ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΣΥΓΧΩΝΕΥΣΗ ΜΟΝΟΓΛΩΣΣΩΝ ΘΗΣΑΥΡΩΝ ΘΕΜΑΤΙΚΩΝ ΟΡΩΝ

Μάριος Μ. Συντιχάκης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΗΡΑΚΛΕΙΟ, Φεβρουάριος 1997

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΣΥΓΧΩΝΕΥΣΗ ΜΟΝΟΓΛΩΣΣΩΝ ΘΗΣΑΥΡΩΝ ΘΕΜΑΤΙΚΩΝ ΟΡΩΝ

Εργασία που υποβλήθηκε από τον
ΜΑΡΙΟ Μ. ΣΥΝΤΙΧΑΚΗ
ως μερική εκπλήρωση των απαιτήσεων για την απόκτηση
ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

Συγγραφέας:

Μάριος Μ. Συντιχάκης

Εισηγητική Επιτροπή:

Πάνος Κωνσταντόπουλος
Αναπληρωτής Καθηγητής, Επόπτης

Γιώργος Γεωργακόπουλος
Επίκουρος Καθηγητής, Μέλος

Γεώργιος Δ. Σταμούλης
Επίκουρος Καθηγητής, Μέλος

Δεκτή:

Πάνος Κωνσταντόπουλος
Πρόεδρος Επιτροπής Μεταπτυχιακών Σπουδών

Ηράκλειο, Φεβρουάριος 1997

Αφιερώνεται στους γονείς μου,
Μιχάλη και Καλλιόπη
ως ελάχιστη ανταπόδοση
όσων μου έχουν προσφέρει

ΠΡΟΛΟΓΟΣ

Το κείμενο που ακολουθεί, περιγράφει το αντικείμενο και τ' αποτελέσματα της δουλειάς μου στο πλαίσιο της εργασίας με θέμα την συγχώνευση μονόγλωσσων θησαυρών θεματικών όρων, για την απόκτηση του Διπλώματος Μεταπτυχιακής Εξειδίκευσης απ' το Τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης.

Στο Κεφάλαιο 1.1 παρουσιάζονται ακροθιγώς οι θησαυροί θεματικών όρων και ορίζεται το πρόβλημα της συγχώνευσής τους και η έκταση στην οποία αυτό μας απασχόλησε.

Στο Κεφάλαιο 2 εντάσσουμε την εργασία στο ευρύτερο ερευνητικό πλαίσιο: αρχικά γίνεται λεπτομερής αναφορά στην οργάνωση θησαυρών, αναλύεται η διαδικασία της συγχώνευσης σε φάσεις και τέλος, ακολουθεί μια επισκόπηση της βιβλιογραφίας σε θέματα σχετικά με κάθε φάση της συγχώνευσης.

Στο Κεφάλαιο 3, εισάγουμε ένα γενικό συνολοθεωρητικό μοντέλο παράστασης θησαυρών και διατυπώνουμε τους περιορισμούς ακεραιότητας που πρέπει να ικανοποιούνται. Εισάγεται επίσης ένα σύνολο στοιχειωδών πράξεων ενημέρωσης θησαυρών το οποίο θα χρησιμοποιηθεί στην συνέχεια για να εκφράσει σύνθετες πράξεις ενημέρωσης όπως η συγχώνευση.

Στο Κεφάλαιο 4, παρουσιάζουμε την μέθοδο συγχώνευσης που σχεδιάσαμε και υλοποιήσαμε: αρχικά γίνεται μια επισκόπηση της μεθόδου και στην συνέχεια, εξετάζεται κάθε φάση συγχώνευσης ξεχωριστά.

Το Παράρτημα I περιέχει μια σύντομη εισαγωγή στην γλώσσα παράστασης γνώσης Telos και το Σύστημα Σημασιολογικού Ευρετηριασμού. Στο Παράρτημα II περιγράφονται οι σημαντικότερες δομές δεδομένων που χρησιμοποιούνται στο Κεφάλαιο 4, οι πράξεις επί αυτών και η πολυπλοκότητά τους. Στο Παράρτημα III αποδεικνύεται οτι η συνάρτηση εννοιολογικής απόστασης όρων που παρουσιάζεται στο Κεφάλαιο 4, είναι ψευδομετρική. Το κείμενο κλείνει με το ελληνικό και αγγλικό γλωσσάρι τα οποία δίνονται στα Παραρτήματα IV και V αντίστοιχα και την σχετική βιβλιογραφία.

ΤΥΠΟΓΡΑΦΙΚΕΣ ΣΥΜΒΑΣΕΙΣ

Τα ονόματα προϊόντων, μεθοδολογιών και συστημάτων γράφονται με χαρακτήρες χωρίς ουρίτσες (sans serif). Με χαρακτήρες γραφομηχανής (`typewriter`) γράφεται οτιδήποτε έχει σχέση με κώδικα όπως ονόματα κλάσεων ή οι συσχετίσεις μεταξύ όρων σ' ένα θησαυρό. Οι δεσμευμένες λέξεις της ψευδογλώσσας που χρησιμοποιούμε στους αλγόριθμους γράφονται με έντονους (**boldface**) χαρακτήρες, ενώ οι μεταβλητές γράφονται όπως στα μαθηματικά π.χ., L. Σημαντικές έννοιες γράφονται με πλάγιους (*slanted*) χαρακτήρες την πρώτη φορά που συναντώνται στο κείμενο, συνήθως μαζί με τον αντίστοιχο αγγλικό όρο σε παρένθεση. Οι συντομογραφίες κατά κανόνα αποφεύγονται, αλλά όταν χρησιμοποιούνται δίνονται στην αγγλική π.χ., DBMS αντί ΣΔΒΔ. Αν και προσπαθούμε να χρησιμοποιούμε πρότυπη ελληνική ορολογία, σε περιπτώσεις που τέτοια δεν υπάρχει —όπως π.χ., ονόματα και συντομογραφίες συσχετίσεων σε θησαυρούς— χρησιμοποιούμε όρους που διαισθητικά είναι "περισσότερο δόκιμοι". Στα παραδείγματα θησαυρών που υπάρχουν στο κείμενο, οι δόκιμοι όροι γράφονται με κεφαλαία ενώ οι αδόκιμοι με πεζά ακολουθώντας τις συμβάσεις του [ISO86], ενώ όταν δεν γίνεται διάκριση οι όροι γράφονται με κεφαλαίο το πρώτο γράμμα κάθε λέξης τους. Τα περισσότερα παραδείγματα που χρησιμοποιούμε προέρχονται απ' το [ISO86], τον θησαυρό AAT και το [Sve89] και δίνονται αυτούσια. Τα υπόλοιπα παραδείγματα δίνονται στα ελληνικά. Τέλος, άλλες —μικρότερης σπουδαιότητας— συμβάσεις δίνονται περιστασιακά στο κείμενο.

ΕΥΧΑΡΙΣΤΙΕΣ

Αισθάνομαι την ανάγκη να ευχαριστήσω θερμότατα τον Αναπληρωτή Καθηγητή του Πανεπιστημίου Κρήτης, ακαδημαϊκό μου σύμβουλο, επόπτη και δάσκαλο κ. Πάνο Κωνσταντόπουλο, για την ευκαιρία που μου έδωσε για την εκπόνηση αυτής της μεταπτυχιακής εργασίας, και ιδιαίτερα για την αμέριστη βοήθεια, καθοδήγηση και κατανόησή του κατά την διάρκειά της.

Ευχαριστώ τα μέλη της εξεταστικής επιτροπής, κκ. Γεώργιο Γεωργακόπουλο και Γεώργιο Σταμούλη Επίκουρους Καθηγητές στο Πανεπιστήμιο Κρήτης, καθώς και τον δρα. Martin Doerr για τις υποδείξεις και παρατηρήσεις τους.

Θερμές επίσης ευχαριστίες, οφείλω στους κκ. Θεόδωρο Καλαμπούκη και Γιάννη Δημητρίου, Αναπληρωτή και Επισκέπτη Καθηγητή αντίστοιχα, στο Οικονομικό Πανεπιστήμιο Αθηνών για την ανεκτίμητη βοήθεια και εμπιστοσύνη τους.

Ευχαριστώ όλους τους συναδέλφους για την βοήθειά τους και ειδικά στους Θανάση Χρυσό, Γιάννη Σουρλατζή, Δέσποινα Βαμβακά, Σαράντη Τούλη, Χρήστο Γεωργή, Σταυρούλα Κιζλαρίδου και Γιάννη Τζίτζικα.

Ευχαριστώ θερμά το Ινστιτούτο Πληροφορικής του Ιδρύματος Τεχνολογίας και Ερευνας και το Τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης για την τεχνική και οικονομική υποστήριξη που μου παρείχαν κατά την διάρκεια της εκπόνησης αυτής της εργασίας.

Στην ολοκλήρωση της εργασίας αυτής συνέβαλαν και άνθρωποι εκτός της πανεπιστημιακής κοινότητας. Θέλω να ευχαριστήσω θερμότατα τη μνηστή μου Καλλιόπη Σκουρέλλου και τους γονείς μου για την υποστήριξη και την συμπαράστασή τους σε κάθε δύσκολη στιγμή αυτά τα δύο χρόνια.

Μ.Μ.Σ.
Ηράκλειο,
Ιανουάριος 1996

ΣΥΓΧΩΝΕΥΣΗ ΜΟΝΟΓΛΩΣΣΩΝ ΘΗΣΑΥΡΩΝ ΘΕΜΑΤΙΚΩΝ ΟΡΩΝ

Μάριος Μ. Συντιχάκης

Μεταπτυχιακή Εργασία

Τμήμα Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

ΠΕΡΙΛΗΨΗ

Ενας θησαυρός είναι μια εννοιολογική δομή που περιγράφει έννοιες ενός πεδίου γνώσης με την χρήση ενός λεξιλογίου όρων και τριών τύπων συσχετίσεων μεταξύ αυτών: συσχετίσεις ισοδυναμίας, ιεραρχικές συσχετίσεις και συσχετίσεις συνάφειας. Στην πράξη έχει αποδειχθεί οτι οι θησαυροί είναι πολύ σημαντικό τμήμα συστημάτων ανάκλησης πληροφορίας. Η κατασκευή τους ωστόσο είναι εξαιρετικά επίπονη εργασία —όπως συμβαίνει με κάθε βάση γνώσης άλλωστε. Στα πλαίσια αυτής της εργασίας, σε μια προσπάθεια να συνεισφέρουμε στο πρόβλημα αυτό, ασχολούμαστε με την συγχώνευση θησαυρών θεματικών όρων.

Η συγχώνευση θησαυρών στοχέυει στην ενοποίηση όρων και των συσχετίσεων δύο ή περισσότερων θησαυρών για την παραγωγή ενός θησαυρού που θα περιγράφει ενα σύνολο εννοιών το οποίο είναι υπερσύνολο των συνόλων εννοιών που περιγράφουν οι συγχωνευόμενοι θησαυροί.

Προσπαθούμε ν' αντιμετωπίσουμε συστηματικά το πρόβλημα εισάγωντας ένα συνολοθεωρητικό πλαίσιο παράστασης θησαυρών ανεξάρτητο από θέματα υλοποίησης, ένα σύνολο περιορισμών ακεραιότητας και ένα σύνολο πράξεων ενημέρωσης θησαυρών. Επιπλέον διακρίνουμε την διαδικασία συγχώνευσης στις φάσεις προενοποίησης, ανάλυσης, αναδιάρθρωσης και ενοποίησης, δίνοντας ιδιαίτερη έμφαση στην φάση της ανάλυσης που στόχο έχει τον εντοπισμό όρων που περιγράφουν κοινές έννοιες μεταξύ των συγχωνευόμενων θησαυρών.

Προκειμένου να εντοπίσουμε όρους οι οποίοι περιγράφουν την ίδια έννοια σε διαφορετικούς θησαυρούς, εισάγουμε ένα μοντέλο υπολογισμού αποστάσεων μεταξύ όρων το οποίο αποτελεί προσαρμογή ενός γενικότερου μοντέλου ομοιότητας αντικειμένων.

Το μοντέλο βασίζεται στις συσχετίσεις μεταξύ όρων. Για να μειώσουμε επιπλέον τα σύνολα όρων που θα πρέπει να συγκριθούν, συνδυάζουμε την εννοιολογική απόσταση με την λεκτική ομοιότητα αξιοποιώντας ταυτόχρονα τις συσχετίσεις ισοδυναμίας των συγχωνευόμενων θησαυρών.

Η διαδικασία της συγχώνευσης, οδηγείται από μια τοπολογική πολιτική διάσχισης του κατευθυνόμενου ακυκλικού γράφου που σχηματίζουν οι συσχετίσεις ιεραρχίας των συγχωνευόμενων θησαυρών, η οποία έχει την ιδιότητα να βελτιώνει προοδευτικά την ποιότητα των υπολογιζόμενων αποστάσεων μεταξύ όρων. Ο τρόπος λειτουργίας μπορεί να είναι διαλογικός ή δεσμικός.

Για την παράσταση και αποθήκευση θησαυρών χρησιμοποιούμε το μοντέλο δεδομένων της γλώσσας Telos του συστήματος Σημασιολογικού Ευρετηριασμού (Semantic Index System, SIS) και ένα απλό εννοιολογικό σχήμα. Χρησιμοποιώντας το SIS ως πλατφόρμα ανάπτυξης, υλοποιήσαμε ενα πρωτότυπο της μεθόδου συγχώνευσης σε C++. Επίσης πραγματοποιήσαμε ένα πείραμα για την συγχώνευση δύο γνωστών θησαυρών: του "Computing Reviews Classification System" και του "Library of Congress Subject Headings". Τα πρώτα αποτελέσματα ήταν αρκετά ενθαρρυντικά και πιστεύουμε πως απλές αλλαγές μπορούν να βελτιώσουν την μέθοδό μας ακόμη περισσότερο.

Επόπτης: Πάνος Κωνσταντόπουλος
Αναπληρωτής Καθηγητής Επιστήμης Υπολογιστών
Πανεπιστήμιο Κρήτης

MONOLINGUAL THESAURI MERGING

Marios M. Sintichakis

Master of Science Thesis

Department of Computer Science
University of Crete

ABSTRACT

A thesaurus is a conceptual structure representing concepts from a particular domain of discourse using a controlled vocabulary and three types of relationships between concepts: equivalence, hierarchical and associative relationships. It has been proven in practice that thesauri can play an essential role as parts of information retrieval systems. Yet, the construction of thesauri is an exceptionally hard and time consuming task. Within the context of this thesis, we deal with the problem of monolingual thesauri merging as a means for thesauri construction and development.

The aim of thesauri merging is the integration of both vocabularies and relationships of two or more thesauri, so as to produce a new thesaurus which describes more concepts than each one of the merging thesauri do.

We try to systematically address the problem: we introduce a set-theoretic framework for the representation of thesauri and the relevant integrity constraints which are independent of any implementation issues. We decompose the merging process in four phases namely pre-integration, analysis, conflict detection and resolution and integration. Our attention is mainly focused to the phase of analysis which is aimed at the detection of terms of the merging thesauri which ascribe the same concept.

