Gazetteers λεξικών που αντιπροσωπεύουν οντότητες διαφορετικών κλάσεων. Για την αποσαφήνιση των ονομάτων των οντοτήτων που αναφέρονται στα κείμενα, πολύτιμες πληροφορίες παρέχουν τα σημασιολογικά δεδομένα και οι σχέσεις που αφορούν τις οντότητες μέσω των γραφικών παραστάσεων τους. Η διάσχιση ενός γραφήματος ξεκινά από τους κόμβους που ταιριάζουν με τις οντότητες για την εξερεύνηση και αποσαφήνιση του σημασιολογικού περιεχομένου τους.

Εκτός της άμεσης αντιστοίχισης και αποσαφήνισης, σύγχρονες τεχνικές εφαρμόζουν πιο σύνθετους κανόνες ταιριάσματος και εξόρυξης σχέσεων όπως είναι τα μοτίβα αντιστοίχισης – matching patterns. Π.χ. Mr. ? works for ? είναι ένα τέτοιο μοτίβο. Σύγχρονα προγράμματα εξόρυξης δεδομένων εκτός από μοτίβα λέξεων χρησιμοποιούν και σύνθετους συνδυασμούς χαρακτηριστικών που παράγονται από (1) τις λέξεις, όπως είναι ορθογραφικά χαρακτηριστικά π.χ. αν υπάρχουν αριθμοί ή παύλες στις λέξεις, συντακτικά χαρακτηριστικά όπως πληροφορίες POS, (2) το περιεχόμενο των λέξεων π.χ. λέξεις μαζί με τις ετικέτες τους και (3) το ήδη γνωστικό υπόβαθρο που προέρχεται από συστήματα πληροφοριών όπως το Gazetteers.

Πολλά μοτίβα και κανόνες χρησιμοποιούνται για την εξαγωγή δεδομένων και μπορούν να θεωρηθούν ως σημασιολογικοί πόροι που τροφοδοτούν ως είσοδοι τα συστήματα εξαγωγής δεδομένων. Η γνώση που αποκτάται από τις αντιστοιχίσεις αποθηκεύεται στις οντολογίες και χρησιμοποιείται ως σημασιολογικός περιορισμός ως φίλτρα αποτελεσμάτων π.χ. σε αυτά που οδηγούν σε σημασιολογικές ασυνέπειες. Π. χ. έχουμε το μοτίβο ο x είναι δήμαρχος του y " ?x is mayor of ?y " και την γνώση ότι το x είναι άτομο και το y πόλη (x, type, Person και y, type, City). Στα αποτελέσματα που παράγονται μπορούμε να αποκλείσουμε αυτά που δεν περιλαμβάνουν σχέσεις του " is mayor of " με ένα πρόσωπο ή μία πόλη.

5.3 Ερμηνεία ερωτημάτων

Η κατανόηση των ερωτημάτων αποτελεί παρόμοια διαδικασία με την κατανόηση κειμένων. Η διαφορά τους έγκειται στο ότι η διατύπωση των ερωτημάτων γίνεται εκτός από τη φυσική γλώσσα και από λέξεις κλειδιά. Κατά συνέπεια τα συντακτικά χαρακτηριστικά που μπορούν να ληφθούν κατά την ερμηνεία των ερωτημάτων είναι λιγότερα, από αυτά που εξάγονται κατά την ερμηνεία του περιεχόμενου των κειμένων στη φυσική γλώσσα.

5.3.1 Ανίχνευση εννοιών σε ερωτήματα λέξεων - κλειδιών

Το πρόβλημα στην ανίχνευση εννοιών στις λέξεις κλειδιά, σχετίζεται με την εύρεση της αμφισημίας τους. Όπως και στην ερμηνεία κειμένων, που περιγράφηκε στην προηγούμενη ενότητα, γίνεται η χρήση θησαυρών για την ερμηνεία και την αποσαφήνιση των λέξεων κλειδιών, χωρίς όμως την χρήση συντακτικών πληροφοριών, όπως αυτή των ετικετών POS.
Εκτός από την χρήση θησαυρών, προτάθηκε μία προσέγγιση που αναφέρεται ως εννοιολογική μοντελοποίηση ερωτήματος (conceptual query modeling) και βασίζεται στην ιδέα της συνάφειας για την αναγνώριση των εννοιών του ερωτήματος. Η τεχνική αυτή μετά την εκτέλεση του ερωτήματος ανακτά τα πιο σχετικά κείμενα. Οι σχολιασμοί των εννοιών που σχετίζονται με αυτά τα κείμενα χρησιμοποιούνται για την κατασκευή της εννοιολογικής αναπαράστασης του μοντέλου συνάφειας.

5.3.2 Ερμηνεία Ερωτημάτων διατυπωμένων με χρήση λέξεων - κλειδιών

Υπάρχουν διάφορες προσεγγίσεις που έχουν προταθεί για την αντιστοίχιση ερωτημάτων λέξεων κλειδιών με πλήρη δομημένα ερωτήματα. Στην ερμηνεία ερωτημάτων, αναγνωρίζονται εκτός από τις έννοιες, οι οντότητες και οι σχέσεις στις οποίες μπορεί να αναφέρεται στο ερώτημα. Η πρόθεση του ερωτήματος συνάγεται από το σχήμα (έννοιες και σχέσεις) και τα στοιχεία των δεδομένων (οντότητες). Σε περιπτώσεις που οι σχέσεις δεν προσδιορίζονται ρητά στο ερώτημα ή δεν επαρκούν ώστε να συνδέσουν τις έννοιες και τις οντότητες που έχουν ανιχνευτεί, χρησιμοποιούνται αλγόριθμοι διάσχισης γραφημάτων για την εύρεση της σύνδεσης στα μονοπάτια (π.χ. έννοιες, σχέσεις και οντότητες αντιστοιχίζονται σε

κατηγορούμενα και σταθερές του ερωτήματος). Η ερμηνεία των ερωτημάτων λαμβάνεται με την συγχώνευση αυτών των μονοπατιών που τελικά ταξινομούνται και αντιστοιχίζονται σε δομημένα ερωτήματα. Όπως έχουμε ήδη πει, η γλώσσα που χρησιμοποιείται για την δημιουργία σημασιολογικών ερωτημάτων (συγκεκριμένα για την έκφραση των μοτίβων γραφημάτων) στην RDF είναι η SPARQL αλλά και άλλοι φορμαλισμοί όπως εννοιολογικά γραφήματα, λογικοί τύποι ή στοιχεία XML.

Το Avatar [57] είναι ένα παράδειγμα ενός συστήματος που ερμηνεύει με αυτόν το τρόπο τα ερωτήματα που αποτελούνται από λέξεις κλειδιά. Διατηρεί ένα ευρετήριο μετάφρασης με όλες τις έννοιες για κάθε μεμονωμένη λέξη κλειδί. Το αποτέλεσμα που επιστρέφει είναι το εννοιολογικό σχήμα και το σχήμα με τα μονοπάτια που αντιστοιχίζονται στις λέξεις κλειδιά. Οι έννοιες που αντιστοιχίζει στις λέξεις κλειδιά συνδυάζονται για να απαριθμήσουν όλες τις πιθανές ερμηνείες του ερωτήματος. Για τα RDF δεδομένα, χρησιμοποιείται ο αλγόριθμος διάσχισης γραφήματος top-k, ώστε να υπολογίσει όλες τις πιθανές ερμηνείες. Κάποια άλλα συστήματα που εφαρμόζουν τον αλγόριθμο διάσχισης γραφήματος είναι τα συστήματα SeamSearch-Pro, Hermes και Tastier.

5.3.3 Ερμηνεία ερωτημάτων διατυπωμένων στη φυσική γλώσσα

Η πρώτη προσέγγιση για την ερμηνεία ερωτημάτων σε φυσική γλώσσα, είναι δημοφιλής σε συστήματα που χρησιμοποιούν ανεξάρτητους λεξιλογικούς πόρους και είναι παρόμοια με αυτή που χρησιμοποιείται για την μετάφραση ερωτημάτων με λέξεις κλειδιά. Συγκεκριμένα, περιλαμβάνει τις διαδικασίες ανίχνευσης εννοιών, οντοτήτων και σχέσεων και την αντιστοίχισή τους σε στοιχεία των δομημένων ερωτημάτων. Ισχυρές τεχνικές για την αντιστοίχιση των λέξεων στα ερωτήματα φυσικής γλώσσας με στοιχεία των σημασιολογικών δεδομένων και σημασιολογικών μοντέλων, επιλύουν τα προβλήματα της αμφισημίας, της ταξινόμησης και της αποσαφήνισης των αποτελεσμάτων αντιστοίχισης.

Η δεύτερη προσέγγιση περιλαμβάνει τη χρήση λεξικών σε συνδυασμό με εργαλεία συντακτικής ανάλυσης. Π.χ. το σύστημα FREyA [58] χρησιμοποιεί το εργαλείο Stanford Parser για τον προσδιορισμό εννοιών σε ερωτήματα φυσικής γλώσσας και το εργαλείο Gate για να αντλήσει την συντακτική ανάλυση των δέντρων. Η χρήση λεξικών συγκεκριμένου τομέα (domain-specific lexicons), παρέχουν ένα τρόπο αντιμετώπισης των αμφισημιών, ειδικά στα παραδοσιακά συστήματα ερμηνείας

φυσικής γλώσσας που εστιάζουν σε έναν τομέα. Γενικά, η χρησιμότητα των λεξικών έγκειται στην αντιστοίχιση των συντακτικών στοιχείων με τα σημασιολογικά στοιχεία. Π.χ. ένα λεξικό προσδιορίζει τις αντιστοιχίσεις μεταξύ ουσιαστικών, ρημάτων κ.τ.λ. μέσω της συντακτικής ανάλυσης και της σημασίας τους στα σημασιολογικά μοντέλα. Επίσης υπάρχουν συστήματα που ζητούν από τους χρήστες να κάνουν τις αντιστοιχίσεις σε περιπτώσεις ασαφειών και τις χρησιμοποιούν για την εκπαίδευση των μοντέλων που χρησιμοποιούνται.

5.4 Αντιστοίχιση (Matching)

Η αναζήτηση μέσω λέξεων-κλειδιών, όπως έχει ήδη αναφερθεί, είναι ο ευρύτερα διαδεδομένος μηχανισμός για την εξαγωγή πληροφορίας από το διαδίκτυο. Τα ερωτήματα που υποβάλλουν οι χρήστες υφίστανται διαφορετική επεξεργασία από τις μηχανές αναζήτησης, σύμφωνα με τον τύπο των δεδομένων που αναζητούν. Υπάρχουν τρεις διαφορετικοί τύποι δεδομένων. Τα μη δομημένα δεδομένα, τα δεδομένα δηλαδή που υπάρχουν σε ελεύθερα κείμενα ή αλλιώς bag-of-words, τα δομημένα δεδομένα δηλαδή τα δεδομένα που υπάρχουν σε βάσεις δεδομένων και τα ημι-δομημένα, δηλαδή δομημένα δεδομένα που έχουν ενσωματωθεί σε κείμενα. Οι παραδοσιακές μηχανές αναζήτησης για την επεξεργασία ερωτημάτων με λέξεις-κλειδιά χρησιμοποιούν το αντεστραμμένο ευρετήριο (inverted index), που είναι αποτελεσματικό για αδόμητα δεδομένα, όχι όμως για δομημένα ή ημι-δομημένα δεδομένα[59]. Το αντεστραμμένο ευρετήριο είναι μία λίστα με τους όρους αναζήτησης που στον καθένα από αυτούς αντιστοιχίζονται τα έγγραφα που την περιέχουν, η θέση τους και η συχνότητα εμφάνισής τους μέσα στα έγγραφα αυτά.

keyword →{<doc1, pos, score, ...>,
<doc2, pos, score, ...>, ...}

Για την βελτιστοποίηση της χρήσης του αντεστραμμένου ευρετηρίου συνδυάζεται η εφαρμογή του αλγόριθμου top-k, μέσω του οποίου ελαχιστοποιείται το εύρος των αποτελεσμάτων που επιστρέφονται στο αίτημα του χρήστη. Με είσοδο το ερώτημα του χρήστη και το αντεστραμμένο ευρετήριο με το σκορ της συχνότητας εμφάνισης του όρου, το σύστημα επιστρέφει ως έξοδο, τις τιμές ομοιότητας μεταξύ της ερώτησης και των κειμένων. Έπεται η ταξινόμηση των ανακτημένων κειμένων με βάση τις τιμές αυτές που υπολογίστηκαν, ώστε στη λίστα που επιστρέφεται να

συμπεριλαμβάνονται τα k έγγραφα που συγκεντρώνουν τον μεγαλύτερο βαθμό ομοιότητας σε φθίνουσα σειρά.

