Στις μέρες μας η εκρηκτική ανάπτυξη του διαδικτύου και ειδικότερα του παγκόσμιου ιστού (WWW) έχει ως αποτέλεσμα ένα μεγάλο πλήθος ετερογενών πηγών πληροφοριών να γίνονται συνεχώς διαθέσιμοι. Ο πιο συνηθισμένος μηχανισμός αναζήτησης στον παγκόσμιο ιστό αλλά και γενικότερα σε ψηφιακές βιβλιοθήκες (π.χ. ψηφιακές εγκυκλοπαίδειες) είναι με τη χρήση αναλυτικών μεθόδων (Information Retrieval (IR) based on analytical query based search strategies). Σύμφωνα με αυτή τη μέθοδο αναζήτησης δημιουργείται ένα κεντρικό ευρετήριο (index) με διάφορες τεχνικές συλλογής (αυτή η διαδικασία ονομάζεται crawling στο διαδίκτυο) και στατιστικής ανάλυσης των κειμένων. Το κεντρικό ευρετήριο χρησιμοποιείται ώστε να υποβάλλονται ερωτήματα τα οποία επεξεργάζονται και πάλι με μία ποικιλία αλγορίθμων και τεχνικών και έτσι τελικά παρουσιάζονται (συνήθως ταξινομημένα ως προς τη σχετικότητα τους) τα αποτελέσματα της αναζήτησης στους τελικούς χρήστες. Οι τελικοί χρήστες μελετούν τα αποτελέσματα και προσπαθούν να εντοπίσουν κείμενα σχετικά με την πληροφοριακή τους ανάγκη.

Ειδικότερα στο περιβάλλον του παγκόσμου ιστού η κεντρικοποιημένη (centralised) αυτή προσέγγιση παρουσιάζει προβλήματα και δυσκολίες κυρίως λόγω του τεράστιου διαθέσιμου όγκου πληροφοριών. Για παράδειγμα, ακόμη και οι μεγαλύτερες μηχανές αναζήτησης όπως η Google και η Yahoo υπολογίζεται ότι ευρετηροποιούν μόνο ένα υποσύνολο του παγκόσμιου ιστού (κάτω από 50%). Επίσης ένα ακόμη βασικό πρόβλημα είναι ότι μεγάλο μέρος από τις πηγές πληροφοριών (web sites) δεν επιτρέπουν την προσπέλαση τους από web crawlers και απαγορεύουν έτσι την κεντρικοποιημένη ευρετηριοποίηση των κειμένων τους από τις μηχανές αναζήτησης. Το σύνολο αυτό του παγκόσμιου ιστού που δεν είναι δυνατόν να ευρητηριαστεί είναι γνωστό και ως hidden web και αποτελεί ένα τεράστιο μέρος του παγκόσμιου ιστού. Ένα άλλο πρόβλημα αποτελεί το γεγονός ότι η διαθέσιμη πληροφορία ανανεώνεται με τέτοιους ρυθμούς που είναι δύσκολο να γίνεαι συνεχής και ενημερωμένη δημιουργία ευρητηρίου. Επίσης σε πολλές περιπτώσεις ένας μεγάλος όγκος πληροφοριών βρίσκεται μέσα σε εταιρικά δίκτυα που η αναζήτηση σε αυτά (γνωστή και ως enterprise search) δεν είναι εφική μέσω των μηχανών αναζήτησης γενικού σκοπού όπως η Google για παράδειγμα.

Μία εναλλακτική πρόταση στο προηγούμενο πρότυπο αναζήτησης είναι η κατανεμημένη αναλυτική αναζήτηση (Distributed Information Retrieval) όπου υπάρχουν πολλά κατανεμημένα (και ετερογενή) ευρετήρια. Ένα πρόβλημα το οποίο αντιμετωπίζουν οι χρήστες που αναζητούν πληροφορίες σε ένα τέτοιο περιβάλλον αναζήτησης, είναι το πρόβλημα της επιλογής πηγών πληροφοριών και σύνθεσης αποτελεσμάτων από τις διάφορες πηγές πληροφοριών το οποίο είναι γνωστό με διάφορους όρους όπως collection fusion problem, metasearch, data fusion ή federated search.

Οι όροι αυτοί χρησιμοποιούνται διεθνώς για να δηλώσουν το πρόβλημα της επιλογής πηγών πληροφοριών από τις πολλές που είναι διαθέσιμες σε ένα κατανεμημένο σύστημα, η/και, τη δημιουργία ενός αποτελέσματος το οποίο να παρουσιάζει συντιθέμενα τα επιμέρους αποτελέσματα των πολλών παράλληλων αναζητήσεων που γίνονται στις επιλεγμένες πηγές πληροφοριών.

Στο παρόν ερευνητικό πρόγραμμα μελετήθηκαν αυτά τα θέματα και δημιουργήθηκαν καινοτόμες μέθοδοι και αλγόριθμοι για την επίλυση αυτών των προβλημάτων. Ειδικότερα, προτάθηκαν, αναπτύχθηκαν και αξιολογήθηκαν αλγόριθμοι βασισμένοι σε link analysis, multiple regression, hybrid methods και integral based. Τα αποτελέσματα των αξιολογήσεων έχουν παρουσιαστεί σε διάφορα συνέδρια και περιοδικά και αναδεικνύουν την αποτελεσματικότητα των προτεινόμενων μεθόδων.

Distributed Information Retrieval (DIR), also known as federated search, offers users the capability of simultaneously searching multiple remote information sources1 (i.e. search engines or specialized web sites) through a single interface. The importance of DIR has particularly augmented in recent years as the prohibitive size and rate of growth of the web make it impossible to be indexed completely. More importantly, a large number of web sites, collectively known as invisible web are either not reachable by search engines or do not allow their content to be indexed by them, offering their own search capabilities. Even publicly available, up-to-date and authoritative government information is often not indexable by search engines. Studies have indicated that the size of the invisible web may be 2?50 times the size of the web reachable by search engines.

The DIR process can be perceived as three separate but interleaved subprocesses: Source representation, in which surrogates of the available remote collections are created. Source selection, in which a subset of the available information collections is chosen to process the query and results merging, in which the separate results are combined into a single merged result list which is returned to the user.

In this research program several methods and algorithms have been suggested to provide solution to the problems of results merging and source selection. More specifically methods for source selection were suggested based on link analysis and on modelling information sources as integrals. Finally algorithms for results merging have been suggested on regression and hybrid methods. Results have been published to various IR conferences (ECIR 2007, CIKM 2007, ECIR 2009) and journals. Systematic evaluation indicates the effectiveness of the suggested methods.