In order to detect terms ascribing the same concept in different thesauri, we introduce a model for the computation of conceptual distance between terms. This model is an adaptation of a more general model of analogical similarity and it is based on the relationships between terms. Moreover we combine conceptual term distances with lexical similarity and equivalence relationships, so as to reduce the size of the term sets which should be considered.

The merging is a top-down procedure based on the topological sorting imposed by

the directed acyclic graph which is formed by the hierarchical relationships of the merging thesauri. This policy has the advantage that it tends to gradually improve the accuracy of the conceptual distances computed at each level. The mode of operation can be either batch or interactive.

We have used the Telos data model of the Semantic Index System for the representation, storage and management of thesauri. Using SIS as a platform, we have implemented a prototype of our method in C++. We have conducted a merging experiment trying to integrate two well-known thesauri: the "Computing Reviews Classification System" and the "Library of Congress Subject Headings". The first results were quite encouraging and we believe that slight modifications can further improve our method.

Supervisor: Panos Constantopoulos
Associate Professor of Computer Science
University of Crete

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Θησαυροί θεματικών όρων

Στο πλαίσιο κάθε πεδίου γνώσης υπάρχει ένα *λεξιλόγιο (vocabulary)*, ένα σύνολο λέξεων ή φράσεων της φυσικής γλώσσας, που χρησιμοποιείται για να περιγράψει *έννοιες (concepts)* —άτομα, υλικά, αφηρημένες οντότητες— που σχετίζονται με το συγκεκριμένο πεδίο γνώσης. Κάθε στοιχείο ενός λεξιλογίου περιγράφει μια ή περισσότερες έννοιες και ονομάζεται *όρος*. Ενα λεξιλόγιο λέγεται οτι είναι *ελεγχόμενο (controlled)* όταν είναι ένα κλειστό, εγκεκριμένο (authorized) προς χρήση, σύνολο όρων ενός πεδίου γνώσης. Στην περίπτωση ενός ελεγχόμενου λεξιλογίου, οι όροι συχνά διακρίνονται σε *δόκιμους (preferred terms)* και *αδόκιμους (non-preferred terms)*. Οι δόκιμοι όροι είναι εγκεκριμένοι προς χρήση, ενώ οι αδόκιμοι είναι εναλλακτικοί όροι που περιγράφουν την ίδια έννοια με κάποιο δόκιμο όρο.

Οι έννοιες κάθε πεδίου γνώσης συνδέονται συνειρμικά μεταξύ τους. Τέτοιες συνδέσεις μεταξύ εννοιών ή όρων ονομάζονται *σημασιολογικές συσχετίσεις*.

Ενας *θησαυρός όρων (thesaurus)* είναι ένα ελεγχόμενο λεξιλόγιο ενός πεδίου γνώσης, επαυξημένο με ρητές σημασιολογικές συσχετίσεις μεταξύ των εννοιών που αυτό περιγράφει [ISO86]. Αν το λεξιλόγιο ενός θησαυρού αποτελείται από όρους που χρησιμοποιούνται αποκλειστικά στο πλαίσιο μιας συγκεκριμένης φυσικής γλώσσας (π.χ., Ελληνική), εξαιρουμένων φυσικά αυτών που είναι δανεισμένοι από άλλες γλώσσες, τότε ονομάζεται *μονόγλωσσος (monolingual thesaurus)* ενώ στην αντίθετη περίπτωση ονομάζεται *πολύγλωσσος (multilingual thesaurus)*.

1.1.1 Ενα παράδειγμα

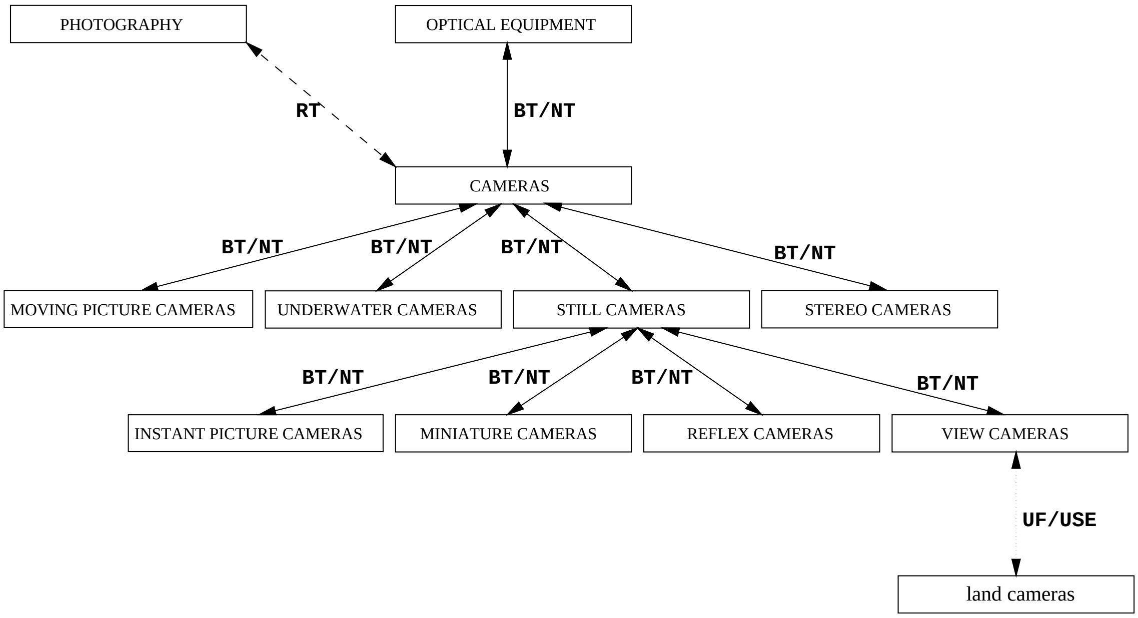
Στο παράδειγμα 1.1 δίνεται ένα απόσπασμα από ένα υποθετικό θησαυρό, που περιγράφει οπτικό εξοπλισμό. Για κάθε έννοια που περιγράφεται από ένα δόκιμο όρο παρατίθενται οι όροι των γενικότερων εννοιών ή απλά γενικότεροι όροι (Broader Terms), οι όροι των ειδικότερων εννοιών ή απλά ειδικότεροι όροι (Narrower Terms), οι όροι συναφών εννοιών ή απλά συναφείς όροι (Related Terms) και οι εναλλακτικοί (αδόκιμοι όροι) (Used For Terms). Κάθε αδόκιμος όρος παραπέμπει (USE) στον ισοδύναμο δόκιμο όρο. Αυτές είναι οι βασικές συσχετίσεις που ορίζονται σε θησαυρούς [ISO86] και οι οποίες είναι ενα κοινά αποδεκτό υποσύνολο των πιθανών συσχετίσεων μεταξύ εννοιών.

ΠΑΡΑΔΕΙΓΜΑ 1.1

```
CAMERAS
    BT    OPTICAL EQUIPMENT
    NT    MOVING PICTURE CAMERAS
          UNDERWATER CAMERAS
          STEREO CAMERAS
          STILL CAMERAS
    RT    PHOTOGRAPHY
INSTANT PICTURE CAMERAS
    BT    STILL CAMERAS
land cameras
    USE   VIEW CAMERAS
MINIATURE CAMERAS
    BT    STILL CAMERAS
MOVING PICTURE CAMERAS
    BT    CAMERAS
OPTICAL EQUIPMENT
    NT    CAMERAS
PHOTOGRAPHY
    RT    CAMERAS
REFLEX CAMERAS
    BT    STILL CAMERAS
SINGLE LENS REFLEX CAMERAS
    BT    REFLEX CAMERAS
STEREO CAMERAS
    BT    CAMERAS
STILL CAMERAS
    BT    CAMERAS
    NT    INSTANT PICTURE CAMERAS
          MINIATURE CAMERAS
          REFLEX CAMERAS
          VIEW CAMERAS
TWIN LENS REFLEX CAMERAS
    BT    REFLEX CAMERAS
```

VIEW CAMERAS
    `UF`   land cameras
    `BT`   STILL CAMERAS
UNDERWATER CAMERAS
    `BT`   CAMERAS

Μια πιο εποπτική παράσταση του του θησαυρού, του παραδείγματος 1.1, δίνεται στο σχήμα 1.1, όπου ο θησαυρός αναπαρίσταται μ' ένα γράφο του οποίου οι κόμβοι είναι όροι και οι ακμές συσχετίσεις μεταξύ των εννοιών που αυτοί παριστάνουν. Τόσο στους όρους όσο και στις συσχετίσεις θ' αναφερθούμε με λεπτομέρεια στο Κεφάλαιο 2, ωστόσο συντομότατα αναφέρουμε εδώ, οτι μια συσχέτιση `UF/USE` εκφράζει ισοδυναμία μεταξύ δύο όρων, μια συσχέτιση `BT/NT` εκφράζει ιεραρχική σχέση μεταξύ δύο εννοιών και μια συσχέτιση `RT` εκφράζει μια οποιαδήποτε άλλη σχέση μεταξύ δύο εννοιών.



Σχήμα 1.1: Γραφική Παράσταση Θησαυρού.
Οι κόμβοι παριστάνουν όρους ενώ οι ακμές συσχετίσεις μεταξύ των αντίστοιχων εννοιών. Οι συσχετίσεις ισοδυναμίας παριστάνονται με γραμμές με κουκίδες, οι συσχετίσεις ιεραρχίας με συμπαγείς γραμμές και οι συσχετίσεις συνάφειας με διακεκομένες γραμμές.

1.1.2 Πεδίο εφαρμογής

Οι θησαυροί χρησιμοποιούνται κυρίως για τον ευρετηριασμό (indexing) και την ανάκληση πληροφορίας (information retrieval) από κείμενα (documents). Σ' αυτά τα πλαίσια χρήσης, τρείς είναι οι βασικές απαιτήσεις από ένα θησαυρό.

1. Η παροχή ενός πρότυπου (standard) λεξιλογίου για ένα συγκεκριμένο πεδίο γνώσης.

2. Η δημιουργία αμφιμονοσήμαντων αντιστοιχίσεων μεταξύ των όρων του λεξιλογίου και των εννοιών του πεδίου γνώσης, κάτι που δεν συμβαίνει σ' ένα λεξιλόγιο φυσικής γλώσσας.

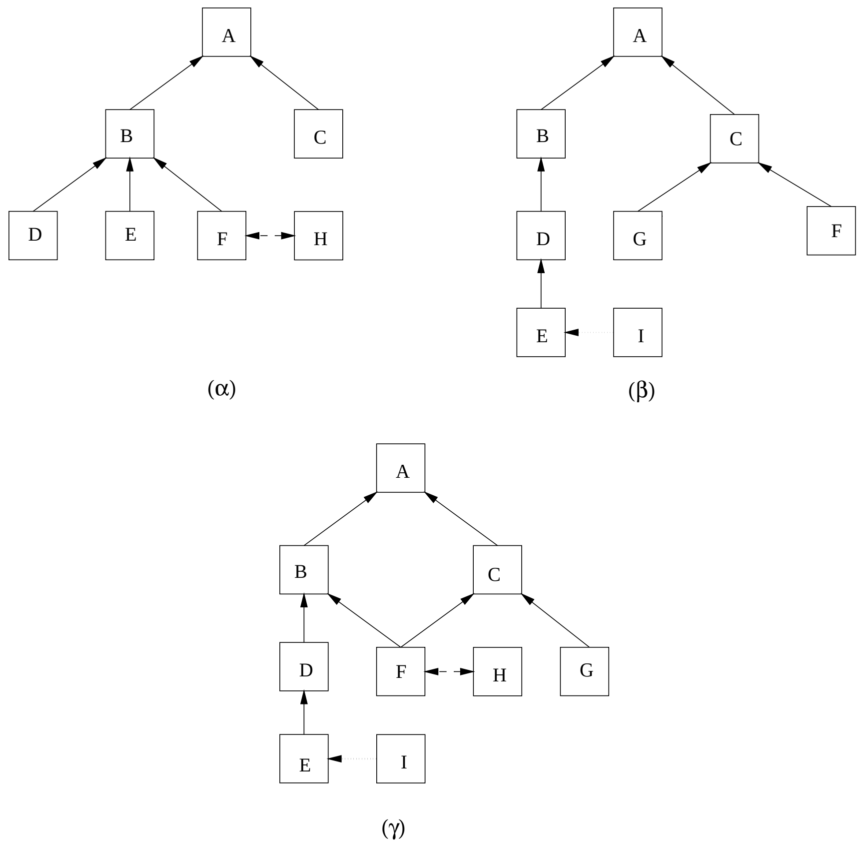
3. Ο ρητός ορισμός των συσχετίσεων μεταξύ των των εννοιών που αντιπροσωπεύουν οι όροι του πεδίου γνώσης προκειμένου να βελτιωθεί η απόδοση του ευρετηριασμού και της ανάκλησης πληροφορίας μέσω του θησαυρού.

Από την πλευρά της ανάκλησης πληροφορίας, η επίτευξη των παραπάνω στόχων είναι εξαιρετικά σημαντική. Το βασικό πρόβλημα των περισσότερων τεχνικών ανάκλησης πληροφορίας είναι η μη ικανοποιητική απόδοση εκφραζόμενη σε όρους *βαθμού ανάκλησης (recall)* και *βαθμού ακρίβειας (precision)* [Pai91], [Sal89], [KS93]. Ο βαθμός ανάκλησης εκράζει το ποσοστό σχετικής (relevant) με μια ερώτηση (query) πληροφορίας το οποίο ανακλήθηκε από την ερώτηση, ενώ ο βαθμός ακρίβειας εκφράζει το ποσοστό της πληροφορίας που ανακλήθηκε από μια ερώτηση, το οποίο είναι σχετικό μ' αυτήν. Η χαμηλή απόδοση (χαμηλός βαθμός ανάκλησης ή/και ακρίβειας) των συστημάτων ανάκλησης πληροφορίας οφείλεται κατά κύριο λόγο στην ύπαρξη μη-αμφιμονοσήματων αντιστοιχίσεων μεταξύ όρων και εννοιών στις φυσικές γλώσσες που εμφανίζεται με την ύπαρξη *συνώνυμων* και *ομόνυμων* εννοιών. Ετσι μια ερώτηση για το "Δία" μπορεί να επιστρέψει πληροφορία σχετικά τόσο με το θεό Δία, όσο και με τον πλανήτη Δία, μειώνοντας έτσι τον βαθμό ακρίβειας της ανάκλησης, ενώ μια ερώτηση για "Αεροσκάφη" ενδέχεται να μην ανακαλέσει πληροφορία σχετική με "Αεροπλάνα", μειώνοντας έτσι το βαθμό ανάκλησης. Αυτό μπορεί ν' αντιμετωπιστεί μέσω ενός θησαυρού που δηλώνει ρητά την ισοδυναμία των όρων "Αεροσκάφη" και "Αεροπλάνα". Οι ιεραρχικές συσχετίσεις επίσης βελτιώνουν την απόδοση των συστημάτων ανάκλησης πληροφορίας. Για παράδειγμα μια ερώτηση για "Θηλαστικά", θ' ανακαλέσει μεταξύ άλλων και πληροφορία για τα "Δελφίνια" κάτι που δεν γίνεται στα συστήματα ανάκλησης πληροφορίας τα οποία δεν υποστηρίζουν θησαυρούς. Πειραματικά αποτελέσματα δείχνουν πως η χρήση θησαυρών σε συστήματα ανάκλησης πληροφορίας, βελτιώνουν σημαντικά την απόδοσή τους [Kri94],[LKL94].