Στις σημασιολογικές μηχανές αναζήτησης, η χρήση του αντεστραμμένου ευρετηρίου δεν είναι αποτελεσματική λόγω της ανεπάρκειάς του να εντοπίσει απαντήσεις με σύνθετη δομημένη πληροφορία που έχουν τα XML κείμενα και οι σχεσιακές βάσεις δεδομένων. Οι μηχανές αναζήτησης νέας γενιάς απαιτούν την ικανότητα ενσωμάτωσης πληροφορίας από συλλογές με ετερογενή δεδομένα. Για τον λόγο αυτό, αντί της χρήσης των αντεστραμμένων ευρετηρίων τα αδόμητα, ημιδομημένα ή δομημένα δεδομένα μοντελοποιούνται ως γραφήματα που έχουν ως κόμβους τους κείμενα, πλειάδες ή στοιχεία και ως ακμές συνδέσμους, σχέσεις γονέα παιδιού ή σχέσεις γονέα παιδιού αντίστοιχα. Τα ερωτήματα μέσω λέξεων κλειδιών απαντώνται αποτελεσματικά καθώς τα ετερογενή δεδομένα συγκεντρώνονται και ομαδοποιούνται σε γραφήματα δημιουργώντας γραφήματα δεικτών.

5.4.1 Αντιστοίχιση Όρων (Term Matching)

Το ταίριασμα όρων στην σημασιολογική αναζήτηση, συνδυάζει προσεγγίσεις που σχετίζονται με την σύνταξη (απόσταση) και το λεξικό. Οι μέθοδοι που είναι βασισμένοι στην σύνταξη είναι η απόσταση Levenstein ή απόσταση σύνταξης (edit distance), η απόσταση Hamming και η απόσταση Jaro-Winkler και εφαρμόζονται ώστε να ταιριάζουν συντακτικά τους όρους αναζήτησης με τους όρους του περιεχομένου του διαδικτύου.

Η απόσταση Levenstein (LD) είναι μία μετρική που υπολογίζει πόσο διαφέρουν δύο ακολουθίες αλφαριθμητικών. Η απόσταση τους είναι ο μικρότερος αριθμός επεξεργασιών που απαιτούνται για να μετασχηματίσουν το ένα αλφαριθμητικό στο άλλο. Οι επιτρεπόμενες επεξεργασίες για τον μετασχηματισμό αυτό είναι η εισαγωγή, διαγραφή, ή αντικατάσταση ενός χαρακτήρα. Για παράδειγμα αν έχουμε δύο αλφαριθμητικά s1= (dog) και s2= (dof), η απόσταση Levenstein που έχουν μεταξύ τους είναι LD(s1,s2)= 1.

Απόσταση Hamming (Hamming Distance) μεταξύ δύο αλφαριθμητικών ίσου μήκους είναι ο αριθμός των θέσεων, στις οποίες τα αντίστοιχα σύμβολα είναι διαφορετικά. Η μέθοδος αυτή δηλαδή, μετράει τον ελάχιστο αριθμό των αντικαταστάσεων που απαιτούνται για την αλλαγή ενός αλφαριθμητικού σε ένα άλλο.

Η απόσταση Jaro-Winkler μετράει την ομοιότητα μεταξύ δύο αλφαριθμητικών. Είναι σχεδιασμένη και λειτουργεί καλύτερα για μικρά αλφαριθμητικά όπως ονόματα. Η απόσταση είναι 0, όταν δεν υπάρχει ομοιότητα μεταξύ των αλφαριθμητικών και 1 όταν ταιριάζουν ακριβώς. Όσο μεγαλύτερη είναι η απόσταση τους τόσο πιο όμοια είναι τα αλφαριθμητικά.

Έπειτα από την συντακτική αντιστοίχιση ακολουθεί η σημασιολογική. Η δομή και η επίσημη σημασιολογία των μεταδεδομένων μπορεί να χρησιμοποιηθεί για την επέκταση, την δημιουργία ή την τροποποίηση του συνόλου των αποτελεσμάτων. Μπορούν να διακριθούν τρεις κατηγορίες σημασιολογικής αντιστοίχισης [36] ανάλογα με την μέθοδο που χρησιμοποιούν δηλαδή τη διάσχιση γράφου (graph traversal), τη χρήση θησαυρών για την γνώση της σημασίας των όρων και των μεταξύ τους σχέσεων και την εξαγωγή συμπερασμάτων με βάση τη σημασιολογία των RDF, RDFS και OWL.

Διάσχιση Γράφου: Η διάσχιση του γράφου δεδομένων είναι απαραίτητη όταν τα μοτίβα ερωτήματος έχουν ελλιπείς συνδέσεις. Όταν δηλαδή, οι μηχανές αναζήτησης δεν μπορούν να συνδυάσουν απευθείας τα αποτελέσματα που βασίζονται σε μεταβλητές join, αλλά χρειάζεται να διερευνήσουν διαφορετικές μεταξύ τους διαδρομές. Επίσης μερικά συστήματα χρησιμοποιούν τους μηχανισμούς διάσχισης γράφου για την αξιολόγηση ερωτημάτων με λέξεις κλειδιά και την μετατροπή τους σε δομημένα ερωτήματα. Παραλείπουν τον υπολογισμό των ερωτημάτων και εξερευνούν απευθείας τους υπογράφους για αποτελέσματα που αντιστοιχούν με τις λέξεις κλειδιά. Τα αποτελέσματα που παράγονται δεν αναφέρονται σε μεμονωμένα κείμενα αλλά σε αρκετές οντότητες που συνδέονται μέσω μονοπατιών σε σημασιολογικά γραφήματα δεδομένων.

Χρήση Θησαυρών:

Παρόμοια με την ερμηνεία κειμένων, οι μέθοδοι ταιριάσματος όρων λαμβάνουν υπ' όψιν τους την σημασιολογία βασισμένοι σε διάφορους θησαυρούς και λεξικά όπως είναι οι οντολογίες, ταξινομήσεις, λεξικά όμοιων λέξεων ή συστήματα μετάφρασης μνήμης δίνοντας βαρύτητα στις σχέσεις των όρων και στην σημασία τους. Π.χ. στο ταίριασμα του όρου εικόνα αναζητείται και ο όρος φωτογραφία ως συνώνυμο του.

Ταίριασμα με πλήρως δομημένα μοτίβα RDF και OWL:

Η αντιστοίχιση αυτή, είναι εφικτή όταν πρόκειται για δομημένα ερωτήματα και κείμενα με σημασιολογικά δεδομένα. Δηλαδή τα ερωτήματα και τα δεδομένα

αναπαριστώνται με όρους οντοτήτων και με τις σχέσεις τους, που συνήθως είναι ερωτήματα SPARQL και δεδομένα RDF. Η πληροφοριακή ανάγκη καλύπτεται μέσω του βασικού μοτίβου γραφήματος (BGP- Basic Graph Pattern) της SPARQL που διαμορφώνει ένα γράφο όπου έχει κόμβους και ακμές με σταθερές ή μεταβλητές. Μέσω των μεταβλητών, οι μηχανές αναζήτησης συνδυάζουν τα ενδιάμεσα αποτελέσματα που λαμβάνουν για υποτμήματα των ερωτημάτων, για τα τρίπτυχα μοτίβα του BGP γράφου και μπορούν να υπολογίσουν τα τελικά αποτελέσματα. Η αντιστοίχιση του BGP γράφου με τα σημασιολογικά δεδομένα ολοκληρώνει την αντιστοίχιση των μοτίβου του γράφου, στο οποίο οι σταθερές αναπαριστούν οντότητες, σχέσεις και ιδιότητες και οι τιμές των ιδιοτήτων αντιστοιχίζονται με τον RDF γράφο για τον εντοπισμό υπογράφων που περιέχουν δεσμούς με τις μεταβλητές στο BGP γράφο.

Επιπλέον στην αντιστοίχιση δομημένων ερωτημάτων, μπορούν να χρησιμοποιηθούν οι οντολογίες που αναπαριστώνται στην γλώσσα OWL. Συγκεκριμένα, χρησιμοποιούνται στην συλλογιστική για να εξετασθούν εκτός από τα διαθέσιμα δεδομένα, αυτά που μπορούν να συναχθούν αυτόματα από την επίσημη σημασιολογία.

5.5 Ταξινόμηση (Ranking)

Οι προσεγγίσεις που αφορούν την κατάταξη των αποτελεσμάτων στα ερωτήματα των χρηστών λαμβάνουν υπ' όψιν τους διάφορους παράγοντες. Σε ένα γενικό πλαίσιο, οι προσεγγίσεις αυτές είτε είναι εξαρτώμενες από το ερώτημα, δηλαδή μελετούν την συνάφεια των αποτελεσμάτων με το ερώτημα, είτε είναι μη εξαρτώμενες από το ερώτημα και λαμβάνουν υπ' όψιν παραμέτρους όπως η δημοτικότητα και η σπανιότητα εμφάνισης των όρων, η αξιοπιστία των πηγών, η προβλεψιμότητα με βάση το περιεχόμενο, η κεντρικότητα, η εγγύτητα κ.α. Παρακάτω θα περιγραφούν τρεις δημοφιλείς κατηγορίες προσεγγίσεων [36], [37]: η ταξινόμηση με βάση την κεντρικότητα, η ταξινόμηση με βάση την εγγύτητα και η ταξινόμηση με βάση την συνάφεια του ερωτήματος.

5.5.1 Αλγόριθμοι ταξινόμησης με βάση την κεντρικότητα (Centrality - Based Rank)

Οι αλγόριθμοι ταξινόμησης που ανήκουν σε αυτή την κατηγορία, είναι οι αλγόριθμοι που χρησιμοποιούνται για τον υπολογισμό της κεντρικότητας των όρων στο κείμενο, μέσω της ανάλυσης των γράφων και των μεταξύ τους συνδέσμων. Οι αλγόριθμοι OntoRank, ObjectRank, PopRank, TripleRAnk και EntityRank αποτελούν επέκταση των δύο βασικών αλγορίθμων που χρησιμοποιήθηκαν στον παγκόσμιο ιστό, των PageRank και HITS. Η βασική διαφορά έγκειται ότι στον σημασιολογικό ιστό οι αλγόριθμοι διαχειρίζονται οντότητες αντί για κείμενα, που συνδέονται με διαφορετικούς τύπους σημασιολογικών συνδέσμων.
Ειδικότερα, ο αλγόριθμος OntoRank μιμείται την πλοηγική συμπεριφορά του χρήστη και παρέχει ποιότητα ανάλογη με αυτή του PageRank. Η ταξινόμηση αφορά τις οντολογίες τις οποίες μπορεί και χειρίζεται και ο ίδιος ο PageRank. Ο ObjectRank έχει δημιουργηθεί και αυτός πάνω στον PageRank, με την διαφορά ότι δεν χρησιμοποιεί την ίδια αρχή ροής για όλες τις ακμές, αλλά ένα γράφο μεταφοράς ώστε να αποτυπώνει τους διαφορετικούς τύπους των ακμών. Οι αλγόριθμοι PopRank και TripleRank ακολουθούν την ίδια λειτουργία με τον ObjectRank, αλλά και οι τρεις αλγόριθμοι διαφοροποιούνται στο τρόπο υπολογισμού των βαρών των ακμών τους. Τέλος ο EntityRank, εκτελείται σε ένα γράφο που περιλαμβάνει ως κόμβους οντότητες αλλά και κείμενα.

5.5.2 Αλγόριθμοι ταξινόμησης με βάση την εγγύτητα

Οι αλγόριθμοι αυτής της κατηγορίας, ταξινομούν στις πρώτες θέσης της κατάταξης τα αποτελέσματα που περιέχουν τους όρους του ερωτήματος σε κοντινές θέσεις μέσα στο κείμενο. Στα σημασιολογικά δεδομένα ο συγκεκριμένος αλγόριθμος εφαρμόζεται αναζητώντας την απόσταση μεταξύ των σημασιολογικών οντοτήτων. Τα αποτελέσματα είναι υπογράφοι που περιέχουν τους κόμβους που έχουν αντιστοιχιστεί στα ερωτήματα. Επίσης, η κατάταξη με βάση την εγγύτητα εφαρμόζεται στα XML κείμενα, όπου ο XRank αλγόριθμος κατατάσσει σε υψηλή προτεραιότητα, τα μικρότερα κείμενα που περιέχουν τις περισσότερες αντιστοιχίσεις των όρων που αναζητούνται.

5.5.3 Αλγόριθμοι Ταξινόμησης με βάση την συνάφεια

Ένας πολύ δημοφιλής αλγόριθμος ταξινόμησης, είναι ο αλγόριθμος που κατατάσσει τα αποτελέσματα με βάση τον βαθμό της συνάφειας τους με το ερώτημα. Ο βαθμός αυτός υπολογίζεται με την μετρική TF-IDF για την εύρεση των βαρών των όρων του ερωτήματος. Ειδικότερα, το TF (Term Frequency), υπολογίζει την συχνότητα ενός όρου στο κείμενο και το IDF (Inverse Document Frequency), είναι ένας δείκτης της σπουδαιότητας του όρου στο κειμένου, σε σχέση με τα υπόλοιπα κείμενα της συλλογής. Ο υπολογισμός της μετρικής αυτής, χρησιμεύει στο να κατανέμει ίσα τα βάρη στις λέξεις, αποφεύγοντας να αναθέτει μεγάλα βάρη στις λέξεις που συναντώνται συχνά μέσα στο κείμενο. Στην σημασιολογική αναζήτηση, η χρήση της μετρικής TF-IDF θα πρέπει να υπολογίζει εκτός της συχνότητας των όρων, την συχνότητα τόσο των οντοτήτων όσο και των σχόλιων. Μία ακόμα λειτουργία ταξινόμησης που χρησιμοποιείται στην ανάκτηση πληροφορίας και υπολογίζει τον βαθμό της συνάφειας των όρων, είναι η BM25F. Στην σημασιολογική έρευνα, η επέκταση του αλγόριθμου βασίζεται στην ιδέα της χρησιμοποίησης διαφορετικών πεδίων για την ευρετηρίαση διαφορετικών ιδιοτήτων των οντοτήτων RDF. Τα πεδία αυτά θα έχουν διαφορετικά βάρη, ώστε να αναγνωρίζονται οι ιδιότητες με την μεγαλύτερη σπουδαιότητα κατά την ταξινόμηση των αποτελεσμάτων.

6. Σημασιολογικές μηχανές Αναζήτησης στην πράξη

6.1 Εισαγωγή

Μια σημασιολογική μηχανή αναζήτησης καλύπτει τις πληροφοριακές ανάγκες των χρηστών μέσω ενός αποτελεσματικού μηχανισμού πρόσβασης σε δημοσιευμένα σημασιολογικά δεδομένα. Οι στόχοι της επικεντρώνονται στην συλλογή του διαθέσιμου σημασιολογικού περιεχόμενου, στην ανάλυση του για την εξαγωγή χρήσιμων ευρετηρίων και μεταδεδομένων και στην παροχή εφαρμογών για την αποτελεσματική διαχείριση των ερωτημάτων που προσπελαύνουν τα δεδομένα. Κάθε μηχανή αναζήτησης αντιμετωπίζει και υλοποιεί τις εργασίες αυτές με διαφορετική στρατηγική. Οι προσεγγίσεις που αναπτύχθηκαν κατατάσσουν τις μηχανές αναζήτησης σε τρεις κατηγορίες. (1) Προσεγγίσεις βασισμένες σε δομημένες γλώσσες ερωτημάτων (π.χ. Swoogle), (2) Προσεγγίσεις βασισμένες σε λέξεις κλειδιά (π.χ. Falcons, SWSE, Sig.Ma) και (3) βασισμένες στη φυσική γλώσσα (π.χ. Power Aqua).

Το κύριο μέρος αυτού του κεφαλαίου καλύπτεται από παραδείγματα των σημαντικότερων μηχανών αναζήτησης, όπως είναι οι Swoogle, Watson, Sindice, Sig.Ma, Falcons και SWSE καθώς και κάποιες σημασιολογικές εφαρμογές που εφαρμόζονται από διάφορες σημασιολογικές μηχανές όπως οι PowerAqua, PowerMagpie κ.α.

6.2 Πληροφοριακές ανάγκες

Ανεξάρτητα από την μεθοδολογία τους, όλες οι σημασιολογικές μηχανές αναζήτησης έχουν ως κοινό στόχο την ικανοποίηση τόσο απλών ερωτημάτων όσο και πιο πολύπλοκων πληροφοριακών αναγκών. Υπάρχουν διαφορετικά είδη αναζητήσεων που διενεργούνται για την κάλυψη των διαφορετικών αναγκών πληροφόρησης, τα οποία είναι [36]:

Αναζήτηση Οντοτήτων: Κατά την αναζήτηση αυτή, αντικείμενο των ερωτημάτων είναι οντότητες (π.χ. πρόσωπα ή εταιρείες) και τα αποτελέσματα που εμφανίζουν απαιτούν επιπλέον περιήγηση ή πλοήγηση για να καλύψουν τις πληροφοριακές ανάγκες. Στον σημασιολογικό ιστό, ενσωματώνονται σημασιολογικά δεδομένα σε

ιστοσελίδες και τα αποτελέσματα που επιστρέφονται περιλαμβάνουν αποσπάσματα με περιγραφές των οντοτήτων.

Αναζήτηση πραγματικών γεγονότων: Η πληροφοριακή ανάγκη αυτών των αιτημάτων απαιτεί την ανάκτηση συγκεκριμένων γεγονότων.

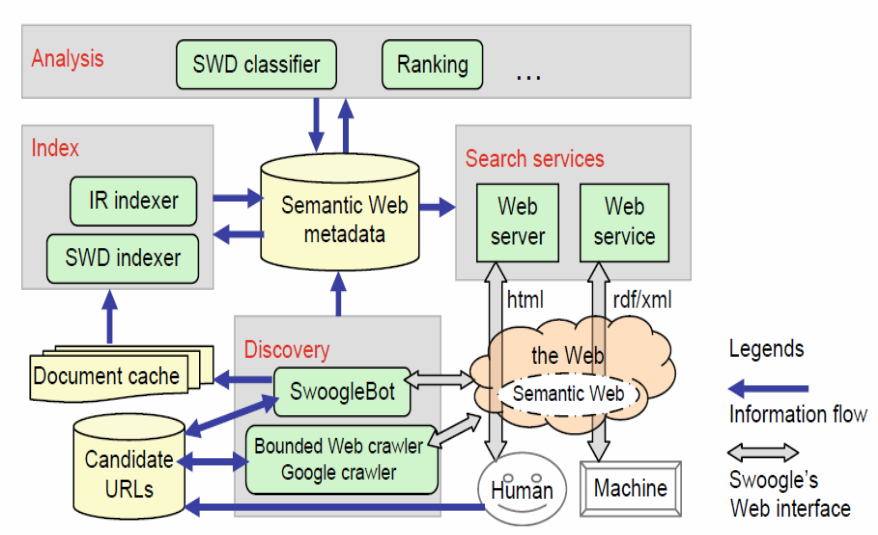
Σχεσιακή Αναζήτηση: Ο τύπος των ερωτημάτων στην σχεσιακή αναζήτηση περιλαμβάνει οντότητες και τις μεταξύ τους σχέσεις. Για το λόγο αυτό, η επεξεργασία τους απαιτεί την κατανόηση των οντοτήτων με τις σχέσεις που τα συνδέουν.

Αναλυτική Αναζήτηση: Πέρα από τις αναζητήσεις σε οντότητες και σχέσεις, υπάρχουν πληροφοριακές ανάγκες που για να επιτευχθούν απαιτούν την διενέργεια επιπρόσθετων αναλύσεων πάνω στα δεδομένα. Στην αναλυτική αναζήτηση εφαρμόζονται διάφορα μαθηματικά και στατιστικά μοντέλα και κάποιες μορφές της μηχανικής μάθησης βασισμένες στην επαγωγική λογική.

6.3 Σημασιολογική Μηχανή Αναζήτησης SWOOGLE

Στη συνέχεια αναλύεται ο σχεδιασμός της αρχιτεκτονικής που υλοποιεί η σημασιολογική μηχανή αναζήτησης Swoogle [34], [64]. Στην αρχιτεκτονική αυτή, η Swoogle προσπαθεί να ανταποκριθεί στο αίτημα του χρήστη, ανιχνεύοντας αρχικά το διαδίκτυο για την εύρεση σημασιολογικών κειμένων (SWDs) και σημασιολογικών όρων (SWTs) και κατόπιν συγκεντρώνει τα σημασιολογικά δεδομένα για την δημιουργία ευρετηρίου. Ο πράκτορας υποβάλλει το ερώτημα με λέξη κλειδί “person” και ενημερώνει την προτεινόμενη URL αναφορά (URLref) -foaf:Person. Στη συνέχεια συνθέτει ένα ερώτημα SPARQL χρησιμοποιώντας τις ανακτημένες URLref και ζητά από την υπηρεσία εύρεσης κειμένων της Swoogle, URL με σημασιολογικά κείμενα σχετικά με το ερώτημα SPARQL που έθεσε. Με τα SWDs των URLs δημιουργεί μία αποθήκη από τρίπτυχα και μέσω των ενσωματωμένων δεδομένων που συνθέτουν τα τρίπτυχα, απαντάει στο ερώτημα SPARQL.

Οι διαδικασίες της ανίχνευσης, ευρετηρίασης, ταξινόμησης και αναζήτησης στην Swoogle είναι παρόμοιες με των συμβατικών μηχανών αναζήτησης. Στην παρακάτω εικόνα εμφανίζεται η δομή της αρχιτεκτονικής αυτής.



Εικόνα 6.3 - Αρχιτεκτονική Swoogle

6.3.1 Ανίχνευση – Crawling

Στο στάδιο της ανίχνευσης η Swoogle, χρησιμοποιεί ένα υβριδικό ανιχνευτή, ώστε να έχει μεγαλύτερη αποτελεσματικότητα στην εύρεση και συγκέντρωση SWDs. Η ροή εργασιών του ανιχνευτή περιλαμβάνει διάφορες διαδικασίες. Για την διευκόλυνση της εύρεσης των URL, παρέχονται οι URL σπόροι που έχουν ανιχνευτεί από τον meta-crawler της μηχανής Google (Google based meta-crawling). Επιπρόσθετα ο ανιχνευτής περιορίζει την ανίχνευση HTML κειμένων (bounded HTML crawling) και επιβάλλει κάποια όρια όσον αφορά π.χ. το βάθος της ανίχνευσης στο διαδίκτυο, των μέγιστο αριθμό των επισκέψιμων URL και το ελάχιστο ποσοστό των SWDs στα επισκέψιμα URLs, με σκοπό να μειώσει το χώρο αναζήτησης και να εξασφαλίσει μεγαλύτερη αποδοτικότητα. Η ανίχνευση RDF βελτιώνει την ανίχνευση HTML κειμένων, εξάγοντας υπερσυνδέσμους με σημασιολογικό περιεχόμενο, κάνοντας επισκέψεις σε νέες σελίδες που ανακαλύπτει ή σελίδες που έχει ήδη επισκεφτεί για να διατηρήσει ενημερωμένα τα μεταδεδομένα που υπάρχουν. Για κάθε URL, φορτώνει το περιεχόμενο του και το αναλύει σε RDF γραφήματα μέσω RDF αναλυτών (π.χ. Jena). Στο τελευταίο στάδιο της ανίχνευσης, λαμβάνεται από τα SWDs και επαληθεύεται από τον ανιχνευτή RDF, ένα δείγμα συνόλου δεδομένων. Με βάση τα χαρακτηριστικά του (π.χ. URL, πόρους ιστότοπου κ.τ.λ.) και τις επισημάνσεις του (π.χ. σημασιολογικό κείμενο, μερικώς σημασιολογικό ή καθόλου σημασιολογικό),

παράγονται νέοι σπόροι που προορίζονται για την ανίχνευση της Google και την ανίχνευση για τον περιορισμό των HTML.

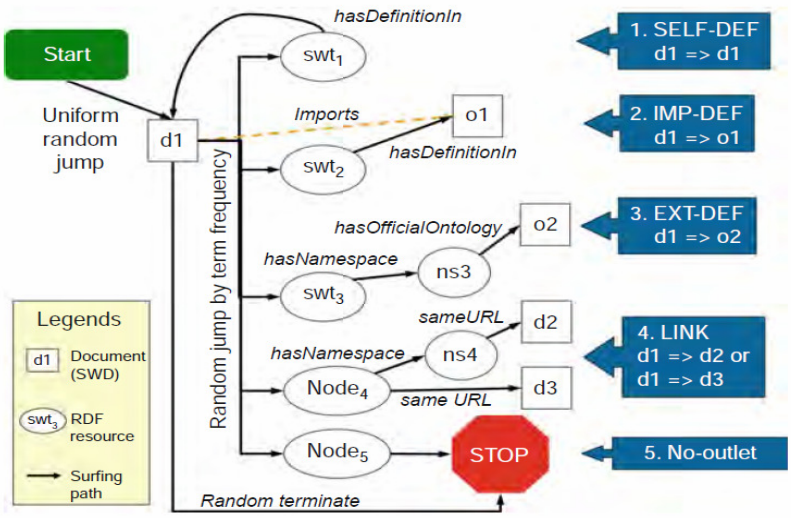
6.3.2 Ευρετηρίαση (Indexing)

Κατά την ευρετηρίαση, αναλύονται τα SWDs και με βάση αυτά παράγεται ο κυριότερος όγκος των μεταδεδομένων του σημασιολογικού διαδικτύου. Τα μεταδεδομένα περιλαμβάνουν τους χαρακτηρισμούς σημασιολογικό κείμενο SWD ή σημασιολογικός όρος SWT και περιγράφουν τις μεταξύ τους σχέσεις.

6.3.4 Ταξινόμηση (Ranking)

Η ταξινόμηση με βάση τον αλγόριθμο Page Rank όπως εφαρμόστηκε από την μηχανή αναζήτησης Google, δεν μπορεί να υιοθετηθεί στο σημασιολογικό διαδίκτυο, εξαιτίας του γεγονότος ότι η περιήγηση βασίζεται στην λογική και όχι στην τυχαία υπερσύνδεση των κειμένων, που απαιτεί την κατανόηση του σημασιολογικού περιεχομένου των σελίδων. Συνεπώς στη θέση του μοντέλου περιήγησης με τυχαίο τρόπο, βρίσκεται ένα μοντέλο περιήγησης λογικής σύνδεσης. Με βάση το μοντέλο αυτό ο πράκτορας μεταφέρεται τυχαία σε ένα από τα SWDs κείμενα με ίση πιθανότητα. Η περιήγηση τερματίζει είτε με ίση πιθανότητα είτε επιλέγοντας έναν RDF κόμβο στο RDF γράφημα του κειμένου με τον κόμβο να έχει επιλεχθεί με βάση την συχνότητα του όρου στην N-τρίπτυχη εκδοχή του κειμένου. Ο πράκτορας είτε περιηγείται σε έναν άλλο κόμβο ή τερματίζει σύμφωνα με την σημασιολογία του επιλεγμένου κόμβου.