1.2 Ορισμός του προβλήματος

Στην εργασία αυτή ασχολούμαστε με το πρόβλημα της συγχώνευσης θησαυρών, ένα θέμα το οποίο δεν έχει απασχολήσει σε ευρεία κλίμακα τους ερευνητές/επιστήμονες πληροφορικής, γεγονός που ίσως να οφείλεται στο οτι μόνο σχετικά πρόσφατα έχει αναγνωριστεί η αξία των θησαυρών στην ανάκληση πληροφορίας [CLBD93], [Pai91], [LKL94], [Kri94]. Ωστόσο η πρακτική σημασία του προβλήματος είναι μεγάλη και έγκειται στο γεγονός οτι η κατασκευή ενός θησαυρού από το μηδέν είναι μια πολύ δύσκολη εργασία ακόμη και για θησαυρούς μέσου μεγέθους [ISO86], [ISO85]. Από την άλλη πλευρά φαίνεται λογικό να προσπαθήσουμε να ενοποιήσουμε θησαυρούς που έχουν κατασκευαστεί ανεξάρτητα, με σκοπό να κατασκευάσουμε ένα νέο θησαυρό ο οποίος διαισθητικά θ' αποτελεί την "ένωση" των αρχικών θησαυρών.

Το πρότυπο [ISO86] παρέχει ένα σύνολο από τύπους κατασκευών (όρους και τύπους συσχετίσεων) και κανόνες (περιορισμούς) για τον συνδυασμό τους με σκοπό την ανάπτυξη θησαυρών. Διαισθητικά, λειτουργεί ανάλογα μ' ένα μοντέλο δεδομένων, ενώ οι θησαυροί αποτελούν το διαισθητικό ισοδύναμο των εννοιολογικών σχημάτων των βάσων δεδομένων. Κάτω απ' αυτό το πρίσμα, η συγχώνευση ενός συνόλου θησαυρών, είναι μια διαδικασία η οποία στόχο έχει την ενοποίηση των κατασκευών τους, και την παραγωγή μιας δομής η οποία είναι επίσης θησαυρός —ικανοποιεί δηλαδή τους περιορισμούς οργάνωσης που εισάγονται απ' το [ISO86]. Ενα απλοποιημένο παράδειγμα συγχώνευσης δύο θησαυρών δίνεται στο σχήμα 1.2.



Σχήμα 1.2: Παράδειγμα συγχώνευσης θησαυρών.
(α),(β) Οι δύο θησαυροί προς συγχώνευση. (γ) Ο παραγόμενος από την συγχώνευση θησαυρός ο οποίος είναι η ένωση του συνόλου όρων και συσχετίσεων των δύο άλλων θησαυρών, αν και αυτό δεν είναι πάντα εφικτό.

1.3 Εκταση και συνεισφορά

Στα πλαίσια αυτής της εργασίας, εστιάζουμε την προσπάθειά μας στην σχεδίαση και υλοποίηση μιας μεθόδου συγχώνευσης μονόγλωσσων θησαυρών οι όροι των οποίων εκφράζονται με την χρήση μιας κοινής φυσικής γλώσσας. Επιπλέον στην προσπάθεια ν' αντιμετωπίσουμε το πρόβλημα με συστηματικό τρόπο, εισάγουμε ένα απλό συνολοθεωρητικό μοντέλο για την παράσταση θησαυρών και περιορισμών ακεραιότητας κ' ένα σύνολο στοιχειωδών πράξεων ενημέρωσης θησαυρών οι οποίες χρησιμοποιούνται ως βάση για την συγχώνευση.

Υλοποιήσαμε τα παραπάνω χρησιμοποιώντας το Σύστημα Σημασιολογικού Ευρετηριασμού (Semantic Index System, SIS) και το μοντέλο δεδομένων της γλώσσας Telos [MBJK90] το οποίο υποστηρίζεται απ' το SIS [DKT95]. Οι λόγοι για τους οποίους το SIS προτιμήθηκε από άλλα συστήματα είναι οι εξής:

1. Η παράσταση θησαυρών με μορφή δικτύου είναι πιο φυσική.
2. Το SIS ξεπερνά την ταχύτητα απόκρισης γνωστών εμπορικών DBMS σε αναδρομικές ερωτήσεις[1] κατά μία ως δύο τάξεις μεγέθους [CD95].
3. Το SIS ικανοποιεί όλες τις προδιαγραφές [Mil91] ώστε να χρησιμοποιηθεί για την διαχείριση θησαυρών.
4. Σε αρκετές περιπτώσεις οι εμπειρίες από την χρήση άλλων συστημάτων σε σχετικές εργασίες, όπως για παράδειγμα του συστήματος FrameKit+ στην εργασία [MR88] ή του συστήματος Ingres στην εργασία [CLBD93] δεν ήταν ενθαρυντικές.

Η εργασία αυτή συνεισφέρει στο πρόβλημα της συγχώνευσης θησαυρών μια νέα μέθοδο η οποία στηρίζεται στην αυστηρή ερμηνεία και τήρηση της σημασιολογίας των συσχετίσεων μεταξύ όρων, στην εισαγωγή ενός απλού και αποδοτικού μηχανισμού εντοπισμού όμοιων όρων και επιπλέον στην χρήση εννοιολογικών κριτηρίων για την καθοδήγηση της διαδικασίας συγχώνευσης.

[1] Οπως θα φανεί στα επόμενα τέτοιου είδους ερωτήσεις αφενός είναι βασικές και αφετέρου εκτελούνται πολύ συχνά στην συγχώνευση και γενικότερα στην διαχείριση θησαυρών.

ΚΕΦΑΛΑΙΟ 2

ΑΝΑΣΚΟΠΗΣΗ ΤΗΣ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

2.1 Οργάνωση θησαυρών

Οι δομικές κατασκευές ενός θησαυρού είναι οι όροι και οι συσχετίσεις μεταξύ εννοιών που αντιπροσωπεύουν οι όροι. Στην ενότητα αυτή παρουσιάζουμε μερικά απ' τα χαρακτηριστικά των κατασκευών αυτών, δίνοντας έμφαση σ' εκείνα που είναι σημαντικά σε σχέση με την συγχώνευση θησαυρών.

2.1.1 Οροι

Κατηγορίες όρων

Συχνά οι όροι ενός θησαυρού, οργανώνονται σε πολύ γενικές, σημασιολογικά συγγενείς κλάσεις εννοιών τις οποίες θα ονομάζουμε *κατηγορίες όρων (facets)*. Κάθε όρος του θησαυρού, θα πρέπει ν' ανήκει σε μία, ή περισσότερες κατηγορίες, πράγμα που δίνει αμέσως πληροφορία για την υπόσταση της έννοιας που περιγράφει.

Η σύνταξη των όρων

Οροι οι οποίοι αποτελούνται από περισσότερες από μία λέξεις, λέγονται *σύνθετοι όροι (compound terms)* σε αντίθεση με τους *απλούς όρους* οι οποίοι αποτελούνται από μια μονάχα λέξη. Ενα σημαντικό θέμα του εισάγουν οι σύνθετοι όροι είναι η σύνταξή τους η διάταξη δηλαδή των λέξεων που αποτελούν τον όρο [Sve89]. Υπάρχουν δύο προσεγγίσεις στο θέμα της σύνταξης των σύνθετων όρων.

1. *Σύνταξη φυσικής γλώσσας*, για παράδειγμα "Digital Computers".

2. *Αντίστροφη σύνταξη*, για παράδειγμα "Computers, Digital".

Αν και η πρώτη προσέγγιση συνιστάται [ISO86] είναι πιθανό να χρησιμοποιείται και η δεύτερη με το επιχείρημα οτι σε μια αλφαβητικά ταξινομημένη παράθεση των όρων του θησαυρού σχετικοί όροι βρίσκονται ομαδοποιημένοι. Για παράδειγμα

...

"Computers, Digital"

"Computers, Mobile"

"Computers, Personal"

...

Εκτός από την σειρά παράθεσης των λέξεων, η σύνταξη των σύνθετων όρων επηρεάζεται από την συντακτική κατηγορία τους. Ενας όρος μπορεί να είναι ουσιαστικό, ρήμα, επίρημα, επίθετο, επιθετική φράση ή προθετική φράση [ISO86]. Οι σύνθετοι όροι προφανώς μπορεί ν' ανήκουν μόνο στις δύο τελευταίες συντακτικές κατηγορίες. Ωστόσο με ποια από τις δύο αυτές συντακτικές κατηγορίες θ' αποδοθεί ένας όρος επηρεάζει την σύνταξή του. Για παράδειγμα η διαδικασία της διαχείρισης των πληροφοριακών συστήματων μπορεί ν' αποδοθεί στην Αγγλική είτε από τον όρο "Information Systems Management" είτε από τον όρο "Management of Information Systems".

Πολυσημικοί όροι

Είναι συχνό φαινόμενο στις φυσικές γλώσσες να υπάρχουν λέξεις ή φράσεις με όμοια ορθογραφία αλλά διαφορετικές σημασίες. Για παράδειγμα στην Αγγλική, η λέξη "Cranes" μπορεί να σημαίνει το πουλί ή τα ανυψωτικά μηχανήματα. Ομοια η λέξη "Mercury" μπορεί να σημαίνει το στοιχείο ή τον πλανήτη. Τέτοιοι όροι ονομάζονται πολυσημικοί, ενώ οι διαφορετικές έννοιες που μπορεί να σημαίνει ένας πολυσημικός όρος ονομάζονται ομώνυμες. Υπάρχουν τρεις μηχανισμοί για την αποσαφήνιση των ομώνυμων ενός πολυσημικού όρου.

1. Ρητή αποσαφήνιση με την παράθεση γενικότερου όρου Για παράδειγμα, "Mercury (Planet)".

2. Υποννοούμενη αποσαφήνιση από κάποια γενικότερη έννοια ή την κατηγορία στην οποία ανήκει ο όρος. Για παράδειγμα αν ο θησαυρός παρέχει την πληροφορία οτι ο "Mercury" είναι πλανήτης, τότε δεν υπάρχει ανάγκη ρητής αποσαφήνισης του ομώνυμου.

3. Αποσαφήνιση με την χρήση *σημείωσης εμβέλειας (Scope Note–SN)*. Για παράδειγμα, "Mercury `SN` The fifth planet of the solar system".

2.1.2 Σημασιολογικές συσχετίσεις

Οπως αναφέρθηκε, ένας θησαυρός πρέπει να ορίζει ρητά τις σημασιολογικές συσχετίσεις[1] μεταξύ των όρων του. Σύμφωνα τόσο με το πρότυπο για την οργάνωση θησαυρών του ISO [ISO86], όσο και με άλλες εργασίες [CLBD93], [MR88], [RM89], [RM87], [Mil91], [Sve89], [Pai91] υπάρχουν τρεις τύποι συσχετίσεων: η *συσχέτιση ισοδυναμίας (equivalence relationship)*, η *ιεραρχική συσχέτιση (hierarchical relationship)* και η *συσχέτιση συνάφειας (associative relationship).*

Συσχέτιση ισοδυναμίας

Θεωρητικά δύο όροι συνδέονται με μια συσχέτιση ισοδυναμίας όταν αντιπροσωπεύουν την ίδια έννοια. Τα συνώνυμα είναι μια κλασική περίπτωση, ωστόσο όπως αναφέρει η Svenonius στο [Sve89], λίγα πραγματικά συνώνυμα υπάρχουν στις φυσικές γλώσσες. Κατά συνέπεια μια συσχέτιση ισοδυναμίας μπορεί μεταξύ άλλων, να συνδέει επίσης [ISO86]:

1. Τον όρο που περιγράφει μια κλάση αντικειμένων και τα μέλη αυτής (παράδειγμα 2.1).

2. Δύο όρους που εκφράζουν έννοιες οι οποίες είναι διαφορετικές τιμές σε μια κλίμακα. Για παράδειγμα "Wetness" και "Dryness".

3. Διαφορετικές μορφές γραφής ενός όρου. Για παράδειγμα "Roumania", "Rumania", "Romania".

4. Επίσημα και κοινά ονόματα. Για παράδειγμα "Polyethylene" και "Polythene"

Από ένα σύνολο ισοδύναμων όρων, ένας επιλέγεται ως αντιπρόσωπος του συνόλου και χαρακτηρίζεται ως δόκιμος όρος ενώ οι υπόλοιποι ως αδόκιμοι. Η συσχέτιση ισοδυναμίας συνδέει πάντα ένα αδόκιμο και ένα ή περισσότερους δόκιμους όρους, ανάλογα με τον τύπο της. Υπάρχουν τρεις τύποι συσχετίσεων ισοδυναμίας.

1. Συσχετίσεις ισοδυναμίας που συνδέουν ένα αδόκιμο και τον ισοδύναμο δόκιμο όρο. Για παράδειγμα:
   boats `USE` SHIPS.

2. Συσχετίσεις ισοδυναμίας που συνδέουν ένα αδόκιμο και ένα σύνολο ισοδύναμων δόκιμων όρων. Για παράδειγμα:
   lifts `USE` ELEVATORS `or` LIFTING EQUIPMENT

[1] Στην εργασία αυτή με τον όρο συσχέτιση (relationship) όρων θα εννοούμε ένα διατεταγμένο ζεύγος όρων, ενώ με τον όρο σχέση (relation) θα εννοούμε το σύνολο των συσχετίσεων ίδιου τύπου.

3. Συσχετίσεις ισοδυναμίας που συνδέουν ένα σύνθετο αδόκιμο όρο με τα συνθετικά του που πρέπει να είναι δόκιμοι όροι. Για παράδειγμα:
   aircraft engines `USE` AIRCRACFS `and` ENGINES
   Αυτή η μορφή συσχέτισης ισοδυναμίας, ονομάζεται συντακτική παραγοντοποίηση (syntactic factoring).

Οι συσχετίσεις ισοδυναμίας είναι ο μοναδικός τύπος συσχέτισης στον οποίο μπορεί να λάβει μέρος ένας αδόκιμος όρος. Από την πλευρά του αδόκιμου όρου η συσχέτιση συμβολίζεται με το σύμβολο `USE`, ενώ από την πλευρά του δόκιμου όρου με το σύμβολο `UF` (Used For).

<u>ΠΑΡΑΔΕΙΓΜΑ 2.1</u>

ROCKS
    `UF`   basalt
            granite
            slate
basalt
    `USE`   ROCKS
granite
    `USE`   ROCKS
slate
    `USE`   ROCKS

Συσχέτιση ιεραρχίας

Οι συσχετίσεις ιεραρχίας είναι το χαρακτηριστικό που ξεχωρίζει ένα θησαυρό από ένα λεξικό ή μια λίστα όρων [Sve89] και ο βασικός μηχανισμός δόμησης θησαυρών [ISO86]. Δύο όροι βρίσκονται σε ιεραρχική συσχέτιση όταν ο ένας απ' αυτούς — ο *γενικότερος όρος (broader term)* αντιπροσωπεύει μια κλάση αντικειμένων, ενώ ο δεύτερος —ο *ειδικότερος όρος (narrower term)* ένα μέλος ή υποσύνολο ή τμήμα του γενικότερου όρου. Γενικά μπορούμε να πούμε οτι δύο όροι συνδέονται με μια ιεραρχική συσχέτιση όταν ο ένας υπάγεται στον άλλο κατά μία έννοια. Η συσχέτιση ιεραρχίας υποδηλώνεται απ' την πλευρά του ειδικότερου όρου με το σύμβολο `BT`, ενώ απ' την πλευρά του γενικότερου όρου με το σύμβολο `NT`. Με βάση τη σημασιολογία τους, οι ιεραρχικές συσχετίσεις διακρίνονται σε τρεις τύπους.