Στο σχήμα απεικονίζεται ο τρόπος με τον οποίο περιηγείται ο πράκτορας λογισμικού. Στα μονοπάτια 1, 2, 3 ο πράκτορας επιδιώκει να βρει έναν ορισμό. Στην περίπτωση που ο κόμβος δεν είναι ανώνυμος και χρησιμοποιείται ως κλάση ή ιδιότητα στο κείμενο, ο πράκτορας επιδιώκει περαιτέρω ορισμό βάση του κειμένου, των οντολογιών ή των namespaces των URI κόμβων. Στη συνέχεια, αν ο κόμβος δεν είναι ανώνυμος, κλάση ή ιδιότητα, η περιήγηση ακολουθεί την διεύθυνση URL που προέρχεται από το URI ή το namespace για να μεταβεί σε ένα άλλο σημασιολογικό κείμενο.



Εικόνα 6.3.4 - Η λειτουργία του πράκτορα λογισμικού στην Swoogle

Το μονοπάτι 5 (No-outlet), περιλαμβάνει τις περιπτώσεις στις οποίες δεν υπάρχει περαιτέρω μονοπάτι που μπορεί να μεταβεί είτε, για παράδειγμα, επειδή ο παρόν κόμβος είναι ανώνυμος, είτε επειδή οδηγεί σε ένα κανονικό κείμενο χωρίς σημασιολογικό περιεχόμενο.

6.3.5 Ανάκτηση (Retrieval)

Όσον αφορά την ανάκτηση πληροφοριών, η Swoogle παρέχει υπηρεσίες τόσο για πράκτορες λογισμικού όσο και για χρήστες, χρησιμοποιώντας τα μεταδεδομένα του ευρετηρίου της. Τα αποτελέσματα ενός σημασιολογικού ερωτήματος μπορούν να είναι εκτός από κείμενα και URI με σημασιολογικούς όρους (κλάσεις, ιδιότητες). Οι υπηρεσίες αναζήτησης των οντολογιών ή σημασιολογικών κειμένων και των σημασιολογικών όρων γίνεται με την χρήση λέξεων κλειδιών, με τη δυνατότητα χρήσης επιπλέον περιορισμών στο ερώτημα.

6.3.6 Αρχείο (Archive)

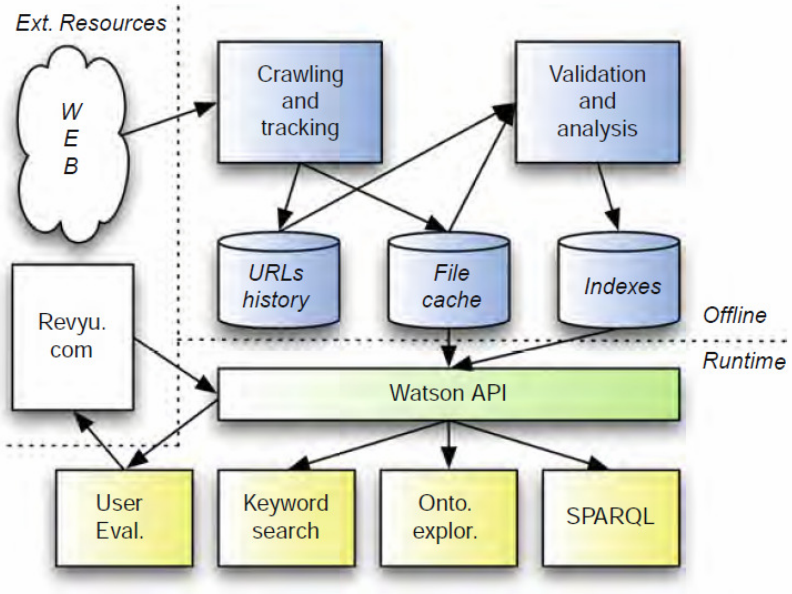
Η μηχανή αναζήτησης Swoogle, διατηρεί:
-ένα χώρο αποθήκευσης για τα διαθέσιμα σημασιολογικά κείμενα,
-αντίγραφα των κειμένων με την αναπαράσταση τους ως σύνολα τρίπτυχων

-ένα αρχείο με όλες τις τρέχουσες και παλιές εκδόσεις των σημασιολογικών κειμένων που διατίθενται στο ευρετήριο της. Το αρχείο μπορεί να χρησιμοποιηθεί για την έρευνα της εξέλιξης των οντολογιών, της αύξησης των RDF δεδομένων και της ανακάλυψης του φυσικού κύκλου ζωής των σημασιολογικών κειμένων.

6.4 Σημασιολογική Μηχανή Αναζήτησης WATSON

Η μηχανή αναζήτησης Watson [34], [49] είναι μία πύλη που συλλέγει, αναλύει και παρέχει πρόσβαση σε οντολογίες και σημασιολογικά δεδομένα που υπάρχουν στο διαδίκτυο. Στόχος της είναι να υποστηρίξει την ανάπτυξη των εφαρμογών σημασιολογικού ιστού νέας γενιάς που συλλέγουν, συνδυάζουν και εκμεταλλεύονται την γνώση δυναμικά.

Η Watson βασίζεται σε πρότυπες, ανοιχτές τεχνολογίες που εμφανίζονται στην εικόνα που ακολουθεί. Για τον εντοπισμό των σημασιολογικών κειμένων, χρησιμοποιείται ένα στοιχείο ανίχνευσης και εντοπισμού, ο ανιχνευτής αρχείων διαδικτύου Heritrix. Τα στοιχεία ανάλυσης και επικύρωσης των αρχείων αυτών, δημιουργούν ένα εξελιγμένο σύστημα ευρετηρίων, μέσω του συστήματος ευρετηρίασης Apache Lucene. Με βάση αυτά τα ευρετήρια, έχει αναπτυχθεί η εφαρμογή API για την παροχή λειτουργιών αναζήτησης, εξερεύνησης και εκμετάλλευσης των συλλεγμένων σημασιολογικών κειμένων.



Εικόνα 6.4 - Αρχιτεκτονική Σημασιολογικής Μηχανής Αναζήτησης Watson

6.4.1 Ανίχνευση (Crawling)

Για την ανίχνευση και τον εντοπισμό των σημασιολογικών κειμένων χρησιμοποιούνται διάφορες πηγές (Google, Swoogle, http://pingthesemanticweb.com, κ.α.) και έχουν σχεδιαστεί ειδικοί ανιχνευτές. Μετά την ανάκτηση, τα κείμενα φιλτράρονται και διατηρούνται μόνο τα σημασιολογικά στοιχεία, εξαλείφοντας όλα τα στοιχεία που δεν μπορούν να αναλυθούν από την εφαρμογή Jena. Συνεπώς διατηρούνται μόνο τα κείμενα που είναι βασισμένα σε δεδομένα RDF.

6.4.2 Διαχείριση περιεχομένου

Από την συλλογή των σημασιολογικών κειμένων, εξάγονται διαφορετικού τύπου πληροφορίες, που έχουν σχέση με τις οντότητες, το κυριολεκτικό περιεχόμενο των κειμένων, τις σχέσεις που υπάρχουν μεταξύ τους κ.α. Για το λόγο αυτό, απαιτείται ανάλυση του περιεχομένου των κειμένων για την λήψη μεταδεδομένων και την χρήση τους από την μηχανή Watson. Συνεπώς, αποτελεί σημαντικό βήμα ο χαρακτηρισμός των όρων που περιέχονται στα κείμενα. Η Watson εξάγει, εκμεταλλεύεται και αποθηκεύει ένα μεγάλο εύρος των δηλωμένων μεταδεδομένων (RDF, RDFS, OWL), πληροφορίες σχετικά με τις οντότητες που περιλαμβάνονται (κλάσεις, ιδιότητες) ή μετρήσεις σχετικά με το πλήθος των γνώσεων που περιέχονται στο κείμενο. Συνδυάζοντας αυτά τα στοιχεία με τις πληροφορίες, η μηχανή Watson αποφασίζει αν τα συγκεκριμένα κείμενα θα πρέπει να αντιμετωπιστούν ως πλούσιες σημασιολογικές οντότητες. Έπειτα αποθηκεύονται και χρησιμοποιούνται για να παρέχουν προηγμένο, ποιοτικά ελεγμένο, ταξινομημένο και αναλυμένο σημασιολογικό περιεχόμενο. Επιπλέον, ενδιαφέρεται για τις σχέσεις μεταξύ των σημασιολογικών κειμένων. Για το λόγο αυτό εξετάζονται οι ανακτημένες οντολογίες, με σκοπό να εξαχθούν πληροφορίες που συνδέονται με τα σημασιολογικά κείμενα ή να ανιχνευτούν πιθανές τοποθεσίες άλλων σημασιολογικών κειμένων.

Ο τρόπος διεξαγωγής ερωτημάτων των χρηστών είναι παρόμοιος με άλλα συνήθη συστήματα αναζήτησης. Οι λέξεις κλειδιά αντιστοιχίζονται με ετικέτες, ονόματα, σχόλια ή την κυριολεκτική σημασία των οντοτήτων των σημασιολογικών κειμένων και παραθέτεται μια λίστα με τις οντολογίες που ταιριάζουν με τα ερωτήματα. Για την διευκόλυνση της επιλογής των οντολογιών, είναι χρήσιμη η δημιουργία περιλήψεων ώστε να διαβάζονται και να κατανοούνται εύκολα. Με κάθε

σημασιολογικό κείμενο υπάρχουν βασικές πληροφορίες που αφορούν το μέγεθος (σε bytes, αριθμό κλάσεων, ιδιοτήτων, τρίπτυχων), την γλώσσα που έχει χρησιμοποιηθεί (OWL, RDF-S), οι σύνδεσμοι με άλλα κείμενα και οι αναφορές από τους χρήστες της εφαρμογής. Στον server Watson έχει αναπτυχθεί ένα SPARQL endpoint και είναι προσαρμόσιμο στο σημασιολογικό κείμενο που θα υποβληθεί το ερώτημα. Μία απλή διεπαφή επιτρέπει την εισαγωγή και την εκτέλεση ενός SPARQL ερωτήματος στο επιλεγμένο σημασιολογικό κείμενο.

6.4.3 Η εφαρμογή Watson API

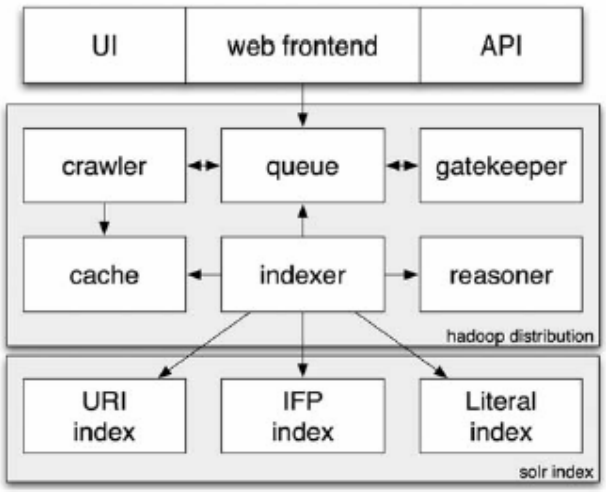
Τα βασικά στοιχεία της μηχανής Watson είναι οι υπηρεσίες και η εφαρμογή API [50] που παρέχουν πολλά πλεονεκτήματα όπως:

- Τον εντοπισμό σημασιολογικών κειμένων μέσω εξελιγμένων αναζητήσεων με λέξεις κλειδιά, που επιτρέπουν στις εφαρμογές να προσδιορίσουν τα ερωτήματα σύμφωνα με τον αριθμό παραμέτρων όπως είναι ο τύπος των οντοτήτων, το επίπεδο αντιστοίχισης με τις λέξεις κλειδιά κτλ
- Την ανάκτηση μεταδεδομένων που υπάρχουν στα κείμενα όπως είναι το μέγεθος, η γλώσσα, η λογική πολυπλοκότητα κτλ
- Τον εντοπισμό συγκεκριμένων οντοτήτων μέσα σε ένα κείμενο όπως κλάσεις και ιδιότητες
- Τον έλεγχο του περιεχομένου του κειμένου όπως είναι η σημασιολογική περιγραφή των οντοτήτων που περιέχει
- Την εφαρμογή ερωτημάτων SPARQL

Το ολοκληρωμένο σύνολο των λειτουργιών της API, επιτρέπει σε κάθε εφαρμογή να χρησιμοποιήσει σε απευθείας σύνδεση τα σημασιολογικά δεδομένα, χωρίς να απαιτείται η λήψη των κειμένων. Επίσης οι εφαρμογές δεν χρειάζεται να διαθέτουν εξειδικευμένους μηχανισμούς και μεγάλους πόρους για να προσπελάσουν τα ευρετήρια της μηχανής Watson, να επιλέξουν και να εκμεταλλευτούν απευθείας τους σημασιολογικούς πόρους. Επιπρόσθετα, χρησιμοποιείται ένα σύνολο λειτουργιών που υπολογίζει την πολυπλοκότητα και το εύρος των οντολογιών και χρησιμεύει στην ταξινόμηση, όπως επίσης και μηχανισμοί που υποστηρίζουν την ανίχνευση των συσχετισμών μεταξύ των οντολογιών. Ωστόσο θα πρέπει να δημιουργηθούν και άλλοι πιο προηγμένοι μηχανισμοί που θα εξετάζουν και άλλου είδους συσχετίσεις όπως η έκδοση, η επέκταση και η συμβατότητα.

6.5 Σημασιολογική Μηχανή Αναζήτησης SINDICE

Πρόκειται για ένα ευρετήριο σημασιολογικού ιστού ή μια υπηρεσία αναζήτησης οντοτήτων, που επικεντρώνεται στην συγκέντρωση μεγάλων ποσοτήτων δεδομένων. Η λειτουργία της μηχανής στην ουσία, είναι να συλλέγει RDF κείμενα από το σημασιολογικό ιστό και να δημιουργεί ευρετήρια χρησιμοποιώντας URIs, Inverse Functional Properties (IFPs) και λέξεις κλειδιά. Ο σχεδιασμός της έχει σκοπό να αποτελέσει μια υπηρεσία που χρησιμοποιείται από αποκεντρωμένες εφαρμογές πελάτη για τον εντοπισμό σχετικών πηγών δεδομένων. Τα αποτελέσματα που βρίσκει ταξινομούνται με σειρά σχετικότητας. Η προσέγγιση στην ευρετηρίαση των RDF εγγράφων γίνεται μέσω τεχνικών ανάκτησης πληροφορίας, όπου οι κυριολεκτικές λέξεις και τα αναγνωριστικά, καταχωρούνται σε ευρετήρια για να γίνουν αναζητήσεις πάνω σε αυτά και επιστρέφει δείκτες στους πόρους που αναφέρονται σε αυτούς τους όρους. Η αρχιτεκτονική της μηχανής περιγράφεται στο ακόλουθο σχήμα[51].



Εικόνα 6.5 - Αρχιτεκτονική Σημασιολογικής Μηχανής Αναζήτησης Sindice

Η Sindice προσφέρει τρεις υπηρεσίες στις εφαρμογές πελάτη. Αναλύει τα αρχεία και τα SPARQL endpoint κατά την διάρκεια της ανίχνευσης, αναζητά πόρους και επιστρέφει τα URL των κειμένων RDF που εμφανίζονται οι πόροι αυτοί και τέλος αναζητά πλήρη κείμενα και επιστρέφει τα URL των πόρων.

Επιπλέον στοχεύει στην μείωση του μεγέθους του ευρετηρίου, στην μείωση των χρόνων αναζήτησης και στην συνεχή ενημέρωση του ευρετηρίου.

Αναλυτικότερα, η αρχιτεκτονική αποτελείται από αρκετά ανεξάρτητα στοιχεία που λειτουργούν σε διάφορους διαύλους ώστε να επιτευχθεί η ανίχνευση, η ευρετηρίαση και η υποβολή ερωτήσεων. Το στάδιο web frontend, που φαίνεται στο σχήμα, είναι το κύριο σημείο εισόδου που διαιρείται σε μία διεπαφή χρήστη για την πρόσβαση των χρηστών στο σύστημα και μία API για την πρόσβαση της μηχανής. Επίσης υπάρχουν αρκετά στοιχεία για την ανίχνευση και την ευρετηρίαση των κειμένων RDF. Ο ανιχνευτής κάνει αυτόνομα την συγκομιδή των RDF δεδομένων από τον ιστό και τα προσθέτει στην ουρά. Ο gatekeeper αξιολογεί κάθε είσοδο στην ουρά και αποφασίζει ποια θα είναι η προτεραιότητα που θα γίνει η καταχώρηση στο ευρετήριο, με βάση κριτήρια όπως αν έχουμε ξαναδεί το συγκεκριμένο κείμενο, την τελευταία ημερομηνία τροποποίησής του, το περιεχόμενό του κτλ. Ο indexer εξάγει URIs, IFPs και literal από κάθε έγγραφο και τα προσθέτει στο αντίστοιχο ευρετήριο. Κατά την διάρκεια της αναζήτησης, τα στοιχεία της διεπαφής πρέπει να δώσουν τα ερωτήματα στο αντίστοιχο ευρετήριο, να συλλέξουν τα αποτελέσματα και να παράγουν τις απαιτούμενες εξόδους, όπως σελίδες HTML, με την κατάλληλη διάταξη.

6.5.1 Υπηρεσία ευρετηρίασης και επερωτήσεων

Η υπηρεσία της Sindice καταχωρεί στο ευρετήριο RDF γραφήματα και μετά επιτρέπει στους χρήστες ή στις εφαρμογές του σημασιολογικού ιστού, να εντοπίσουν τις τοποθεσίες των πόρων μέσω επερωτήσεων. Οι εργασίες αυτές γίνονται μέσω ενός αγωγού ευρετηρίασης και ενός αγωγού επερωτήσεων.

Ο αγωγός ευρετηρίασης εκτελεί μια σειρά από εργασίες για την ευρετηρίαση των γραφημάτων RDF. Αρχικά ο Scheduler λαμβάνει ως είσοδο δεδομένα RSS που προσδιορίζουν την τοποθεσία του RDF γραφήματος. Το URL που αντιστοιχεί στην τοποθεσία του γραφήματος καταχωρείται στον προγραμματιστή που λειτουργεί ως συλλέκτης URLs, τα οποία αναμένουν να υποβληθούν σε επεξεργασία, ώστε να αποφευχθεί η υπερφόρτωση των Web servers. Ο scheduler διατηρεί ένα μικρό hashtable με μεταδεδομένα για κάθε πόρο που επισκέπτεται (π.χ. το χρόνο επίσκεψης). Η επίσκεψη στους ίδιους πόρους πραγματοποιείται εφόσον επέλθει ένα όριο χρόνου αναμονής και η νέα ανάλυση των πόρων επαναλαμβάνεται μόνο αν έχει αλλάξει το περιεχόμενο του hashtable.

Τα URLs στον scheduler μπορούν να είναι δύο ειδών: RDF πόροι και SPARQL endpoints. Για την εξαγωγή των γραφημάτων στους RDF πόρους, ανακτάται το

αρχείο και έπειτα στέλνεται στον αναλυτή. Ο αναλυτής αρχικά επαληθεύει την εγκυρότητα του RDF γραφήματος. Στην συνέχει εξάγει όλα τα URIs από το γράφημα και τα τοποθετεί στον scheduler, ακολουθώντας την αρχή των συνδεδεμένων δεδομένων ώστε κάθε URI να είναι ένας υπερσύνδεσμος σε επιπλέον πληροφορίες. Στην περίπτωση των SPARQL endpoints στέλνεται στην βάση δεδομένων ένα ή περισσότερα ερωτήματα για την εξαγωγή του πλήρους περιεχομένου του.

Για την επεξεργασία των γραφημάτων, ο graph processor, εξάγει και ευρετηριάζει ολόκληρο το κείμενο και τα αναγνωριστικά των πόρων στο γράφημα. Επίσης εξάγει έμμεσα αναγνωριστικά, σε ζεύγη (p,o), για όλες τις αρχές συμπερασμού IFP (Inverse Functional Properties - Αντίστροφες Λειτουργικές Ιδιότητες) που αναφέρονται στο γράφημα. Για την εξαγωγή των IFP απαιτείται η επέκταση του γραφήματος με όλες τις πληροφορίες του σχήματος, που ακολουθείται από λογικά συμπεράσματα OWL για την αναγνώριση όλων των IFPs στα σχήματα.

Ο αγωγός επερωτήσεων, χωρίζεται σε τρία στάδια: Την ανάκτηση ευρετηρίου, την φάση ταξινόμησης και την παραγωγή αποτελεσμάτων. Κατά την ανάκτηση του ευρετηρίου, το ερώτημα αναζητείται σε ένα αντιστραμμένο ευρετήριο, το οποίο μπορεί να υλοποιηθεί είτε ως ένα hashmap είτε ως μια μηχανή ανάκτησης πληροφορίας. Ο κατάλογος με τα αποτελέσματα είναι προσωρινής αποθήκευσης και ανανεώνεται καθημερινά για να διατηρεί τα αποτελέσματα ανανεωμένα. Κατά την φάση της ταξινόμησης, τα αποτελέσματα ταξινομούνται σύμφωνα με διάφορες μετρικές. Η λειτουργία ταξινόμησης απαιτεί μόνο λίγα μεταδεδομένα για κάθε πόρο και είναι σχετικά εύκολη να υπολογιστεί. Αυτό επιτυγχάνεται κυρίως με την επιλογή hostnames ίδιων με τα hostnames των πόρων, πόρων που βρίσκονται σε υψηλές θέσεις ταξινομήσεων σε τοποθεσίες που χρησιμοποιούν παραδοσιακούς αλγορίθμους κατάταξης και πόρων που μοιράζονται σπάνια όρους.

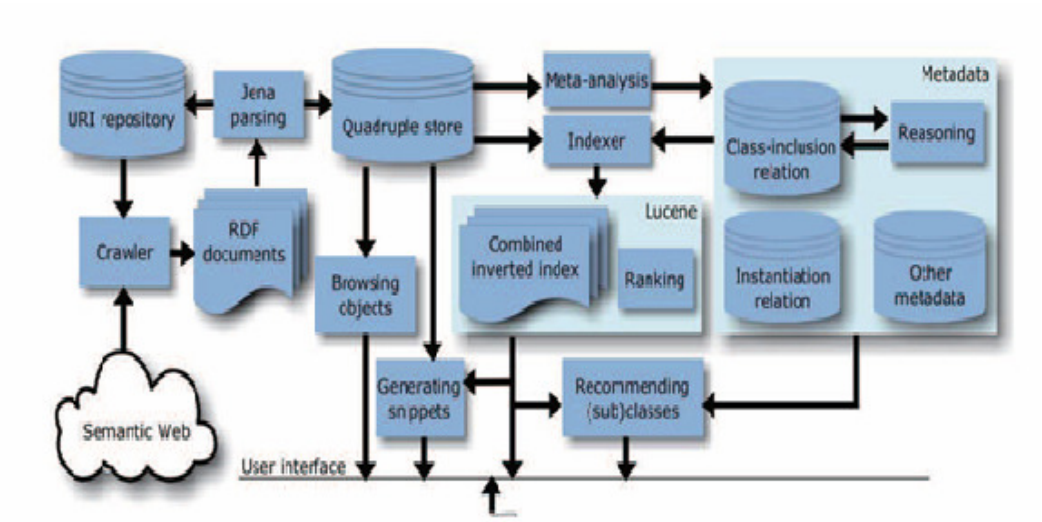
6.6 Σημασιολογική Μηχανή Αναζήτησης Sig.Ma

Η Sig.Ma είναι ένα εργαλείο που επιτρέπει στους χρήστες να εντοπίζουν RDF οντότητες, μέσω της αναζήτησης με λέξεις κλειδιά. Η Sig.Ma βασίζεται στην Sindice για να παρέχει μια συγκεντρωτική άποψη πάνω σε διαθέσιμα σημασιολογικά δεδομένα για μία οντότητα ή πηγή. Η Sig.Ma σε ένα ερώτημα λέξης κλειδιού, εμφανίζει τις ιδιότητες των οντοτήτων σε ένα ευρύ φάσμα συνδεδεμένων πηγών δεδομένων, με κάθε δεδομένο να αντιστοιχίζεται στον πόρο από τον οποίο

προέρχεται. Για παράδειγμα στο ερώτημα ενός ονόματος κάποιου προσώπου, η Sig.Ma θα προβάλλει φωτογραφίες, πληροφορίες επικοινωνίας, τοποθεσία, το χώρο εργασίας και ημερομηνία γέννησης, με κάθε πληροφορία να προέρχεται από διαφορετικό πόρο. Αξιοσημείωτο είναι επίσης ότι η Sig.Ma αποτελεί βάση για την υλοποίηση και άλλων εφαρμογών.

6.7 Σημασιολογική Μηχανή Αναζήτησης FALCONS

Η Falcon[65] είναι μία ακόμα σημασιολογική μηχανή αναζήτησης που παρέχει αναζήτηση με λέξεις κλειδιά για URIs αντικείμενα, κλάσεις, ιδιότητες και κείμενα του σημασιολογικού ιστού. Στη σελίδα των αποτελεσμάτων για κάθε αντικείμενο ή κλάση, παρέχονται ο τίτλος, το URI, οι τύποι και περιγραφές RDF. Η ανάλυση των RDF/XML κειμένων χρησιμοποιεί την εφαρμογή Jena, η αποθήκευση δεδομένων την MySQL και η ευρετηρίαση δεδομένων την Apache Lucene. Στο παρακάτω σχήμα απεικονίζεται η αρχιτεκτονική της μηχανής Falcon.