1. *Ιεραρχική συσχέτιση γενίκευσης (generic)* η οποία συνδέει ένα όρο που περιγράφει ένα σύνολο οντοτήτων κ' ένα όρο που περιγράφει ένα υπερούνολό αυτού του συνόλου και υποδεικνύεται απ' τα σύμβολα `BTG/NTG` (Broader/Narrower Term Generic) όπως φαίνεται στο παράδειγμα 2.2.

2. *Ιεραρχική συσχέτιση μέρους-όλου (part-whole)*, η οποία συνδέει τον όρο που περιγράφει ένα τμήμα μιας οντότητας και τον όρο που περιγράφει την οντότητα αυτή, η οποία υποδεικνύεται απ' τα σύμβολα BTP/NTP (Broader/Narrower Term Partitive) όπως φαίνεται στο παράδειγμα 2.3.

3. *Ιεραρχική συσχέτιση παραδείγματος (instance-of)* η οποία συνδέει τον όρο που περιγράφει ένα σύνολο οντοτήτων και τον όρο που περιγράφει μια οντότητα αυτού του συνόλου. Για παράδειγμα
   ALPS BT MOUNTAIN REGIONS.

<u>ΠΑΡΑΔΕΙΓΜΑ 2.2</u>

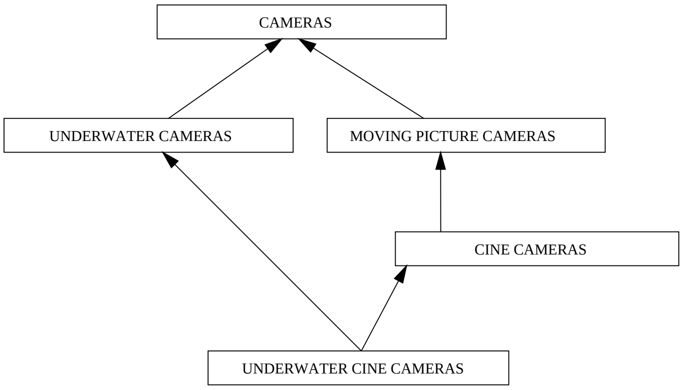
RATS
  BTG RODENTS
RODENTS
  NTG RATS

Στους περισσότερους θησαυρούς οι ιεραρχικές συσχετίσεις έχουν την σημασιολογία της ιεραρχικής συσχέτισης γενίκευσης, αν και μπορεί να υποδηλώνονται με τα σύμβολα BT/NT. Για το λόγο αυτό στα επόμενα θα αναφερόμαστε σε μια ιεραρχική συσχέτιση με τον όρο συσχέτιση γενίκευσης/ειδίκευσης (generalization/specialization relationship) ή απλά συσχέτιση γενίκευσης. Η σχέση γενίκευσης —το σύνολο δηλαδή όλων των συσχετίσεων γενίκευσης ενός θησαυρού— κατέχει την μεταβατική και την αντισυμμετρική ιδιότητα. Κατά συνέπεια, σχηματίζει μια ιεραρχία η οποία μπορεί να παρασταθεί μ' έναν ακυκλικό κατευθυνόμενο γράφο (Directed Acyclic Graph, DAG).

<u>ΠΑΡΑΔΕΙΓΜΑ 2.3</u>

CENTRAL NERVOUS SYSTEM
  BTP NERVOUS SYSTEM
NERVOUS SYSTEM
  NTP CENTRAL NERVOUS SYSTEM

Οριομένοι συγγραφείς [Maz94] θεωρούν πως κάθε όρος πρέπει να έχει το πολύ ένα γενικότερο όρο, ενώ άλλοι [MR88] θεωρούν πως όλοι οι γενικότεροι όροι κάθε όρου πρέπει να βρίσκονται στο ίδιο επίπεδο ιεραρχίας —άποψη εξίσου περιοριστική με την προηγούμενη αφού το απλό παράδειγμα του σχήματος 2.1 δεν θα μπορούσε ν' αποδοθεί, καθώς ο όρος "UNDERWATER CINE CAMERAS" έχει γενικότερους όρους οι οποίοι δεν βρίσκονται στο ίδιο επίπεδο ιεραρχίας. Το πρότυπο [ISO86] ωστόσο επιτρέπει πολλαπλή ιεραρχικότητα.



Σχήμα 2.1: Γραφική παράσταση συσχετίσεων γενίκευσης.

Οι συσχετίσεις BT σχηματίζουν ένα ακυκλικό κατευθυνόμενο γράφο που θα ονομάζουμε *ιεραρχία γενίκευσης/ειδίκευσης (generalization/specialization hierarchy).*

Συσχέτιση συνάφειας

Δύο όροι συνδέονται με μια συσχέτιση συνάφειας όταν στην καθημερινή τους χρήση, συνδέονται με μια συσχέτιση η οποία ωστόσο δεν είναι ούτε συσχέτιση ισοδυναμίας ούτε συσχέτιση ιεραρχίας. Η σχέση συνάφειας σ' ένα θησαυρό είναι κατά συνέπεια η ένωση ενός συνόλου σχέσεων. Για το λόγο αυτό η σχέση συνάφειας κατέχει μονάχα την ιδιότητα της συμμετρίας με την έννοια οτι όποιου τύπου συσχέτιση και να συνδέει δύο όρους πάντα μπορούμε να ορίσουμε την αντίστροφη αυτής, η οποία εφόσον δεν είναι συσχέτιση ισοδυναμίας ή συσχέτιση ιεραρχίας, θα είναι αναγκαστικά συσχέτιση συνάφειας. Για παράδειγμα οι όροι "TEMPERATURE CONTROL" και "THERMOSTATS" είναι συναφείς κατά την έννοια οτι ο έλεγχος της θερμοκρασίας εκτελείται από θερμοστάτες και οι θερμοστάτες εκτελούν τον έλεγχο θερμοκρασίας. Η συσχέτιση συνάφειας εδώ εκφράζει την ύπαρξη της συσχέτισης *εκτελεί* και *εκτελείται από* μεταξύ των δύο όρων. Η συσχέτιση συνάφειας μεταξύ δύο όρων υποδηλώνεται με το σύμβολο RT.

ΠΑΡΑΔΕΙΓΜΑ 2.4

TEMPERATURE CONTROL
    RT   THERMOSTATS
THERMOSTATS
    RT   TEMPERATURE CONTROL

2.2 Συγχώνευση θησαυρών

Η συγχώνευση θησαυρών εμπλέκει μια σειρά θεμάτων που θα πρέπει ν' αντιμετωπιστούν επιτυχώς. Αρχικά θα πρέπει να εντοπίσουμε όρους οι οποίοι περιγράφουν την ίδια έννοια στους συγχωνευόμενους θησαυρούς. Αυτοί οι όροι θα πρέπει να ενοποιηθούν. Θα πρέπει δηλαδή να δημιουργηθεί ένας νέος όρος ο οποίος θα συνδυάζει τις συσχετίσεις που έχουν οριστεί σε κάθε θησαυρό για κάθε έναν από τους όρους που ενοποιούνται. Στην περίπτωση που μια τέτοια ενοποίηση έρχεται σε σύγκρουση με κάποιον απ' τους κανόνες δόμησης θησαυρών, τότε αυτή η σύγκρουση θα πρέπει να αρθεί. Προκειμένου να μελετήσουμε τρόπους αντιμετώπισης των παραπάνω θεμάτων, διακρίνουμε την συγχώνευση θησαυρών σε μια ακολουθία φάσεων, τις οποίες ονομάζουμε *προενοποίηση (preintegration)*, *ανάλυση (analysis)*, *αναδιάρθρωση (conformation)* και *ενοποίηση (integration)*[2].

Προενοποίηση. Η διαδικασία συγχώνευσης διευκολύνεται αν οι συγχωνευόμενοι θησαυροί ακολουθούν το ίδιο μοντέλο και σχήμα δεδομένων. Αντικείμενο της φάσης προενοποίησης είναι η μετάφραση των θησαυρών σε κοινό μοντέλο δεδομένων και εννοιολογικό σχήμα.

Ανάλυση. Κατά κανόνα η διαδικασία συγχώνευσης θησαυρών υλοποιείται ως επαναληπτική ενοποίηση όρων που αποδίδουν την ίδια έννοια καθώς και των συσχετίσεων που αυτοί ορίζουν. Ο εντοπισμός τέτοιων όρων στους συγχωνευόμενους θησαυρούς είναι ο στόχος της φάσης ανάλυσης.

Αναδιάρθρωση. Οταν υπάρχουν όροι που αναφέρονται στην ίδια έννοια, αλλά ωστόσο οι συσχετίσεις που ορίζουν δεν είναι συμβατές μεταξύ τους διότι παραβιάζουν μια ή περισσότερες αρχές οργάνωσης θησαυρών, τότε είναι απαραίτητη η τροποποίησή τους ώστε να γίνουν συμβατές. Τέτοιου είδους τροποποιήσεις εκτελούνται στην φάση της αναδιάρθρωσης.

Ενοποίηση. Πρόκειται συγκριτικά για την πιο απλή φάση της διαδικασίας συγχώνευσης. Στόχος αυτής είναι η παραγωγή ενός νέου όρου που συνδυάζει τις συσχετίσεις των όρων που ενοποιούνται.

Στην ενότητα αυτή θ' ασχοληθούμε με την ανασκόπηση της βιβλιογραφίας σχετικά με τα θέματα εντοπισμού και ανάλυσης ομοιοτήτων, αναδιάρθρωσης και ενοποίησης. Η βιβλιογραφία που θα παρουσιάσουμε καλύπτει ένα ευρύτερο φάσμα απ' αυτό της συγχώνευσης θησαυρών, γεγονός που οφείλεται σε δύο κυρίως λόγους:

1. Θέματα που εμπλέκονται στην διαδικασία συγχώνευσης θησαυρών, παρουσιάζονται και σε διαφορετικά ερευνητικά πεδία όπως η ενοποίηση σχημάτων (schema integration), η ανάκληση πληροφορίας, η αναλογική ομοιότητα (analogical similarity) κλπ.

2. Απ' όσο γνωρίζουμε, η συγχώνευση θησαυρών δεν έχει αντιμετωπιστεί συστηματικά στην βιβλιογραφία, πράγμα που αφήνει ελεύθερο πεδίο για την εξερεύνηση άλλων ερευνητικών περιοχών με σκοπό την άντληση ιδεών για μια βελτιωμένη

[2] Δανειζόμαστε την διάκριση αυτή από την βιβλιογραφία ενοποίησης όψεων και βάσεων δεδομένων.

επίλυση του προβλήματος.

2.2.1 Ανάλυση

Η ανακάλυψη των κοινών εννοιών μεταξύ διαφορετικών θησαυρών είναι μια επίπονη και χρονοβόρα διαδικασία για ανθρώπους [ISO85] κυρίως λόγω του όγκου πληροφορίας που πρέπει να επεξεργαστούν. Μια προσέγγιση για τον μηχανικό εντοπισμό κοινών εννοιών μεταξύ θησαυρών είναι να χρησιμοποιήσουμε μια συλλογή κειμένων τα οποία έχουν ευρετηριαστεί με όρους που προέρχονται απ' τους συγχωνευόμενους θησαυρούς. Χρειαζόμαστε επιπλέον για κάθε κείμενο, το σύνολο των όρων κάθε θησαυρού με τους οποίους έχει ευρετηριαστεί. Συγκρίνοντας τα σύνολα αυτά μπορούμε να εντοπίσουμε όρους οι οποίοι είναι πιθανό να περιγράφουν την ίδια έννοια. Η προσέγγιση αυτή έχει το μειονέκτημα οτι απαιτεί μια συλλογή κειμένων τα οποία έχουν ευρετηριατεί ξεχωριστά με όρους από δύο ή περισσότερους θησαυρούς, κάτι που δεν είναι πάντα διαθέσιμο. Επιπλέον δεν λαμβάνει υπ' όψη του την δομή των θησαυρών.

Μια άλλη προσέγγιση είναι να θεωρήσουμε κάθε όρο και τις συσχετίσεις που αυτός ορίζει σαν μια αφηρημένη περιγραφή μιας έννοιας, και ν' αναπτύξουμε ένα μοντέλο υπολογισμού ενός μέτρου ομοιότητας μεταξύ των περιγραφών των εννοιών, στηριζόμενοι στην διαίσθηση που υπαγορεύει πως μια έννοια θα πρέπει λογικά να περιγράφεται παρόμοια σε διαφορετικούς θησαυρούς.

Αυτή η προσέγγιση χρησιμοποιείται πολύ συχνά σε πεδία εφαρμογών στα οποία απαιτείται ο εντοπισμός κοινών αντικειμένων μέσω δομημένων περιγραφών τους. Τέτοια πεδία εφαρμογών είναι η αναχρησιμοποίηση αντικειμένων λογισμικού [SC96], [GI94], η ενοποίηση εννοιολογικών σχημάτων βάσεων δεδομένων [BL86], [GLN92], η ανάκληση πληροφορίας [Pai91], [Sal89] κλπ. Το πρόβλημα με την προσέγγιση αυτή έγκειται στο γεγονός οτι ένα αντικείμενο μπορεί να παρασταθεί με περιγραφές που ενδέχεται να διαφέρουν σημαντικά μεταξύ τους. Αυτό οφείλεται στους εξής λόγους:

1. Στον καθορισμό των χαρακτηριστικών γνωρισμάτων και των συσχετίσεων μεταξύ των αντικειμένων ενός κόσμου, εμπλέκονται υποκειμενικά κριτήρια.

2. Ο σκοπός χρήσης, η εμβέλεια και επιλογές σχεδίασης μοντέλων ενός κόσμου διαφέρουν από εφαρμογή σ' εφαρμογή.

Παρά το βασικό αυτό μειονέκτημα η προσέγγιση αυτή έχει ευρεία χρήση λόγω της συγγένειάς της με τον τρόπο, με τον οποίο οι άνθρωποι αναγνωρίζουν όμοια ή ανάλογα αντικείμενα και γι' αυτό την υιοθετούμε και στην εργασία μας.