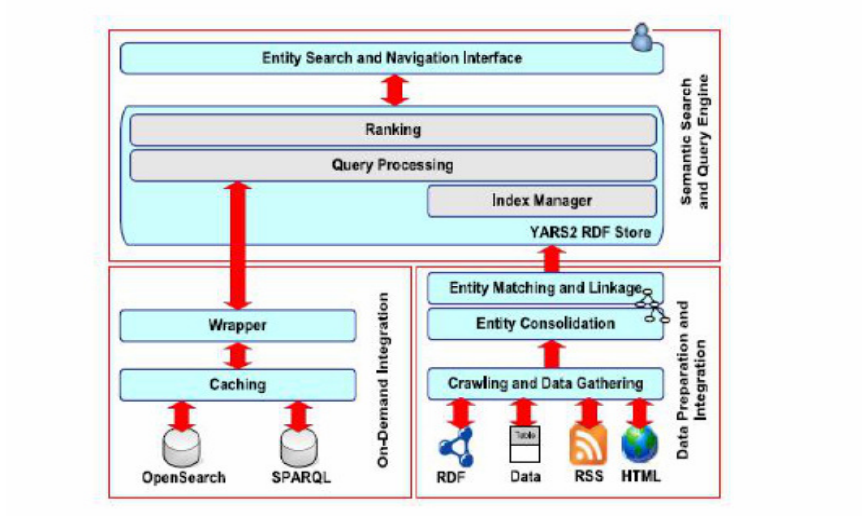


Εικόνα 6.7 - Αρχιτεκτονική Σημασιολογικής Μηχανής Falcon

6.8 Σημασιολογική μηχανή αναζήτησης SWSE

Η SWSE[65] μία σημασιολογική μηχανή αναζήτησης βασισμένη στην αναζήτηση με λέξεις-κλειδιά, η οποία έχει υιοθετήσει τον αλγόριθμο Page Rank στα σημασιολογικά δεδομένα συνδυάζοντας ταξινομήσεις από RDF γραφήματα και ταξινομήσεις από τα

γραφήματα δεδομένων των πόρων. Η υλοποίηση της μηχανής SWSE υπονοεί δύο μεγάλες ερευνητικές προκλήσεις. Το σύστημα πρέπει να συγκεντρώνει μεγάλες ποσότητες δεδομένων και να είναι ανεκτό σε ετερογενή, θορυβώδη και πιθανόν αντικρουόμενα δεδομένα που έχουν συλλεχθεί από ένα μεγάλο αριθμό πηγών. Η SWSE περιέχει στοιχεία για την ανίχνευση, ταξινόμηση και ευρετηρίαση αλλά και στοιχεία για τον χειρισμό των δεδομένων RDF. Η SWSE εκτελεί σημασιολογική ενσωμάτωση των δομημένων δεδομένων από το διαδίκτυο, πόρους με XML βάσεις δεδομένων, στατικά δεδομένα κ.α. Παρακάτω εμφανίζεται η αρχιτεκτονική της μηχανής SWSE.



Εικόνα 6.8 - Αρχιτεκτονική Σημασιολογικής Μηχανής SWSE

6.9 Σημασιολογική Μηχανή Αναζήτησης HAKIA

Η σημασιολογική μηχανή Hakia[49], προσπαθεί να προβλέψει τις ερωτήσεις που είναι σχετικές με ένα κείμενο και μπορούν να ερωτηθούν από τους χρήστες και τις χρησιμοποιεί ως πύλες για το περιεχόμενό του. Τα ερωτήματα αναζήτησης αντιστοιχίζονται με τα αποτελέσματα και κατατάσσονται, χρησιμοποιώντας έναν αλγόριθμο που τα βαθμολογεί σύμφωνα με την ανάλυση της πρότασης και το πόσο πολύ ταιριάζουν με τις έννοιες που σχετίζονται με το ερώτημα.

Η Hakia παρουσιάζει τα αποτελέσματα σε κατηγορίες που έχουν διαφορετικό περιεχόμενο, για να καλύψουν το ερώτημα. Για παράδειγμα σε ένα ερώτημα σχετικά με τον Μότσαρτ στην μηχανή Hakia, ο χρήστης λαμβάνει ως αποτέλεσμα μία σελίδα

που περιέχει την φωτογραφία του Μότσαρτ και επιπλέον στοιχεία από διάφορες κατηγορίες όπως, βιογραφίες, αποσπάσματα, μουσικό έργο, χρονολόγιο κ.α.

Ένα από τα πλεονεκτήματα της Hakia είναι η αναζήτηση σε φυσική γλώσσα. Το αποτέλεσμα που επιστρέφεται στα ερωτήματα αυτού του τύπου, είναι η ευθεία απάντηση στην ερώτηση και όχι μία λίστα με ιστοσελίδες όπου είναι πιθανό να περιέχουν την απάντηση αλλά θα πρέπει να εντοπιστεί με περαιτέρω πλοήγηση.

Η απάντηση στα ερωτήματα φυσικής γλώσσας καθίσταται εφικτή με τον συνδυασμό διαφορετικών στοιχείων των επιστημονικών κλάδων, όπως οι αρχές της φιλοσοφίας, η μαθηματική λογική, η γνωστική επιστήμη που καλείται οντολογική σημασιολογία, μία επίσημη και περιεκτική γλωσσική θεωρία της σημασίας στη φυσική γλώσσα.

Τέλος, η Hakia έχει δημιουργήσει ένα νέο σύστημα για την αναπαράσταση εννοιών, το Qdexing. Το σύστημα περιέχει την ανάλυση ολόκληρου του περιεχομένου μιας ιστοσελίδας και στη συνέχεια την εξαγωγή όλων των πιθανών ερωτημάτων που μπορούν να ζητηθούν. Τα ερωτήματα αυτά γίνονται πύλες, για τα αρχικά κείμενα, τις παραγράφους και τις προτάσεις κατά την διάρκεια της λειτουργίας ανάκτησης.

6.10 Εφαρμογές Σημασιολογικού Ιστού

Τα τελευταία χρόνια δημιουργήθηκε μια νέα γενιά εφαρμογών σημασιολογικού ιστού, οι οποίες εν αντιθέσει με τις εφαρμογές πρώτης γενιάς που περιορίζονταν στη παραγωγή και χρήση των δικών τους δεδομένων, εκμεταλλεύονται τον σημασιολογικό ιστό ως μια ετερογενή πηγή πληροφορίας ευρείας κλίμακας. Η σημαντικότερη διαφορά τους, όσον αφορά τη λειτουργία τους, είναι ότι οι νέας γενιάς εφαρμογές εξερευνούν το διαδίκτυο και εντοπίζουν τις οντολογίες εκ των προτέρων και όχι κατά τη διάρκεια του σχεδιασμού του. Οι εφαρμογές αυτές, θα πρέπει να διακρίνονται από κάποια στοιχεία:

- Την ικανότητα εύρεσης πόρων με σημασιολογικές πληροφορίες δυναμικά, διότι κατά το χρόνο σχεδιασμού οι προγραμματιστές μπορεί να μην έχουν την ικανότητα να κρίνουν μια συγκεκριμένη πηγή σχετική με το στόχο.
- Την ικανότητα επιλογής της κατάλληλης γνώσης, από το σύνολο των ήδη εντοπισμένων σημασιολογικών κειμένων με βάση κριτήρια όπως η ποιότητα και η επάρκεια του θέματος.

- Την ικανότητα να συνδυάζει οντολογίες και πόρους. Μια σημασιολογική μηχανή νέας γενιάς πρέπει να επιλέγει και να ενσωματώνει τμήματα γνώσης από διαφορετικές πηγές και να τις εκμεταλλεύεται από κοινού.

Στη συνέχεια περιγράφονται κάποιες σημαντικές εφαρμογές σημασιολογικού ιστού [34], [50].

6.10.1 Ο Browser Power Magpie

Ο browser σημασιολογικού ιστού Power Magpie, χρησιμοποιεί διαθέσιμα σημασιολογικά δεδομένα για να βοηθήσει τους χρήστες να ερμηνεύσουν το περιεχόμενο των ιστοσελίδων του διαδικτύου. Αναγνωρίζει και χρησιμοποιεί την γνώση που παρέχεται από πολλαπλές οντολογίες σε χρόνο εκτέλεσης.

Από την πλευρά των χρηστών, η Power Magpie είναι μια επέκταση του κλασσικού browser. Παρέχει αρκετές λειτουργίες που επιτρέπουν στους χρήστες να εξερευνήσουν τις σημασιολογικές πληροφορίες των τρεχόντων ιστοσελίδων. Συγκεκριμένα, συνοψίζει εννοιολογικές οντότητες σχετικές με την ιστοσελίδα, τονίζει τα σημεία τους στο κείμενο και αφήνει τους χρήστες να εξερευνήσουν την πληροφορία γύρω από αυτές με διάφορους τρόπους. Επιπρόσθετα, όταν εντοπίζει απευθείας σημασιολογικές πληροφορίες που σχετίζονται με το κείμενο, τις ενσωματώνει ως σχόλια σε RDFa. Οι χρήστες μπορούν να αποθηκεύσουν αυτούς τους σχολιασμούς σε μια τοπική βάση και να την χρησιμοποιήσει για να μεσολαβήσει στην αλληλεπίδραση των σημασιολογικών συστημάτων.

6.10.2 Σύστημα απάντησης επερωτήσεων PowerAqua

Η PowerAqua εκτελεί ερωτήσεις για έναν απεριόριστο αριθμό οντολογιών και μπορεί αυτόματα να συνδυάσει πληροφορίες από διαφορετικές οντολογίες σε χρόνο εκτέλεσης. Οι χρήστες θέτουν ένα ερώτημα σε φυσική γλώσσα και στη συνέχεια το σύστημα εντοπίζει και επιλέγει τις οντολογίες με την μηχανή Watson, επιλέγει την κατάλληλη οντολογία έπειτα από την αποσαφήνιση του ερωτήματος με την χρήση των σημασιολογικών πληροφοριών και εξάγει τις απαντήσεις σε μορφή οντολογικών οντοτήτων.

Η απόδοση των μηχανισμών της Power Aqua για την εύρεση οντολογιών και το ταίριασμά τους, είναι σημαντική, ωστόσο επηρεάζεται από την ποιότητα των σημασιολογικών δεδομένων του σημασιολογικού ιστού.

6.10.3 Σύστημα εύρεσης συσχετισμών Scarlet

Η εφαρμογή Scarlet, επιλέγει και εξερευνά online οντολογίες, με σκοπό να ανακαλύψει τις σχέσεις που υπάρχουν μεταξύ δύο δοσμένων εννοιών. Για να επιτευχθεί αυτό, αναγνωρίζει τις οντολογίες που παρέχουν πληροφορίες σχετικά με τον τρόπο που συσχετίζονται οι δύο έννοιες και έπειτα συγκρίνει αυτές τις πληροφορίες για να συναχθεί η σχέση τους.

Υπάρχουν δύο στρατηγικές που βοηθούν στον εντοπισμό των σχέσεων. Σύμφωνα με την πρώτη στρατηγική, παράγεται μία σχέση μεταξύ δύο εννοιών εάν η σχέση έχει οριστεί μέσα σε μία οντολογία. Π.χ. υπάρχει μία σχέση μεταξύ των όρων κτίριο και κατάστημα, εάν η οντολογία ορίζει ότι: κατάστημα $\subseteq$ κτίριο.

Η δεύτερη στρατηγική συνδυάζει τις πληροφορίες που εμπίπτουν σε δύο ή περισσότερες οντολογίες π.χ. Ψωμί $\subseteq$ Σιτηρά είναι η μία οντολογία, Σιτηρά $\subseteq$ Υδατάνθρακες η δεύτερη.

Για την υποστήριξη αυτής της λειτουργίας η Scarlet χρειάζεται μία πύλη σημασιολογικού ιστού ώστε να έχει πρόσβαση στις οντολογίες και για αυτό χρησιμοποιεί τις λειτουργίες της μηχανής Watson.

6.10.4 Σύστημα σημασιολογικού εμπλουτισμού των Folksonomies, Flor

Τα Folksonomies είναι συστήματα του Web 2.0 που έχουν ως βασικά στοιχεία τους χρήστες, τους πόρους και τα tags. Τα Tags μπορούν να είναι μία οποιαδήποτε ακολουθία από χαρακτήρες, που ένας χρήστης μπορεί να επισυνάψει σε έναν πόρο και χρησιμοποιούνται στην ταξινόμηση περιεχομένου, με την απουσία όμως σημασιολογικής πληροφορίας. Το σύστημα Flor, FoLksonomy Ontology enRichment, δέχεται ως είσοδο ένα σύνολο από tags και αυτόματα τα συσχετίζει με σχετικές σημασιολογικές οντολογίες (όπως κλάσεις, σχέσεις) όπως ορίζονται σε online οντολογίες.

Το σύστημα Flor εκτελεί τρία βασικά βήματα. Κατά το πρώτο βήμα λαμβάνει χώρα η λεξιλογική επεξεργασία, κατά την οποία επιλέγονται tags που έχουν σημασία και θα πρέπει να συμπεριληφθούν ως βάση στον σημασιολογικό εμπλουτισμό. Για παράδειγμα tags που περιέχουν, αριθμούς, χαρακτήρες, διαστήματα, ειδικούς χαρακτήρες κ.α. δεν έχουν επιπλέον επεξεργασία και απομονώνονται από το σύστημα. Στη συνέχεια από τα tag που απομένουν, προσπαθεί να γεφυρώσει συμβατά ονόματα που χρησιμοποιούνται σε οντολογίες, folksonomies και άλλες πηγές, παράγοντας μια λίστα από πιθανές λεξιλογικές αναπαραστάσεις για κάθε tag που μπορεί να εμπλουτιστεί.

Στο δεύτερο βήμα γίνεται η αποσαφήνιση των tags και η σημασιολογική επέκτασή τους. Εξαιτίας της πολυσημίας ένα tag μπορεί να έχει διαφορετικές σημασίες σε διαφορετικά περιεχόμενα. Π.χ. το tag jaguar μπορεί να περιγράφει ένα ζώο, ένα αυτοκίνητο ή ένα λειτουργικό σύστημα. Για την αποσαφήνιση και τον ορισμό των tag ακολουθούνται διάφορες στρατηγικές. Όσον αφορά την επέκταση της σημασιολογίας, χρησιμοποιείται ένας συνδυασμός θησαυρών και άλλων πόρων γνώσης.

Στη τελευταία φάση το σύστημα Flor αναγνωρίζει τις σημασιολογικές οντότητες, που είναι σχετικές για κάθε tag, με τη συμμετοχή των αποτελεσμάτων των προηγούμενων βημάτων. Οι σχετικές σημασιολογικές οντότητες επιλέγονται με την πύλη Watson, που παραχωρεί πρόσβαση σε όλες τις online οντολογίες και αναγνωρίζονται αυτές που ανταποκρίνονται στα tags. Στο τέλος, μέσω της εφαρμογής Scarlet, αναγνωρίζονται οι σχέσεις μεταξύ των σημασιολογικών οντοτήτων.

6.10.5 Λεξικό Οντολογίας Swoogle

Πρόκειται για μία πρόσθετη εφαρμογή στην μηχανή Swoogle, που συλλέγει όλους τους σημασιολογικούς όρους που υπάρχουν στα σημασιολογικά κείμενα, για την δημιουργία ενός λεξικού σημασιολογικού ιστού. Το λεξικό παρέχει μία συνολική άποψη του σημασιολογικού περιεχομένου στο διαδίκτυο. Εκτός από τους όρους που έχουν οριστεί ως οντολογίες, συλλέγει επίσης και όρους που έχουν αρχικά οριστεί ως κλάσεις ή ιδιότητες.

Το λεξικό οντολογίας Swoogle, προσφέρει δύο διεπαφές χρηστών για τον εντοπισμό των όρων του σημασιολογικού ιστού.

Το αλφαβητικό ευρετήριο όρων, που οργανώνει όλους τους σημασιολογικούς όρους του ιστού με βάση το αλφαβητικό τους πρόθεμα και την αναζήτηση όρων, που

επιτρέπει στους χρήστες να κάνουν αναζητήσεις στους σημασιολογικούς όρους με βάση τα URI, namespaces, κυριολεκτικές περιγραφές, τοπικά ονόματα ή σημασιολογικούς ορισμούς.

6.10.6 Επαναχρησιμοποίηση γνώσης με την εφαρμογή Watson Plug-in

Η επαναχρησιμοποίηση μιας οντολογίας είναι μια πολύπλοκη διαδικασία που περιλαμβάνει δραστηριότητες όπως η αναζήτηση σχετικών οντολογιών, η αξιολόγηση της ποιότητας της γνώσης, η επιλογή των μερών της και η ενσωμάτωση της στην υπό επεξεργασία οντολογία. Η εφαρμογή Watson plug-in στοχεύει να διευκολύνει την επαναχρησιμοποίηση της γνώσης, ενσωματώνοντας τις δυνατότητες αναζήτησης της μηχανής Watson στο περιβάλλον σύνταξης της οντολογίας. Το αποτέλεσμα της εφαρμογής επιτρέπει στον χρήστη να εκτελέσει όλα τα βήματα που είναι αναγκαία για την επαναχρησιμοποίηση της online γνώσης, μέσα στο ίδιο περιβάλλον όπου η γνώση επεξεργάστηκε και σχεδιάστηκε. Πρακτικά η εφαρμογή επιτρέπει στον προγραμματιστή της οντολογίας να εντοπίσει στις υπάρχουσες online οντότητες, τις περιγραφές των οντοτήτων που υπάρχουν στις υπό επεξεργασία οντολογίες, να ελέγξει τις περιγραφές που υπάρχουν στις οντότητες αυτές και να τις ενσωματώσει στη βάση οντολογίας. Για παράδειγμα για την επέκτασης μιας οντολογίας με δηλώσεις σχετικά με την κλάση Παπαγάλος, η εφαρμογή Watson plug-in εντοπίζει μέσω της Watson άλλες υπάρχουσες οντολογίες που περιέχουν σχετικές δηλώσεις όπως:

- Ο παπαγάλος είναι μια υποκλάση των Πτηνών
- Μακάο είναι μια υποκλάση του Παπαγάλου
- Ό παπαγάλος είναι το πεδίο ορισμού της ιδιότητας Τροπικό Κλίμα.

Με την χρήση αυτών των δηλώσεων, μπορεί να επεκταθεί η υπό επεξεργασία οντολογία για να εξασφαλίσει για παράδειγμα ότι η κλάση Μακάο είναι υποκλάση μιας νέας ενσωματωμένης κλάσης Παπαγάλος.

6.10.7 Πλαίσιο εξέλιξης οντολογιών Evolva

Η εφαρμογή Evolva είναι ένα ολοκληρωμένο πλαίσιο που έχει σχεδιαστεί για την εξέλιξη των οντολογιών. Οι οντολογίες διαμορφώνουν την βάση για τα σημασιολογικά συστήματα, αλλά για να παραμείνουν χρήσιμες θα πρέπει να

διατηρούνται ενημερωμένες. Για τον λόγο αυτό η έρευνα της τεχνολογίας σχετικά με την εξέλιξη των οντολογιών έχει αυξανόμενο ενδιαφέρον.

Η εξέλιξη των οντολογιών ορίζεται ως η έγκαιρη προσαρμογή των οντολογιών στις αλλαγές που προέκυψαν σε αυτές και η συνεπής διαχείρισή τους. Το πλαίσιο Evolva καλύπτει έναν πλήρη κύκλο εξέλιξης των οντολογιών που περιλαμβάνει την εκτέλεση και την διαχείριση των αλλαγών με βάση εξωτερικές πηγές δεδομένων. Οι πηγές αυτές μπορεί να είναι κείμενα, βάσεις δεδομένων, folksonomies ή άλλες οντολογίες. Κάθε μία από αυτές απαιτεί διαφορετική μέθοδο για την εξαγωγή του περιεχομένου και την ανακάλυψη καινούργιων πληροφοριών. Το στοιχείο επικύρωση δεδομένων - data validation αναγνωρίζει νέους όρους που είναι σχετικοί με τις οντολογίες. Επίσης ελέγχει την ποιότητα του περιεχομένου και φιλτράρει τον θόρυβο που παράγεται από το στοιχείο ανακάλυψης πληροφοριών - information discovery component. Στην επικύρωση των πληροφοριών σημαντικό ρόλο έχει το γνωστικό υπόβαθρο για την ενσωμάτωση των νέων πληροφοριών στην οντολογία. Επίσης η επικύρωση της οντολογίας – validated ontology, περνά στο στοιχείο διαχείριση της εξέλιξης – evolution management component, όπου ο χρήστης έχει τον έλεγχο πάνω στην εξέλιξη και οι αλλαγές που έγιναν, καταγράφονται και μεταδίδονται σε οντολογίες με εξάρτηση σε αυτή.

Η εφαρμογή Evolva κάνει χρήση της μηχανής Watson μέσω της εφαρμογής Scarlet, για να εντοπίσει τους εξωτερικούς πόρους της υπάρχουσας γνώσης. Σκοπός της είναι να δημιουργήσει σχέσεις μεταξύ των όρων και της γνώσης που ήδη υπάρχει στην οντολογία, παρέχοντας τα μέσα για να ενσωματώσει και τους νέους όρους στην οντολογία. Για αυτό το λόγο, διενεργείται η διαδικασία ανακάλυψης σχέσεων που συνδυάζει διάφορους πόρους από το γνωστικό υπόβαθρο, με σκοπό την βελτιστοποίηση της απόδοσης όσον αφορά τον χρόνο και την ακρίβεια.

6.10.8 SWAML

Το πρόγραμμα SWAML (Semantic Web Archive for Mailing List) εξάγει το περιεχόμενο μιας λίστας ηλεκτρονικών μηνυμάτων σε αρχείο μορφής κατάλληλης για το σημασιολογικό διαδίκτυο. Το SWAML διαβάζει μία συλλογή από ηλεκτρονικά μηνύματα που είναι αποθηκευμένα σε ένα mailbox μορφής Unix και παράγει από αυτά μία RDF περιγραφή χρησιμοποιώντας την οντολογία SIOC. Ωστόσο οι πληροφορίες του ηλεκτρονικού ταχυδρομείου δεν αρκούν για την οργάνωσή τους.

Π.χ. πληροφορίες σχετικά με πηγές που αφορούν ανθρώπους, βρίσκονται σε διάφορες πηγές στο σημασιολογικό ιστό και είναι βασισμένες στο αρχείο FOAF. Για τον λόγο αυτό χρησιμοποιείται η μηχανή αναζήτησης Sindice, ώστε να συλλέγονται σημασιολογικές πληροφορίες που σχετίζονται με τους ανθρώπους, με βάση τις διευθύνσεις email τους. Ένα από τα πλεονεκτήματα της Sindice είναι η ικανότητα εξαγωγής συμπερασμάτων IFP. Στα αρχεία FOAF, η σχέση που συνδέει ένα άτομο με τη διεύθυνση του email του, ορίζεται ως IFP, που σημαίνει ότι μια διεύθυνση σχετίζεται με ένα μόνο άτομο. Με αυτόν τον τρόπο, όταν αρκετοί πόροι συνδέονται με την ίδια διεύθυνση ηλεκτρονικού ταχυδρομείου, η μηχανή Sindice εξάγει το συμπέρασμα ότι αυτοί οι πόροι απευθύνονται στο ίδιο άτομο.

6.10.9 Επέκταση ερωτήματος Wahoo/Gowgle

Η Wahoo and Gowgle είναι δύο οδηγοί που δείχνουν πως η μηχανή Watson μπορεί να χρησιμοποιηθεί ως μια απλή εφαρμογή, για να εκτελέσει την επέκταση των ερωτημάτων σε μία κλασσική μηχανή αναζήτησης. Π.χ. σε μια λέξη-κλειδί, τα εργαλεία αυτά μπορούν να εντοπίσουν άλλους όρους που προσδιορίζουν το ερώτημα και να τους προτείνουν σε μία Web μηχανή αναζήτησης. Χωρίς την μηχανή Watson, θα ήταν απαραίτητο κάποιος να ενσωματώνει μία ή περισσότερες οντολογίες σχετικά με τον τομέα του ερωτήματος και μία υποδομή για την αποθήκευση, την διερεύνηση και την υποβολή ερωτημάτων. Αν η εξεταζόμενη μηχανή είναι μια γενική μηχανή αναζήτησης όπως η Google, η περιοχή των ερωτημάτων δεν θα μπορεί να προβλεφθεί και η κατάλληλη οντολογία θα μπορεί να επιλέγει μόνο σε χρόνο εκτέλεσης, ανάλογα με το ερώτημα που δίνεται. Επιπρόσθετα, αυτή η εφαρμογή απαιτεί βαριά υποδομή που θα είναι ικανή να χειριστεί αποδοτικά μεγάλες οντολογίες και να τους υποβάλει ερωτήματα. Οι Gowgle και Wahoo βασίζονται σε οντολογίες σημασιολογικού ιστού που διερευνούνται με την χρήση της Watson. Η συνολική αρχιτεκτονική τους έχει κατασκευαστεί σε Javascript/HTML για την είσοδο ερωτημάτων και την προβολή των αποτελεσμάτων, που επικοινωνεί με τον server της Watson χρησιμοποιώντας αρχές της AJAX. Στην περίπτωση της Gowgle, η Google χρησιμοποιείται ως μηχανή αναζήτησης και οι υπηρεσίες SOAP Web της Watson, χρησιμοποιούνται για την εξερεύνηση οντολογιών. Στην περίπτωση της Wahoo, χρησιμοποιείται η Yahoo! και οι υπηρεσίες Watson REST API. Και οι δύο εφαρμογές χρησιμοποιούν την Watson για να εκμεταλλευτούν τις online οντολογίες με σκοπό να προτείνουν όρους

σχετικούς με το ερώτημα. Αυτό συνεπάγεται πως αν το ερώτημα περιέχει τη λέξη 'αποθεραπεία' (1) εντοπίζει τις οντολογίες που αναφέρουν την έννοια της, (2) εντοπίζει στις οντολογίες αυτές τις οντότητες που αντιστοιχούν στη λέξη, και (3) επιθεωρεί τις σχέσεις αυτών των οντοτήτων για να εντοπίσει σχετικούς όρους.