Εντοπισμός ομοιοτήτων στην ενοποίηση εννοιολογικών σχημάτων

Οι Gotthard, Lockemann και Neufeld, στο άρθρο [GLN92] παρουσιάζουν μια μεθοδολογία για την ενοποίηση εννοιολογικών σχημάτων που ακολουθούν το μοντέλο οντοτήτων–συσχετίσεων[3]. Στην εργασία αυτή, χρησιμοποιούνται τα αναγνωριστικά και τα χαρακτηριστικά των κατασκευών προκειμένου να συναχθούν συμπεράσματα σχετικά με την ομοιότητά τους. Ενα σύνολο κανόνων χρησιμοποιείται για την αναγνώριση ομοιοτήτων μεταξύ κατασκευών. Συγκεκριμένα:

1. Δύο γνωρίσματα είναι όμοια αν έχουν έχουν το ίδιο αναγνωριστικό ή πεδίο τιμών.
2. Δύο ρόλοι είναι όμοιοι αν έχουν το ίδιο αναγνωριστικό ή οι οντότητες οι οποίες συμμετέχουν σ' αυτούς είναι όμοιες.
3. Δύο οντότητες είναι όμοιες αν έχουν το ίδιο αναγνωριστικό ή όμοια γνωρίσματα ή συμμετέχουν σε όμοιους ρόλους ή είναι ειδικεύσεις ή γενικεύσεις όμοιων οντοτήτων.

Στο άρθρο [BL84] οι Batini και Lenzerini —οι οποίοι επίσης χρησιμοποιούν το μοντέλο οντοτήτων συσχετίσεων— εφοδιάζουν κάθε κατασκευή με ένα σύνολο συνωνύμων εκτός από το αναγνωριστικό. Συμπεράσματα σχετικά με την ομοιότητα αντικειμένων εξάγονται από την σύγκριση των αναγνωριστικών και των συνωνύμων τους, σε συνδυασμό με τα αντίστοιχα σύνολα ιδιοτήτων τους. Πιο συγκεκριμένα δύο κατασκευές είναι όμοιες αν έχουν κοινό αναγνωριστικό ή κοινό συνώνυμο ή κοινό σύνολο ιδιοτήτων. Είναι φανερό οτι η ιδέα είναι βασικά όμοια με την προηγούμενη, με την διαφορά οτι τα κριτήρια είναι σαφώς πιο περιοριστικά.

Ο εντοπισμός όμοιων αντικειμένων στην εργασία των Dayal και Hwang [DH84], είναι πανομοιότυπος με την εργασία των Batini και Lenzerini, με την διαφορά ότι δεν χρησιμοποιούνται συνώνυμα, αλλά μόνο αναγνωριστικά για τις κατασκευές.

Εντοπισμός ομοιοτήτων από περιγραφές σε φυσική γλώσσα

Οι Girardi και Ibrahim στο άρθρο [GI94], περιγράφουν το μηχανισμό ανάκλησης του συστήματος αναχρησιμοποίησης λογισμικού ROSA (Reuse Of Software Artifacts) στο οποίο, αντικείμενα λογισμικού (προδιαγραφές, λεπτομερή σχέδια, κώδικας) ανακαλούνται από ερωτήσεις (περιγραφές) φυσικής γλώσσας. Η ομοιότητα μεταξύ περιγραφών

[3] Οι κατασκευές αυτού του μοντέλου για την αναπαράσταση αντικειμένων, είναι οι οντότητες, οι συσχετίσεις μεταξύ οντοτήτων, οι ρόλοι οντοτήτων που λαμβάνουν μέρος σε μια συσχέτιση και τα γνωρίσματα οντοτήτων και συσχετίσεων. Κάθε κατασκευή έχει ένα αναγνωριστικό και ένα σύνολο χαρακτηριστικών. Για παράδειγμα χαρακτηριστικά ενός γνωρίσματος είναι το πεδίο τιμών του, ενώ για ένα τύπο οντοτήτων, χαρακτηριστικά είναι τα γνωρίσματά του και οι συσχετίσεις του [TL82].

ορίζεται ως μια μονότονα φθίνουσα συνάρτηση της απόστασής τους. Βασικό ρόλο στην προσέγγιση αυτή παίζει ο υπολογισμός της απόστασης μεταξύ επιθετικών φράσεων, που αποτελούνται από ένα ουσιαστικό που ονομάζεται *κεφαλή* και ένα σύνολο επιθετικών προσδιορισμών που ονομάζονται *προσδιοριστές*. Για παράδειγμα στον σύνθετο όρο "Moving Picture Cameras", η κεφαλή είναι ο όρος "Cameras" ενώ οι προσδιοριστές είναι οι όροι "Moving" και "Picture".

Αν ένας θησαυρός είναι διαθέσιμος, τότε η απόσταση ανάμεσα σε δύο απλούς όρους ορίζεται να είναι (α) μηδέν αν οι όροι είναι (αλφαβητικά) ίσοι ή είναι συνώνυμα, (β) το μήκος του μικρότερου ιεραρχικού μονοπατιού (αν τέτοιο υπάρχει) που τους συνδέει, (γ) άπειρη σε κάθε άλλη περίπτωση. Η απόσταση ανάμεσα σε δύο σύνολα απλών όρων A και B ορίζεται ως το ελάχιστο άθροισμα των αποστάσεων μεταξύ ζευγών όρων απ' όλα τα δυνατά σύνολα διαφορετικών ζευγών όρων των A και B. Συνδυάζοντας τις δύο παραπάνω αποστάσεις, η απόσταση δύο φράσεων ορίζεται ως το άθροισμα της απόστασης των κεφαλών τους με την απόσταση των συνόλων προσδιοριστών τους.

Αναλογική Ομοιότητα Αντικειμένων

Ο Σπανουδάκης στην διδακτορική του διατριβή [Spa94], προτείνει ένα μοντέλο εντοπισμού αναλογικών ομοιοτήτων μεταξύ αντικειμένων βασιζόμενο σε εννοιολογικές/σημασιολογικές περιγραφές που κατασκευάζονται με την χρήση των τριών μηχανισμών αφαίρεσης του μοντέλου δεδομένων της γλώσσας Telos [MBJK90], [DKT95]. Συγκεκριμένα το μοντέλο αυτό υποθέτει οτι τα συγκρινόμενα αντικείμενα περιγράφονται μέσω ταξινόμησης (classification), γενίκευσης (generalization) και γνωρισματοδότησης (attribution).

Η ομοιότητα αντικειμένων ορίζεται ως μια μονότονα φθίνουσα συνάρτηση της απόστασής τους. Η απόσταση δύο αντικειμένων αποτελεί την συνάθροιση τριών διαφορετικών αποστάσεων: ταξινόμησης, γενίκευσης και γνωρισμάτων.

Η απόσταση ταξινόμησης δύο αντικειμένων ορίζεται ως το άθροισμα της σπουδαιότητας των διαφορετικών κλάσεών τους. Η σπουδαιότητα μιας κλάσης εξαρτάται από το πόσο γενική είναι, με τις γενικές κλάσεις να θεωρούνται πιο σπουδαίες διότι συγκεντρώνουν αφηρημένα γνωρίσματα τα οποία κληρονομούνται από τις ειδικότερες κλάσεις. Η απόσταση γενίκευσης ορίζεται ως το άθροισμα της σπουδαιότητας των διαφορετικών υπερκλάσεών τους. Η απόσταση γνωρισμάτων υπολογίζεται αναδρομικά με βάση την απόσταση ταξινόμησης, γενίκευσης, γνωρισμάτων καθώς επίσης και την απόσταση των τιμών τους.

Εντοπισμός ομοιοτήτων σε θησαυρούς

Στα άρθρα τους για την συγχώνευση και επαύξηση θησαυρών οι Mili, Rada και Martin [MR88], [RM89], [RM87], περιορίζουν τον εντοπισμό ομοιοτήτων μονάχα στους όρους και κατά συνέπεια μονάχα σε λεκτικά κριτήρια. Η σύγκριση δύο όρων γίνεται αφού εφαρμοστεί ένας αλγόριθμος αφαίρεσης καταλήξεων πληθυντικού αριθμού και παθητικής φωνής. Οι συσχετίσεις μεταξύ των εννοιών δεν λαμβάνονται υπόψη, αν και ορίζεται ένα μέτρο απόστασης μεταξύ όρων, ως το ελάχιστο πλήθος ιεραρχικών συσχετίσεων που τους συνδέει[4].

Σύνοψη

Στην παράγραφο αυτή θα προσπαθήσουμε να συνοψίσουμε τα πλεονεκτήματα και τα μειονεκτήματα κάθε μιας από τις προσεγγίσεις στο θέμα της ομοιότητας που παρουσιάστηκαν στα προηγούμενα, εστιάζοντας κυρίως στην εφαρμοσιμότητά τους στο πρόβλημα της συγχώνευσης θησαυρών. Τα κριτήρια με βάση τα οποία θα προσπαθήσουμε ν' αξιολογήσουμε τις προσεγγίσεις, είναι η γενικότητα, η πολυπλοκότητα, η προσαρμοσιμότητα και η ασάφεια της ομοιότητας.

Γενικότητα. Ανεξαρτησία δηλαδή μιας προσέγγισης από συγκεκριμένα πεδία εφαρμογών.

Πολυπλοκότητα. Λόγω του μεγάλου μεγέθους των θησαυρών η πολυπλοκότητα μιας μεθόδου υπολογισμού ομοιότητας πρέπει να είναι μικρή ή μέση. Μέση θα θεωρήσουμε μια πολυπλοκότητα που φράσσεται από ένα πολυώνυμο.

Προσαρμοσιμότητα. Θεωρούμε πως ένα μοντέλο υπολογισμού ομοιότητας εννοιών που περιγράφονται σε θησαυρούς, θα πρέπει να διαθέτει παραμέτρους ελέγχου της συμπεριφοράς του, ανάλογα με την περίπτωση.

Ασάφεια της ομοιότητας. Ενα μοντέλο υπολογισμού ομοιότητας εννοιών που περιγράφονται σε θησαυρούς θα πρέπει να έχει την δυνατότητα να δίνει μια ένδειξη της έντασης (aptness) της ομοιότητας και όχι να διαχωρίζει απλά μεταξύ όμοιων και ανόμοιων εννοιών.

Ξεκινώντας από τις προσεγγίσεις στο θέμα της ομοιότητας στο πλαίσιο ενοποίησης εννοιολογικών σχημάτων, πρέπει να παρατηρήσουμε πως ένα πλεονέκτημα το οποίο είναι κοινό και στις τρεις προσεγγίσεις είναι η απλότητα των κριτηρίων που χρησιμοποιούνται. Οταν η ομοιότητα εντοπίζεται στη βάση απλών κριτηρίων τότε και η υπολογιστική

[4] Εδώ θα πρέπει να παρατηρήσουμε μια ασυνέπεια μεταξύ του ορισμού αυτού και της απαίτησης οι γενικότεροι όροι ενός όρου να βρίσκονται στο ίδιο επίπεδο ιεραρχίας. Εφόσον όλοι οι γενικότεροι όροι κάθε όρου βρίσκονται στο ίδιο επίπεδο, το πλήθος των ιεραρχικών συσχετίσεων που συνδέει δύο τυχαίους όρους θα είναι το ίδιο, όποιο μονοπάτι κ' αν ακολουθηθεί.

πολυπλοκότητα —η οποία είναι σε πολλές περιπτώσεις το τελικό κριτήριο αποδοχής ή εφαρμοσιμότητας ενός αλγόριθμου— είναι σχετικά χαμηλή. Ωστόσο η απλότητα αυτή κοστίζει σ' ένα άλλο σημείο: Ενα προφανές κοινό μειονέκτημα των παραπάνω προσεγγίσεων είναι οτι η ομοιότητα μεταξύ δύο αντικειμένων είναι μια ιδιότητα η οποία είτε ισχύει, είτε όχι. Δεν παρέχεται κανένα είδος έντασης της ομοιότητας. Επιπλέον στις περιπτώσεις των άρθρων [BL84] και [DH84] ο υπολογισμός της ομοιότητας βασίζεται σε κριτήρια τόσο περιοριστικά ώστε να είναι απαγορευτική η χρήση τους για τον εντοπισμό κοινών εννοιών μεταξύ θησαυρών. Για παράδειγμα, σύμφωνα με τα κριτήρια ομοιότητας που υιοθετούν οι Batini και Lenzerini, δύο αντικείμενα με διαφορετικά αναγνωριστικά και σύνολα ιδιοτήτων που διαφέρουν κατά μία μονάχα ιδιότητα, θα θεωρηθούν διαφορετικά.

Η προσέγγιση των Girardi και Ibrahim [GI94] ταιριάζει απόλυτα στο περιβάλλον συγχώνευσης θησαυρών, διότι ένα βασικό στοιχείο για την περιγραφή μιας έννοιας σ' έναν θησαυρό είναι ο όρος που την παριστάνει. Επομένως η ομοιότητα μεταξύ όρων και μάλιστα βασισμένη σε σημασιολογικά κριτήρια είναι εξαιρετικά ελκυστική. Ωστόσο το πρόβλημα που εισάγει η συγκεκριμένη προσέγγιση, είναι οτι απαιτείται ένας επιπλέον θησαυρός για τον ορισμό των αποστάσεων των λέξεων που αποτελούν τους σύνθετους όρους. Δεδομένου οτι ένας τέτοιος θησαυρός θα είναι αναγκαστικά προσανατολισμένος στο πεδίο γνώσης που καλύπτουν οι συγχωνευόμενοι θησαυροί, η προσέγγιση υστερεί σε γενικότητα. Επιπλέον δεν πρέπει να ξεχνάμε οτι προκειμένου να αναγνωριστούν η κεφαλή και οι προσδιοριστές ενός σύνθετου όρου, θα πρέπει να προηγηθεί συντακτική ανάλυση η οποία απαιτεί μία ακόμη βάση γνώσης που θα περιγράφει την συντακτική κατηγορία κάθε απλού όρου και η οποία αναγκαστικά θα είναι προσανατολισμένη σ' ένα συγκεκριμένο πεδίο γνώσης. Ενα ακόμη πρόβλημα που υπάρχει είναι η σχετικά υψηλή πολυπλοκότητα του αλγόριθμου υπολογισμού της απόστασης μεταξύ σύνθετων όρων που οφείλεται αφενός στο οτι υπολογίζονται όλοι οι συνδυασμοί αποστάσεων μεταξύ των προσδιοριστών τους και αφετέρου στην απαιτούμενη συντακτική ανάλυση.

Η προσέγγιση των Mili, Rada και Martin, [MR88], [RM87] παρά το γεγονός οτι έχει χρησιμοποιηθεί σε συγχώνευση θησαυρών, έχει το σοβαρό μειονέκτημα οτι βασίζεται αποκλειστικά και μόνο στους όρους και δεν αξιοποιεί τις συσχετίσεις μεταξύ των εννοιών. Επίσης θεωρούμε οτι είναι μάλλον ανεπαρκής για την αναγνώριση όμοιων όρων αφού δεν αντιμετωπίζει το πρόβλημα της σύνταξης των σύνθετων όρων. Αντιθέτως η χρήση αλγορίθμων αποκοπής καταλήξεων είναι μια δοκιμασμένη τεχνική για την μείωση της πολυπλοκότητας της προσεγγιστικής σύγκρισης όρων με σχετικά καλά αποτελέσματα και φαίνεται να είναι μια ιδέα που αξίζει να υιοθετηθεί.