Η αύξηση στον αριθμό των σημασιολογικών κειμένων οδήγησε στην αύξηση τόσο των σημασιολογικών μηχανών αναζήτησης όσο και στην αύξηση των εφαρμογών σημασιολογικών μηχανών αναζήτησης. Επιπλέον με τις μηχανές αναζήτησης που αναφέρθηκαν σε αυτή την ενότητα, υπάρχει μια μεγάλη πληθώρα μηχανών που προσφέρουν πολλές δυνατότητες ανάκτησης πληροφορίας. Κάποιες από αυτές είναι οι μηχανές η SenseBot, PowerSet, DeepDyve, Duck duck go κ.α.

Οι σημασιολογικές μηχανές αναζήτησης εκτός από εργαλεία που συλλέγουν, αναλύουν και καταχωρούν σε ευρετήρια οντολογίες, και σημασιολογικά δεδομένα μπορούν να λειτουργήσουν και ως πλατφόρμες έρευνας που υποστηρίζουν την εξερεύνηση του σημασιολογικού ιστού για την καλύτερη κατανόηση των χαρακτηριστικών του. Τα συστήματα αυτά παρέχουν στατιστικές για τα κείμενα και τις οντότητες που έχουν συλλέξει και οι ερευνητές που εμπλέκονται στην ανάπτυξη των σημασιολογικών μηχανών αναζήτησης χρησιμοποιούν τα σημασιολογικά δεδομένα και οντολογίες που είναι διαθέσιμα μέσω αυτών των συστημάτων.

Συμπεράσματα - Μελλοντικές Κατευθύνσεις

Η σημασιολογική έρευνα επικεντρώνει τις προσπάθειες της στην εξέλιξη του σημερινού τρόπου πρόσβασης στην πληροφορία και αντιμετωπίζει την ανάγκη για πληροφόρηση σε σημασιολογικό επίπεδο, λαμβάνοντας υπ' όψιν την έννοια των ερωτημάτων των χρηστών και τους διαθέσιμους πόρους. Τα τελευταία έτη, έγιναν εντυπωσιακά βήματα στην ανάπτυξη και εφαρμογή των σημασιολογικών τεχνολογιών και έχουν αναπτυχθεί πολλές μεθοδολογίες και ερευνητικές κατευθύνσεις που αφορούν όλα τα επίπεδα της διαδικασίας αναζήτησης. Οι σημασιολογικές μηχανές καθιστούν μια νέα γενιά εφαρμογών, που επιτρέπουν στους χρήστες να εξερευνήσουν την γνώση με αποτελεσματικούς τρόπους.

Παρόλη όμως την μεγάλη εξέλιξη των σημασιολογικών μηχανών αναζήτησης, πολλά θέματα έρευνας εξακολουθούν να διερευνώνται ώστε να βρεθούν νέες λύσεις που θα παρέχουν έγκυρη και ενημερωμένη γνώσης. Η επέκταση της σημασιολογικής έρευνας επικεντρώνεται στην βελτίωση των τρωτών σημείων των συστημάτων ή στην ορθή λήψη αποφάσεων για την αποτελεσματικότερη λειτουργία τους. Εν παραδείγματι, παρόλο που πολλές προσεγγίσεις και συστήματα υποστηρίζουν ερωτήματα σε φυσική γλώσσα, η έρευνα που αφορά αυτόν τον τομέα βρίσκεται σε αρχικό στάδιο και χρήζει τεχνολογικής επέκτασης και υποστήριξης. Η αυτόματη μετάφραση ενός φυσικού ερωτήματος σε οντολογικό, η αυτόματη προσθήκη σημασιολογικών σχολιασμών στο περιεχόμενο του διαδικτύου ή αυτόματη εξαγωγή γνώσης είναι μερικά κρίσιμα θέματα που πρέπει να αντιμετωπίσει η κοινότητα της σημασιολογικής έρευνας.

Ένα ακόμα θέμα έρευνας που πρέπει να διευρυνθεί, αφορά τον τρόπο δημιουργίας και διατήρησης των οντολογιών. Αν δηλαδή θα δημιουργηθούν με χειροκίνητο τρόπο από ειδικούς (π.χ. στην Wikipedia), αν θα εξαχθούν αυτόματα από το διαδίκτυο (π.χ. από ήδη υπάρχουσες οντολογίες, RDFa, mikroformats), ή αν θα δημιουργηθούν και με τους δύο τρόπους σε συνδυασμό.

Επιπλέον, κάποιοι περιορισμοί, όπως το μέγεθος και η ανομοιογένεια του διαδικτύου, ο συγκριτικά μικρός όγκος της σημασιολογικής γνώσης, η έλλειψη επίσημων κριτηρίων αξιολόγησης, είναι μερικές αιτίες που δυσκολεύουν την εξολοκλήρου εφαρμογή των ερευνητικών κατακτήσεων, στις σημασιολογικές μηχανές αναζήτησης.

Η υλοποίηση των σημασιολογικών μηχανών αναζήτησης αποτελούν προϊόν πολυετούς ερευνητικής μελέτης και τεχνολογικής υποστήριξης. Παρά το γεγονός της απαιτούμενης μελλοντικής δουλειάς για την ολοκλήρωση και βελτίωση των λειτουργιών τους, έχουν κατορθώσει να παρέχουν έναν μεγάλο βαθμό ποιότητας και αποδοτικότητας στις υπηρεσίες τους, αποδεικνύοντας ότι προσεγγίζουν τον τελικό τους στόχο, την εξέλιξη τους σε μηχανές ανθρώπινης σκέψης.

Βιβλιογραφία

[1] Google Annual Search Statistics.
http://www.statisticbrain.com/google-searches

[2] Junaidah Mohamed Kassim, Mahathir Rahmany " Introduction to Semantic Search Engines ".

[3] History of Search Engines: From 1945 to Google today. http://www.searchenginehistory.com

[4] ΩΡΙΩΝ, Βιβλιοθήκη Αλεξάνδρειου Τεχνολογικού Εκπαιδευτικού Ιδρύματος, Μετά-μηχανές, http://orion.lib.teithe.gr/index.php?page=web-metasearch

[5] Certified Knowledge
http://certifiedknowledge.org/blog/goto-to-overture-to-ysm-timeline/

[6] How do search Engine works?
http://pingnatcha.blogspot.gr/2011/10/how-do-search-engines-work.html

[7] Πύλη για την ελληνική γλώσσα: Σημασιολογία (Semantics)
http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/glossology/show.html?id=25

[8] Wikipedia, The free Encyclopedia
http://en.wikipedia.org/wiki/Semantic_search

[9] Thanh Tran, Daniel M. Herzig, Gunter Ladwig "SemSearchPro – Using Semantics throughout the Search Process" – linguistic model conceptual model

[10] Media Corpus
http://mediacorpus.com/smarter-searching-semantic-literal/

[11] R. Guha, Rob McCool, Eric Miller "Semantic Search"

[12] Semantic Search: Factors considered by Search Engines
http://www.techulator.com/resources/5933-What-Semantic-Search.aspx

[13] Concept of Semantic Web
http://www.indiastudychannel.com/resources/150620-Semantic-search-An-overview.aspx

[14] Μηχανή Αναζήτησης Google
http:// www.google.com

[15] Σημασιολογική Μηχανή Αναζήτησης
http:// www.duckduckgo.com

[16] The Semantic Web
http://www.w3schools.com/web/web_semantic.asp

[16] Web Crawler
http://codeglobe.blogspot.gr/2009/02/web-crawler-or-webrobot-or-web-spider.html

[17] Σύγχρονη Τεχνική Επιθεώρηση «Τα οφέλη του Σημασιολογικού Ιστού στο e-Επιχειρείν
http://www.technicalreview.gr/index.php?option=com_content&task=view&id=684

[18] Web 1.0, 2.0, 3.0
http://linnnordahl.wordpress.com/2011/09/29/web-1-0-2-0-and-3-0/

[19] Linked Data
http://www.w3.org/DesignIssues/LinkedData.html

[20] Ten years of building Semantic Web
http://www.fgcsic.es/lychnos/en_EN/articles/ten_years_of_building_a_semantic_web

[21] Roberto Garcia "A Semantic Web Approach to Digital Rights Management"

[22] Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Μονάδα Σημασιολογικού Ιστού
http://www.swu.auth.gr/el/glossary

[23] Σημασιολογικό Διαδίκτυο
http://courses.dbnet.ntua.gr/fsr/2508/Intro_to_Semantic_Web(lecture4).pdf

[24] Wikipedia, The free Encyclopedia
http://en.wikipedia.org/wiki/Unicode

[25] XML Images
http://images.yourdictionary.com/xml

[26] RDF Model
http://www.xml.com/pub/a/2003/02/05/brownsauce.html

[27] Cambridge Semantics
http://www.cambridgesemantics.com/semantic-university/sparql-by-example

[28] Ontologies and the Semantic Web
http://www.obitko.com/tutorials/ontologies-semantic-web/owl-dl-semantics.html

[29] Intelligent agents and the Semantic Web
http://www.ibm.com/developerworks/library/wa-intelligentage/

[30] Wikipedia, The free Encyclopedia
http://en.wikipedia.org/wiki/Intelligent_agent

[31] Agents and the Semantic Web
http://www.it.hiof.no/prosjekter/hoit/html/nr1_05/kvh.html

[32] Competing for the Future with the Intelligent Agents
http://home1.gte.net/pfingar/agents_doc_rev4.htm

[33] Α. Μπάτζιος «Εξόρυξη και Διαχείριση Σημασιολογικής Πληροφορίας στον Σημασιολογικό Ιστό»

[34] John Domingue, Dieter Fensel, James Hendler "Handbook of Semantic Web Technologies", 2011

[35] Eetu Makela "Survey of Semantic Search Research", 2005.

[36] Tranh Tran, Peter Mika, "A survey of Semantic Search Approaches"

[37] M. Hildebrand "End-user for access to heterogeneous linked data"

[38] Γ. Στοΐλος, Μετσόβειο Πολυτεχνείο, " Γλώσσες Αναπαράστασης Γνώσης στο Σημασιολογικό Ιστό"

[39] RDFa
http://www.w3.org/TR/xhtml-rdfa-primer/

[40] Semantic Web Primer
http://www.ics.forth.gr/isl/swprimer/presentation.htm

[41] Andreas Harth, Aidan Hogan, Jürgen Umbrich, Stefan Decker "Building a Semantic Web Search Engine: Challenges and Solutions"

[42] R.Guha, Eric Miller and R. McCool : Semantic Search

[43] Π. Κατσιούλη «Απεικόνιση Σχεσιακού Μοντέλου σε Οντολογία Σημασιολογικού Ιστού»

[44] Thanh Tran, Gunter Ladwig "Structure Index for RDF Data"

[45] Joel Coffman, Alfred C. Weaver "Learning to Rank Results in Relational Keyword Search"

[46] Krzystof Janowicz, Pennsylvania State University, Pascal Hitzler, Wright State University "Semantic Search on the Web" (keyword natural approaches)

[47] Ghislain Auguste Atemezing "Analyzing and Ranking Multimedia Ontologies for their Reuse"

[48] G. Li, B.C. Ooi, J. Feng, J. Wang, L. Zhou "Ease: an Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured Data" ranking

[49] International Hellenic University "Knowledge Search Engines"

[50] Mathieu d' Aquin, Enrico Motta, Marta Sabou, Sofia Angeletou, Laurian Gridinoc, Vanessa Lopez, Davide Guidi, Open University "Toward a New Generation of Semantic Web Applications"

[51] Tummarello, Delbru, Oren "Sindice.com: Weaving the Open Linked Data"

[52] A.Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, S. Decker "Searching and Browsing Linked Data with SWSE: the semantic Web Search Engine"

[53] Mathieu d'Aquin, Li Ding, Enrico Motta "Semantic Web Search Engines"

[54] M.Fernandez, I. Cantador, V. Lopez, D. Vallet, P.Castells, E. Motta "Semantically enhanced Information Retrieval: an ontology-based approach"

[55] S.Grimm, A.Abecker, J.Volker, R.Studer "Ontologies and the Semantic Web"

[56] F. Giunchiglia, U. Kharkevich, I. Zaihrayeu, "Concept search", 2009.

[57] E. Kandogan, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, H. Zhu, "Avatar semantic search: a database approach to information retrieval,", 2006.

[58] D. Damljanovic, M. Agatonovic, H. Cunningham "Natural language interfaces to ontologies: Combining syntactic analysis and ontology based lookup through the user interaction", 2010.

[59] I. Μανωλόπουλος, Α. Παπαδόπουλος "Ανάκτηση Πληροφορίας"

[60] SLOR "Semantic Learning Objects Repository"
http://slor.sourceforge.net/e_store.htm

[61] R. De Virgilio, F. Giunchiglia, L. Tanca "Semantic Web Information Management", 2010

[62] B. Haslhofer, E. Momeni, B. Schandl, S. Zander " Europeana RDF Store Report", 2011

[63] U.Straccia "Fuzzy Logic, Annotation Domains and Semantic Web Languages"

[64] L.Ding, T.Finin, A.Joshi, Y.Peng, R.Pan, P.Reddivari "Search on the Semantic Web", University of Maryland, Baltimore, 2005.

[65] G. A. Atemezing "Analyzing and Ranking Multimedia Ontologies for their Reuse".