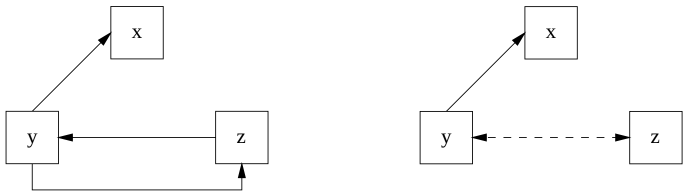
Η προσέγγιση του Σπανουδάκη, συνδυάζει μερικά πολύ επιθυμητά χαρακτηριστικά. Αρχικά μπορεί να μεταφερθεί στο πλαίσιο των θησαυρών. Υπενθυμίζουμε ότι η σημασιολογία της συσχέτισης γενίκευσης στην γλώσσα Telos είναι ίδια μ' αυτήν της γενικής ιεραρχικής σχέσης σε θησαυρούς. Επίσης η έννοια της ταξινόμησης υπάρχει και στο πλαίσιο των θησαυρών και αφορά την κατάταξη μιας έννοιας σε μια όψη. Το μοντέλο ομοιότητας μπορεί επιπλέον να εφοδιαστεί με παραμέτρους προκειμένου να ρυθμίζεται η συμπεριφορά του ανάλογα με την περίπτωση. Τέλος η πολυπλοκότητα του μοντέλου είναι μέτρια δεδομένου οτι στην περίπτωση θησαυρών δεν χρειάζεται να χρησιμοποιηθεί ο αλγόριθμος υπολογισμού της απόστασης γνωρισμάτων, ο οποίος είναι το πολυπλοκότερο τμήμα του μοντέλου.

	Γενικότητα	Προσαρ/τητα	Ασάφεια	Πολυπλοκότητα
[BL86]	++	-	-	+
[GLN92]	++	-	-	+
[DH84]	++	-	-	+
[MR88]	+	+	-	++
[Spa94]	++	++	++	+
[GI94]	-	++	++	-

Πίνακας 2.1: Σύγκριση των χαρακτηριστικών διαφόρων μεθόδων εντοπισμού ομοιοτήτων

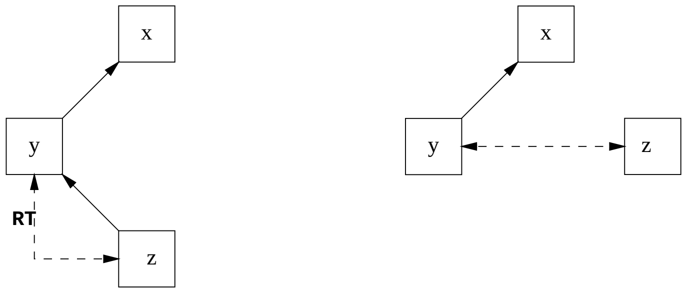
2.2.2 Αναδιάρθρωση

Πριν την ενοποίηση δύο όρων οι οποίοι παριστάνουν την ίδια έννοια και των αντίστοιχων συσχετίσεων, θα πρέπει να ανιχνευθούν και να επιλυθούν τυχόν συγκρούσεις ονομασίας ή δομικές συγκρούσεις. Μια σύγκρουση ονομασίας συμβαίνει αν οι όροι που αναπαριστούν την έννοια είναι διαφορετικοί. Στην περίπτωση αυτή η σύγκρουση επιλύεται με τον καθορισμό ενός μοναδικού όρου για την αναπαράσταση της έννοιας. Μια δομική σύγκρουση συμβαίνει όταν οι συσχετίσεις που ορίζουν οι συγχωνευόμενοι όροι, δεν είναι συνεπείς με τις αρχές δόμησης θησαυρών. Κατά κανόνα τέτοιες συγκρούσεις επιλύονται με την συμβολή ανθρώπινου δυναμικού. Ωστόσο είναι δυνατό ν' ακολουθηθούν και προκαθορισμένοι κανόνες επίλυσης. Την προσέγγιση αυτή ακολουθούν οι Mili και Rada, οι οποίοι διακρίνουν σε τρεις τύπους δομικών συγκρούσεων που παρουσιάζονται στα σχήματα 2.2-2.4 μαζί με τον τρόπο επίλυσής τους. Είναι φανερό οτι ο τρόπος επίλυσης που υιοθετείται ενδέχεται να εισάγει προβλήματα στην δομή του θησαυρού —ας μην ξεχνάμε οτι η βασικός μηχανισμός δόμησης των θησαυρών είναι οι ιεραρχικές συσχετίσεις. Μια τέτοια περίπτωση παρουσιάζεται στο σχήμα 2.5.



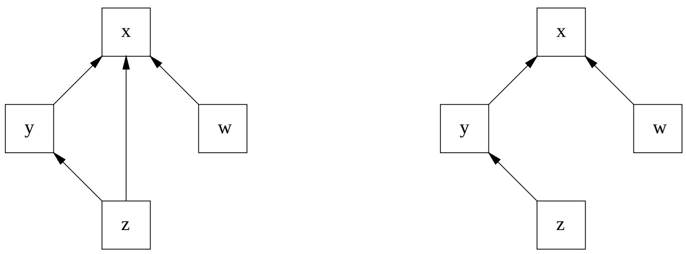
(α) (β)

Σχήμα 2.2: Σύγκρουση ιεραρχικών συσχετίσεων
(α) Ο όρος y είναι ταυτόχρονα γενικότερος και ειδικότερος απ' τον όρο z. (β) Επίλυση στην εργασία [MR88]: Οι συγκρουόμενες ιεραρχικές συσχετίσεις διαγράφονται και εισάγεται μια συσχέτιση συνάφειας.



(α) (β)

Σχήμα 2.3: Σύγκρουση τύπου BT/RT
(α) Ο όρος y είναι ταυτόχρονα γενικότερος και συναφής όρος του όρου z. (β) Επίλυση στην εργασία [MR88]: Η ιεραρχική συσχέτιση μεταξύ των y και z διαγράφεται.



(α) (β)

Σχήμα 2.4: Επίλυση σύγκρουσης ιεραρχικών συσχετίσεων στην εργασία [MR88]
(α) Ο όρος z έχει δύο γενικότερους όρους οι οποίοι βρίσκονται σε διαφορετικά επίπεδα της ιεραρχίας.
(β) Επίλυση: Η συσχέτιση προς τον πιο γενικό απ' τους γενικότερους όρους του z διαγράφεται.

2.2.3 Ενοποίηση

Στην εργασία [MR88], ένας όρος και οι συσχετίσεις που αυτός ορίζει υλοποιούνται με πλαίσια. Η ενοποίηση όρων που αναπαριστούν την ίδια έννοια γίνεται με την ενοποίηση των αντίστοιχων πλαισίων, με τον υπολογισμό δηλαδή της ένωσης των αντίστοιχων σχισμών (slots). Το αποτέλεσμα παριστάνεται γραφικά στο σχήμα 2.6.

Οι Mannino, Navathe και Effelsberg στο άρθρο τους "A Rule-Based Approach for Merging Generalization Hierarchies" [MNE88], περιγράφουν μια μέθοδο συγχώνευσης ιεραρχιών γενίκευσης στο πλαίσιο ενοποίησης όψεων. Η συσχέτιση γενίκευσης στην εργασία τους έχει σημασιολογία υποσυνόλου όπως και η γενική ιεραρχική συσχέτιση σε θησαυρούς. Ορίζεται ένα σύνολο πράξεων ενημέρωσης των ιεραρχιών το οποίο μπορεί είτε να χρησιμοποιηθεί σε συνδυασμό με κανόνες συγχώνευσης, είτε ανεξάρτητα για την διαλογική συγχώνευση των ιεραρχιών. Το σύνολο των πράξεων ενημέρωσης παρέχει την δυνατότητα

- Δημιουργίας υπερκλάσης ενός συνόλου κλάσεων.
- Διαγραφή μιας κλάσης και των υποκλάσεών της από την ιεραρχία.
- Μετακίνηση μιας κλάσης στην ιεραρχία.
- Διαγραφή όλων των κλάσεων ενός επιπέδου γενίκευσης.

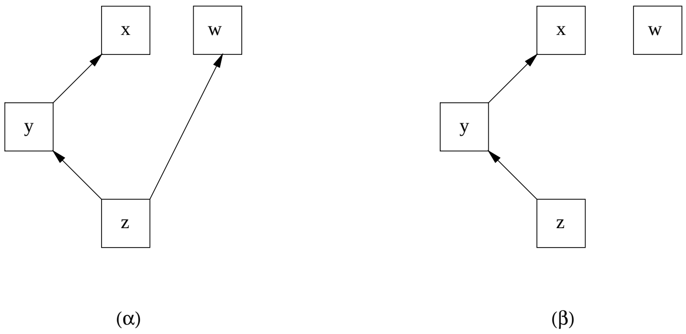
2.3 Σύνοψη

Κλείνοντας το Κεφάλαιο αυτό, θα σκιαγραφήσουμε τα βασικά χαρακτηριστικά που πρέπει να έχει μια μέθοδος συγχώνευσης θησαυρών. Πρόκειται για τις κατευθυντήριες γραμμές που ακολουθούμε στην εργασία αυτή.

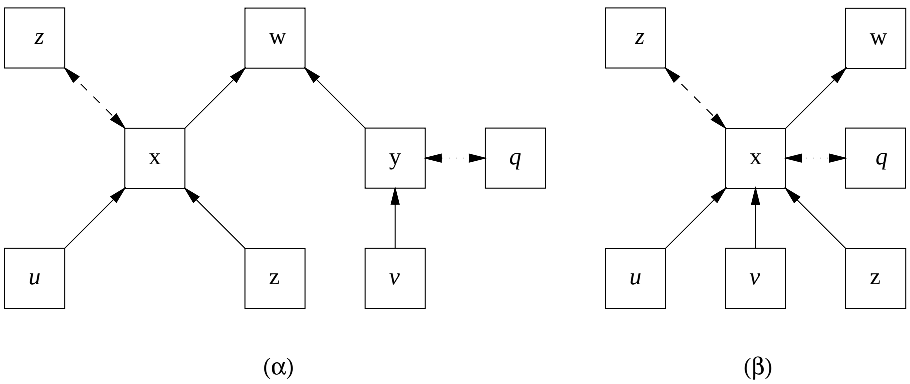
Η κατασκευή ενός θησαυρού, απ' την επιλογή του λεξιλογίου ως τον ορισμό των συσχετίσεων, είναι μια εργασία καθαρά γνωσιακή. Είναι κατά συνέπεια εξαιρετικά δύσκολο να κατασκευαστεί μηχανικά ένας "καλός" θησαυρός, χωρίς την παρέμβαση ανθρώπων στην διαδικασία της κατασκευής. Επομένως μια μηχανική μέθοδος κατασκευής ενός θησαυρού —όπως η συγχώνευση— δεν πρέπει να στοχεύει κατά την άποψή μας στην παραγωγή ενός "τέλειου" αποτελέσματος, αλλά στην υποστήριξη των ειδικών επιστημόνων που ασχολούνται με την κατασκευή θησαυρών. Πιο συγκεκριμένα αυτό που θα πρέπει να παρέχει μια μέθοδος μηχανικής συγχώνευσης θησαυρών είναι η γρήγορη κατασκευή του σκελετού ενός θησαυρού συνδυάζοντας την δομή άλλων. Οι όποιες ρυθμίσεις βελτίωσης στην συνέχεια ανήκουν στους ειδικούς. Τί είναι όμως ο σκελετός ενός θησαυρού; Αναμφίβολα ο βασικός μηχανισμός δόμησης θησαυρών είναι οι ιεραρχικές συσχετίσεις [Sve89], [ISO86]. Σε πολλές περιπτώσεις ο βαθμός ανάλυσης της ιεραρχίας που σχηματίζουν, είναι το βασικό κριτήριο της ποιότητας ενός θησαυρού [Soe95].

Ωστόσο ακόμη και αυτός ο περιορισμός της εμβέλειας μιας μεθόδου συγχώνευσης θησαυρών, δεν επιτρέπει την ολοκλήρωση της διαδικασίας χωρίς την ανάμειξη ανθρώπινου δυναμικού. Η αιτία περιγράφηκε στα προηγούμενα: είναι δυνατό να παρουσιαστούν περιπτώσεις κατά τις οποίες η ενοποίηση όρων δεν μπορεί να πραγματοποιηθεί γιατί παραβιάζονται αρχές δόμησης θησαυρών. Μια προκαθορισμένη πολιτική μπορεί να οριστεί σ' αυτές τις περιπτώσεις, όπως γίνεται στην εργασία [MR88], αλλά κατά την άποψή μας δεν είναι αποτελεσματική, διότι καταστρέφει τον σκελετό που προσπαθούμε να σχηματίσουμε. Συνεπώς κατά την άποψή μας δύο είναι οι δυνατές επιλογές: Αν η διαδικασία της συγχώνευσης εκτελείται διαλογικά, τότε θα πρέπει να παρέχεται μέσω της διεπαφής χρήσης, ένα σύνολο πράξεων με την εφαρμογή των οποίων ειδικοί θα μπορούν να επιλύουν τυχόν συγκρούσεις, ενώ αν η συγχώνευση είναι δεσμική τότε οι συγκρούσεις θα πρέπει ν' αναφέρονται απλά και να μην γίνεται καμία προσπάθεια μηχανικής επίλυσής τους.

Απ' την μελέτη της σχετικής βιβλιογραφίας είναι ίσως φανερό πως το θέμα στο οποίο, μια μέθοδος συγχώνευσης μπορεί να προσφέρει περισσότερο είναι ο εντοπισμός αντίστοιχων, όμοιων ή ταυτόσημων όρων μεταξύ των συγχωνευόμενων θησαυρών. Οσο πιο αποδοτικά, γίνει αυτό, τόσο ουσιαστικότερη είναι η προσφορά της μεθόδου στο πολύπλοκο έργο της κατασκευής ή επαύξησης θησαυρών. Κατά την άποψή μας, στην προσπάθειά αυτή πρέπει ν' αξιοποιηθεί η πληροφορία (ή γνώση) που οι ίδιοι οι συγχωνευόμενοι θησαυροί περιέχουν. Επιπλέον θεωρούμε πως είναι αναγκαίο η διαδικασία εντοπισμού όμοιων όρων να έχει μια τάση βελτίωσης καθώς η συγχώνευση προχωρεί. Μια τέτοια ιδιότητα έχει το σημαντικό πλεονέκτημα, ότι η μέθοδος συγχώνευσης μπορεί να υποστηριχθεί αρχικά από ανθρώπινο δυναμικό και έπειτα από κάποιο σημείο προόδου να εκτελεστεί δεσμικά, αποδεσμεύοντας έτσι τους ανθρώπινους πόρους.



Σχήμα 2.5: Απώλεια πληροφορίας από την επίλυση σύγκρουσης ιεραρχικών συσχετίσεων
(α) Ο όρος z έχει δύο γενικότερους όρους οι οποίοι βρίσκονται σε διαφορετικά επίπεδα της ιεραρχίας. (β) Η διαγραφή της συσχέτισης απ' τον z στον w προκαλεί απώλεια πληροφορίας.



Σχήμα 2.6: Ενοποίηση όρων στην εργασία [MR88]
(α) Η κατάσταση πριν την ενοποίηση πλαισίων: Οι όροι x και y είναι ταυτόσημοι και τα αντίστοιχα πλαίσια θα ενοποιηθούν με τον υπολογισμό της ένωσης των αντίστοιχων σχισμών τους. (β) Η κατάσταση μετά την ενοποίηση.

<u>ΚΕΦΑΛΑΙΟ 3</u>

ΠΑΡΑΣΤΑΣΗ, ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΘΗΣΑΥΡΩΝ

3.1 Παράσταση θησαυρών[1]

<u>ΟΡΙΣΜΟΣ 3.1</u>

Ενας *θησαυρός* είναι μια εξάδα

$$\theta = (T, P, \mathcal{F}, G, E, A)$$

όπου: T είναι το σύνολο των όρων (λεξιλόγιο) του θ, P είναι το σύνολο των δόκιμων όρων,

$$P = \{x_1, x_2, \cdots, x_n\}$$

$T - P$ το σύνολο των αδόκιμων όρων,

$$T - P = \{u_1, u_2, \cdots, u_m\}$$

$\mathcal{F}$ είναι το σύνολο των κατηγοριών των όρων

$$\mathcal{F} = \{f_1, f_2, \cdots, f_l\}$$

και $G \subseteq P \times P$, $E \subseteq (T - P) \times P$, $A \subseteq P \times P$ είναι αντίστοιχα οι σχέσεις γενίκευσης, ισοδυναμίας και συνάφειας μεταξύ των όρων του θ.

[1]Στους ορισμούς που ακολουθούν τα σύμβολα $\boldsymbol{R}$ και $\boldsymbol{N}$ παριστάνουν το σύνολο των πραγματικών και φυσικών αριθμών αντίστοιχα.

ΟΡΙΣΜΟΣ 3.2

Κάθε όρος t ενός θησαυρού θ, απεικονίζεται σ' ένα φυσικό αριθμό $\#t$, τον οποίο θα ονομάζουμε *αναγνωριστή* του t, μέσω της αμφιμονοσήμαντης αντιστοιχίας $I_\theta : T \longrightarrow \boldsymbol{N}$, έτσι ώστε $I_\theta(t) = \#t,\ \forall t \in T$.

ΟΡΙΣΜΟΣ 3.3

Σε κάθε *δόκιμο όρο* x, αντιστοιχίζεται μια τετράδα

$$\boldsymbol{x} = (\#x, F, B, R)$$

όπου F, είναι το σύνολο κατηγοριών στις οποίες ανήκει ο x,

B είναι το σύνολο των άμεσα γενικότερων όρων του x,

$$B = \{\#y : y \in P \text{ είναι άμεσα γενικότερος όρος του } x\}$$

και R είναι το σύνολο των συναφών όρων του x.

$$R = \{\#y : y \in P \text{ είναι συναφής όρος του } x\}$$

ΟΡΙΣΜΟΣ 3.4

Σε κάθε *αδόκιμο όρο* u, αντιστοιχίζεται μια τριάδα

$$\boldsymbol{u} = (\#u, r, U)$$

όπου

$$U = \begin{cases} \{x \in P : x \text{ είναι ισοδύναμος δόκιμος όρος του } u\} & \text{αν } r = 0 \\ \{x \in P : x \text{ είναι συντακτικός παράγοντας του } u\} & \text{αν } r = 1 \end{cases}$$

Στο εξής, θα θεωρούμε τους συμβολισμούς x και $\boldsymbol{x}$ ως ισοδύναμους και κατά συνέπεια θα γράφουμε $x.B$ αντί $\boldsymbol{x}.B$, $u.U$ αντί $\boldsymbol{u}.U$ κλπ. Επίσης για λόγους ευκολότερης ανάγνωσης του κειμένου[2], στο εξής όταν είναι προφανές οτι αναφερόμαστε σ' ένα θησαυρό θ, δεν θα γράφουμε π.χ., $\theta.T$ ή $\theta.P$ αλλά T ή P αντίστοιχα.

3.2 Περιορισμοί ακεραιότητας

Για να είναι μια εξάδα $\theta = (T, P, \mathcal{F}, G, E, A)$ αποδεκτός θησαυρός θα πρέπει να πληρούνται συγκεκριμένες συνθήκες που ονομάζουμε περιορισμούς ακεραιότητας. Στην

[2]Ισως και για λόγους διευκόλυνσης του συγγραφέα.

εργασία αυτή υιοθετούμε τους περιορισμούς ακεραιότητας που περιγράφονται στο πρότυπο για την οργάνωση θησαυρών [ISO86] και στην εργασία [Mil91]. Οι περιορισμοί αυτοί, θα πρέπει να ισχύουν από την αρχική κατασκευή και να εξακολουθούν να ισχύουν έπειτα από κάθε πράξη ενημέρωσης του θησαυρού.

3.2.1 Συσχετίσεις γενίκευσης

Ενα μονοπάτι γενίκευσης μήκους l απ' τον όρο x_i στον όρο x_{i+l}, είναι μια ακολουθία όρων $x_i, x_{i+1}, \cdots, x_{i+l}$ έτσι ώστε $l \geq 1$ και $\#x_{j+1} \in x_j.B$, $\forall j \in [i, l-1]$. Δηλαδή ένα μονοπάτι γενίκευσης είναι μια ακολουθία όρων καθένας απ' τους οποίους ανήκει στο σύνολο των άμεσα γενικότερων όρων του προηγούμενού του. Αν υπάρχει ένα μονοπάτι γενίκευσης απ' τον x στον y, Θα γράφουμε $x \rightsquigarrow y$.

Κανένα μονοπάτι γενίκευσης δεν πρέπει να καταλήγει στον όρο απ' τον οποίο ξεκινά. Θα πρέπει συνεπώς να ισχύει

$$x \rightsquigarrow y \Rightarrow x \neq y \tag{3.1}$$

Αν ο θησαυρός θ ικανοποιεί την (3.1) θα γράφουμε $\theta \models \boldsymbol{A}$.

Αν x, y, z είναι δόκιμοι όροι τότε θα πρέπει να ισχύει

$$\#y, \#z \in x.B \Rightarrow y \not\rightsquigarrow z \text{ και } z \not\rightsquigarrow y \tag{3.2}$$

Αν ο θησαυρός θ ικανοποιεί την (3.2) θα γράφουμε $\theta \models \boldsymbol{M}$.

<u>ΟΡΙΣΜΟΣ 3.5</u>

Η *σχέση γενίκευσης* G, των όρων ενός θησαυρού θ, είναι ένα υποσύνολο του $P \times P$, που ορίζεται ως εξής

$$G = \{(x, y) : x \rightsquigarrow y\} \tag{3.3}$$

Η αντίστροφη της G, $S = G^{-1}$ είναι η σχέση ειδίκευσης. Παρατηρούμε οτι η G, είναι μια διάταξη του P: η συνθήκη (3.1) εξασφαλίζει την αντισυμμετρικότητα της G, η οποία είναι και μεταβατική αφού αν (x, y) και $(y, z) \in G$, τότε $x \rightsquigarrow y$ και $y \rightsquigarrow z$ άρα $x \rightsquigarrow z$ και συνεπώς $(x, z) \in G$. Θα συμβολίζουμε με $G^+(x)$ και $S^+(x)$ το *σύνολο των γενικότερων* και *ειδικότερων όρων* του όρου x αντίστοιχα. Δηλαδή,

$$\begin{aligned} G^+(x) &= \{y \in P : (x, y) \in G\} && (3.4) \\ S^+(x) &= \{y \in P : (y, x) \in G\} && (3.5) \end{aligned}$$

3.2.2 Συσχετίσεις ισοδυναμίας

Οι αδόκιμοι όροι ενός θησαυρού υπάρχουν πάντα σε αντιστοιχία μ' έναν ή περισσότερους δόκιμους όρους. Κατά συνέπεια κάθε αδόκιμος όρος πρέπει να ορίζει μια συσχέτιση ισοδυναμίας μ' ένα τουλάχιστον δόκιμο όρο. Θα πρέπει δηλαδή να ισχύει

$$u \in T - P \Rightarrow u.U \neq \emptyset \tag{3.6}$$

Αν ο θησαυρός θ ικανοποιεί την (3.6) θα γράφουμε $\theta \models \boldsymbol{E}$.

ΟΡΙΣΜΟΣ 3.6

Η *σχέση ισοδυναμίας* E των όρων ενός θησαυρού θ, είναι ένα υποσύνολο του $(T-P)\times T$, που ορίζεται ως εξής:

$$\mathrm{E} = \{(u,x) : u \in T - P \text{ και } x \in P \text{ και } \#x \in u.U\}$$

3.2.3 Συσχετίσεις συνάφειας

ΟΡΙΣΜΟΣ 3.7

Η *σχέση συνάφειας* A των όρων ενός θησαυρού θ, είναι ένα υποσύνολο του $P \times P$, κάθε στοιχείο του οποίου ικανοποιεί τις παρακάτω σχέσεις

$$\begin{aligned} y \in P \text{ και } \#y \in x.R &\Rightarrow (x,y) \in A \\ (x,y) \in A &\Rightarrow (y,x) \in A \end{aligned}$$

3.2.4 Αλλες συνθήκες

Δύο όροι που λαμβάνουν μέρος σε μια συσχέτιση ενός τύπου δεν μπορούν να λαμβάνουν μέρος σε συσχέτιση διαφορετικού τύπου. Δεδομένου οτι οι αδόκιμοι όροι δεν λαμβάνουν μέρος παρά σε συσχετίσεις ισοδυναμίας, η παρακάτω σχέση αρκεί για να τηρηθεί αυτή η αρχή.

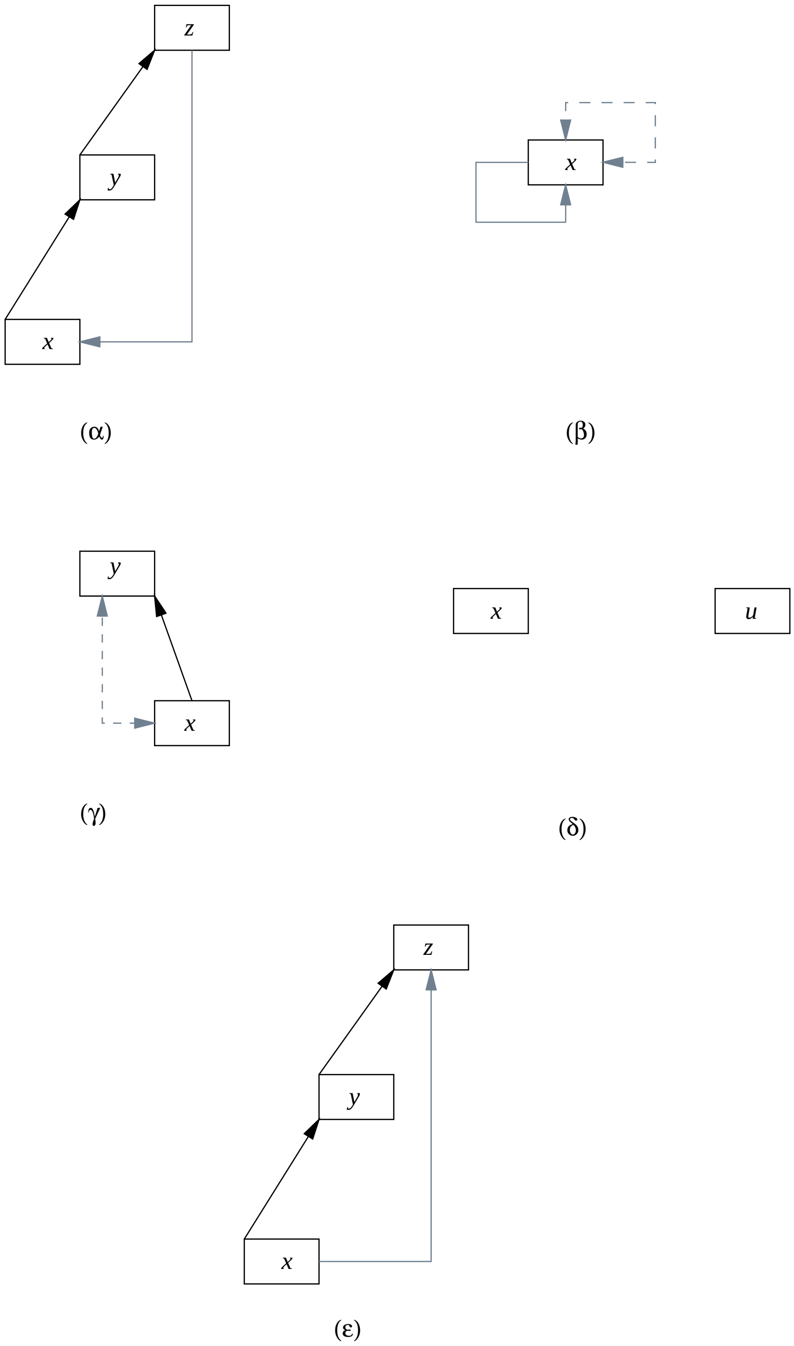
$$G \cap A = \emptyset \tag{3.7}$$

Αν ο θησαυρός θ ικανοποιεί την (3.7) θα γράφουμε $\theta \models \boldsymbol{D}$.

Τέλος δεν επιτρέπεται ένας όρος να ορίζει καμία συσχέτιση με τον εαυτό του. Τυπικά:

$$\{\#x\} \cap x.B = \{\#x\} \cap x.R = \emptyset \ \forall x \in P \tag{3.8}$$

Αν ο θησαυρός θ ικανοποιεί την (3.8) θα γράφουμε $\theta \models \boldsymbol{S}$.



Σχήμα 3.1: Συνθήκες ακεραιότητας θησαυρών.
Οι μαύρες ακμές παριστάνουν συσχετίσεις που ήδη υπάρχουν ενώ οι γκρι ακμές παριστάνουν συσχετίσεις οι οποίες δεν μπορούν να εισαχθούν στο θησαυρό διότι παραβιάζουν κάποια συνθήκη ακεραιότητας. (α) Παραβίαση της συνθήκης (3.1). (β) Παραβίαση της συνθήκης (3.8). (γ) Παραβίαση της συνθήκης (3.7). (δ) Παραβίαση της συνθήκης (3.6). (ε) Παραβίαση της συνθήκης (3.2).

3.3 Πράξεις ενημέρωσης θησαυρών

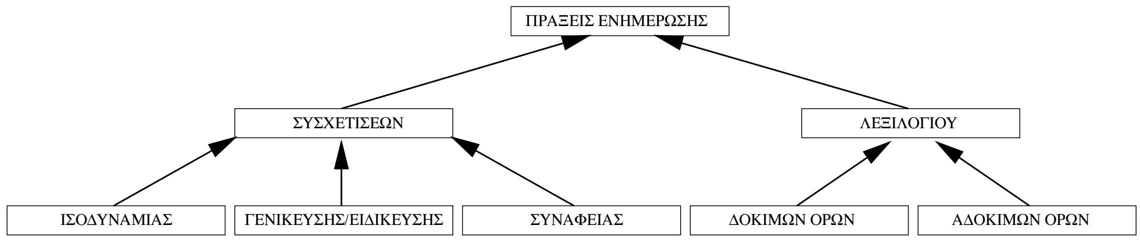
Εστω ότι Θ είναι το σύνολο όλων των θησαυρών $\theta = (T, P, \mathcal{F}, G, E, A)$ και $\mathcal{A}$ είναι ένα σύνολο ορισμάτων, τότε η εξέλιξη ενός θησαυρού επιτυγχάνεται μέσω ενός συνόλου *πράξεων ενημέρωσης*

$$\mathcal{E} = \{\epsilon : \Theta \times 2^{\mathcal{A}} \longrightarrow \Theta\}$$

Μια πράξη ενημέρωσης δέχεται ένα σύνολο ορισμάτων $\alpha \subseteq \mathcal{A}$ και εφαρμοζόμενη σ' ένα θησαυρό θ παράγει ένα νέο θησαυρό $\theta' = \epsilon(\theta, \alpha)$. Μια πράξη ενημέρωσης ϵ θα πρέπει να παράγει ένα θησαυρό ο οποίος είναι συνεπής με τους περιορισμούς ακεραιότητας $\mathcal{C} = \{\boldsymbol{A}, \boldsymbol{E}, \boldsymbol{D}, \boldsymbol{S}, \boldsymbol{M}\}$. θα πρέπει συνεπώς να ισχύει

$$C \in \mathcal{C} \text{ και } \theta \models C \Rightarrow \epsilon(\theta, \alpha) \models C$$

Στην ενότητα αυτή παρουσιάζεται ένα σύνολο τέτοιων πράξεων το οποίο είναι το ελάχιστο σύνολο πράξεων το οποίο μας επιτρέπει να ορίζουμε οποιαδήποτε σύνθετη πράξη ενημέρωσης. Οι πράξεις ενημέρωσης μπορούν να ταξινομηθούν σύμφωνα με την οντότητα στην οποία επιδρούν όπως απεικονίζεται στο σχήμα 3.2.



Σχήμα 3.2: Πράξεις ενημέρωσης θησαυρών

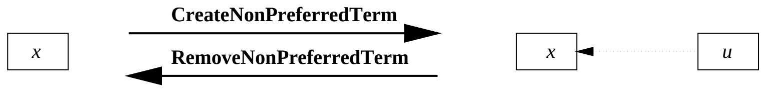
Δημιουργία δόκιμου όρου. Με την πράξη αυτή ένας νέος δόκιμος όρος με κενά σύνολα κατηγοριών, άμεσα γενικότερων και συναφών όρων, δημιουργείται και εισάγεται στον θησαυρό.

Διαγραφή δόκιμου όρου. Ας υποθέσουμε οτι θέλουμε να καταργήσουμε απ' το λεξιλόγιο ενός θησαυρού, ένα δόκιμο όρο. Στην περίπτωση που αυτός έχει ειδικότερους όρους, εισάγεται το πρόβλημα τί θα γίνουν αυτοί. Ανάλογα με την έννοια που αντιπροσωπεύει ο όρος, υπάρχουν διάφορες επιλογές για την κατανομή στην ιεραρχία γενίκευσης/ειδίκευσης των ειδικότερων όρων του. Για το λόγο αυτό ένας όρος μπορεί να καταργηθεί μόνο εφόσον έχει ολοκληρωθεί αυτή η κατανομή και δεν έχει πλέον ειδικότερους όρους. Η κατάργηση του όρου συνεπάγεται και την κατάργηση όλων των συσχετίσεων συνάφειας στις οποίες λαμβάνει μέρος.

Δημιουργία αδόκιμου όρου. Με την πράξη αυτή ένας νέος αδόκιμος όρος δημιουργείται και εισάγεται στον θησαυρό. Επειδή κάθε αδόκιμος όρος δεν είναι αυθύπαρκτο

Σχήμα 3.3: Διαγραφή δόκιμου όρου

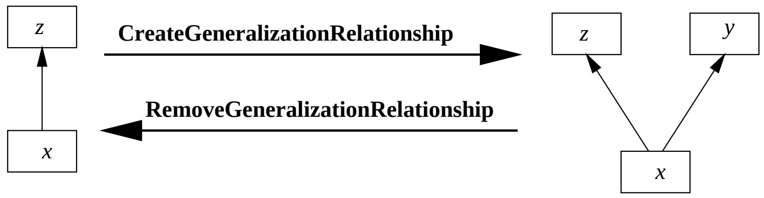
αντικείμενο, αλλά συνδέεται πάντα μ' έναν δόκιμο όρο, η δημιουργία του συνεπάγεται και την δημιουργία μιας συσχέτισης ισοδυναμίας στην οποία λαμβάνει μέρος ο νέος αδόκιμος όρος και κάποιος υπάρχων δόκιμος όρος.



Σχήμα 3.4: Δημιουργία και διαγραφή αδόκιμου όρου

Διαγραφή αδόκιμου όρου. Η κατάργηση ενός αδόκιμου όρου u από το λεξιλόγιο ενός θησαυρού, συνεπάγεται και την κατάργηση (διαγραφή) της συσχέτισης ισοδυναμίας μεταξύ του u και του δόκιμου όρου x με τον οποίο συσχετίζεται.

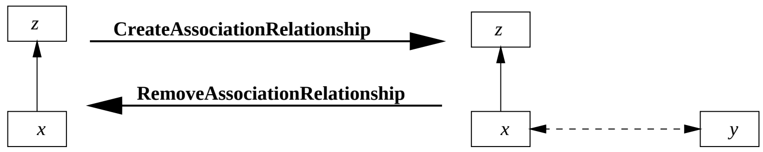
Δημιουργία και διαγραφή συσχέτισης γενίκευσης. Η πράξη αυτή δημιουργεί μια συσχέτιση γενίκευσης από το δόκιμο όρο x προς το δόκιμο όρο y, εισάγοντας το δεύτερο στο σύνολο των άμεσα γενικότερων όρων του πρώτου. Με την αντίστροφη πράξη, διαγράφεται μια συσχέτιση γενίκευσης. Και στις δύο περιπτώσεις είναι αναγκαία η ισχύς των συνθηκών (3.1), (3.2), (3.7) και (3.8).



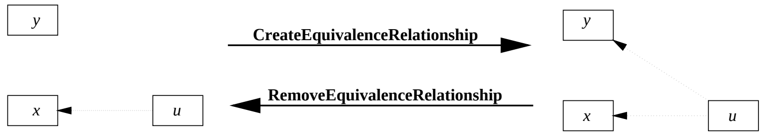
Σχήμα 3.5: Δημιουργία και διαγραφή συσχέτισης γενίκευσης

Δημιουργία και διαγραφή συσχέτισης συνάφειας. Μια συσχέτιση συνάφειας μεταξύ δύο δόκιμων όρων x και y δημιουργείται με την εισαγωγή του y στο σύνολο των συναφών όρων του x και αντίστροφα. Με ακριβώς αντίστροφο τρόπο, λειτουργεί η διαγραφή μιας συσχέτισης συνάφειας. Αναγκαίες συνθήκες και στις δύο περιπτώσεις είναι οι (3.7) και (3.8).

Δημιουργία και διαγραφή συσχέτισης ισοδυναμίας. Μια συσχέτιση ισοδυναμίας μεταξύ του αδόκιμου όρου u και του δόκιμου όρου x, δημιουργείται με την εισαγωγή του x στο σύνολο ισοδύναμων όρων του u. Η αντίστροφη πράξη, διαγράφει μια συσχέτιση ισοδυναμίας. Και στις δύο περιπτώσεις είναι αναγκαία η ισχύς της συνθήκης (3.6).



Σχήμα 3.6: Δημιουργία και διαγραφή συσχέτισης συνάφειας



Σχήμα 3.7: Δημιουργία και διαγραφή συσχέτισης ισοδυναμίας

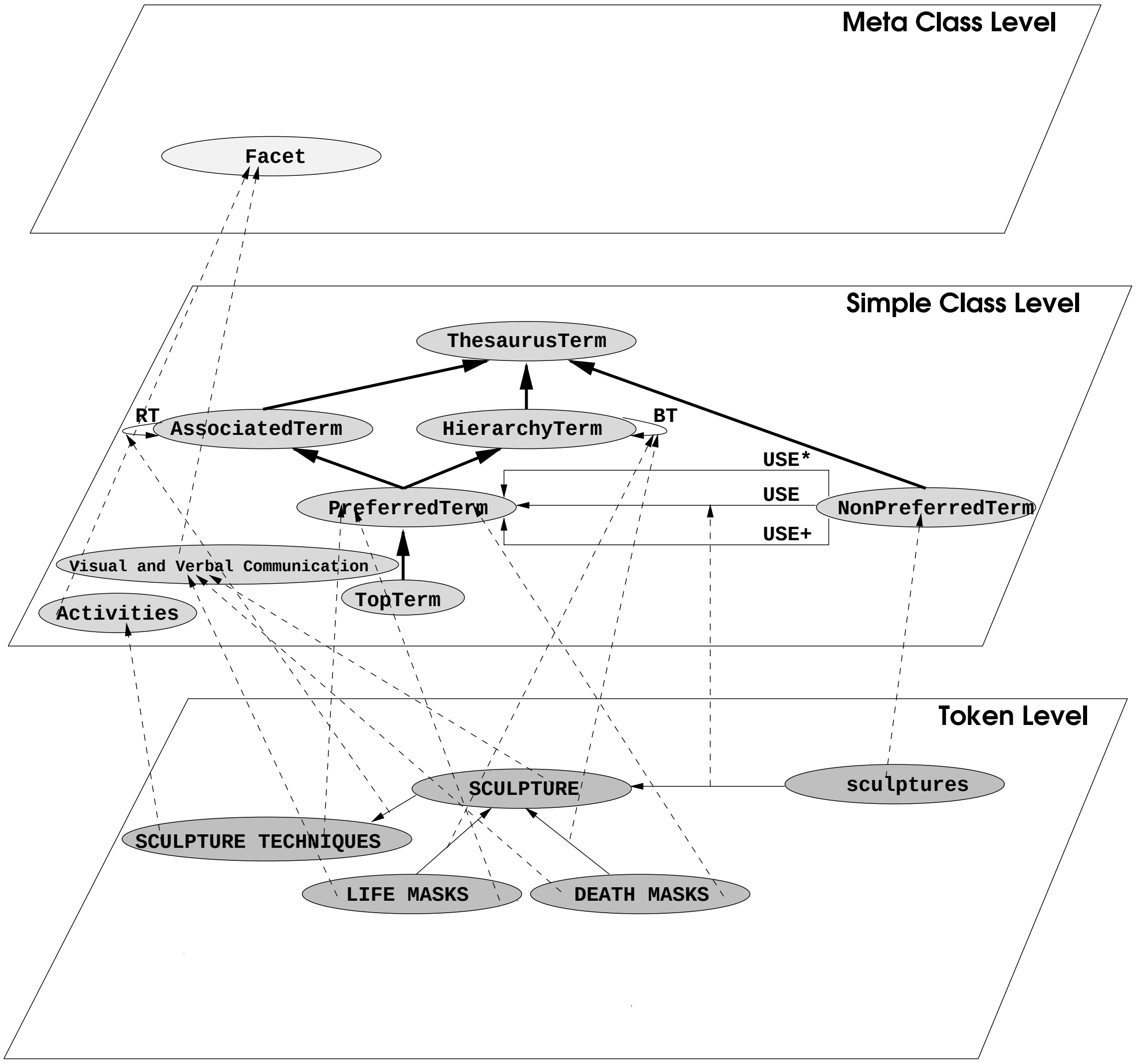
3.4 Παράσταση θησαυρών στην SIS/Telos

Στο σχήμα 3.8 παρουσιάζεται το μοντέλο για την παράσταση θησαυρών στην γλώσσα SIS/Telos μια σύντομη παρουσίαση της οποίας, δίνεται στο Παράρτημα I.

Η μετακλάση `Facet` είναι το σύνολο όλων των κατηγοριών όρων ενός θησαυρού. Κάθε κατηγορία όρων είναι μια κλάση η οποία είναι περίπτωση της `Facet`.

Η κλάση `ThesaurusTerm` είναι το σύνολο των όρων του θησαυρού (το σύνολο T στο συνολοθεωρητικό μοντέλο). Η `ThesaurusTerm` διαμερίζεται στις κλάσεις `AssociatedTerm`, `HierarchyTerm` και `NonPreferredTerm`. Η πρώτη ορίζει την κατηγορία γνωρισμάτων `RT` που παριστάνει συσχετίσεις συνάφειας και η δεύτερη την κατηγορία γνωρισμάτων `BT` που παριστάνει συσχετίσεις γενίκευσης. Η τρίτη ορίζει τις κατηγορίες γνωρισμάτων `USE`, `USE+` και `USE*` οι οποίες αντιπροσωπεύουν τους τρείς τύπους συσχετίσεων ισοδυναμίας που παρουσιάστηκαν στο Κεφάλαιο 2 και παίρνουν τιμές στην κλάση `PreferredTerm`. Η τελευταία, ως υποκλάση των κλάσεων `AssociatedTerm` και `HierarchyTerm`, κληρονομεί τα γνωρίσματα `RT` και `BT`. Η `TopTerm` είναι μια υποκλάση της `PreferredTerm` και αποτελεί το σύνολο των δόκιμων όρων που δεν έχουν γενικότερους όρους.

Στο επίπεδο των ατομικών αντικειμένων (Token Level) δίνονται πέντε όροι προκειμένου να γίνει πιο κατανοητή η παράσταση θησαυρών σ' αυτό το μοντέλο.



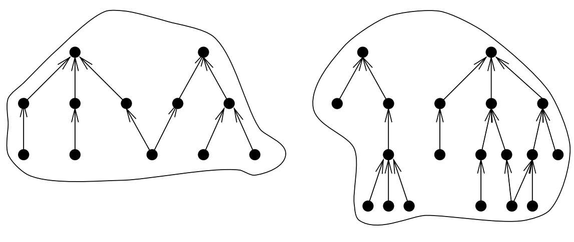
Σχήμα 3.8: Η οντολογία του μοντέλου παράστασης θησαυρών στην γλώσσα Telos
Οι διακεκομμένες γραμμές υποδεικνύουν ταξινόμηση σε μια κλάση, οι έντονες συμπαγείς γραμμές, γενίκευση και οι λεπτές συμπαγείς γραμμές γνωρισματοδότηση.

ΚΕΦΑΛΑΙΟ 4

ΣΥΓΧΩΝΕΥΣΗ ΘΗΣΑΥΡΩΝ

4.1 Επισκόπηση της μεθόδου συγχώνευσης

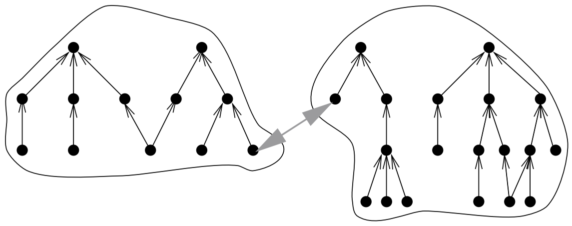
Στην μέθοδό μας η διαδικασία συγχώνευσης είναι δυαδική. Μπορεί να εφαρμοστεί για να συγχωνεύσει δύο θησαυρούς κάθε φορά. Μια δυαδική μέθοδος συγχώνευσης έχει το πλεονέκτημα της απλότητας, αλλά στην περίπτωση που περισσότεροι από δύο θησαυροί πρόκειται να συγχωνευθούν, θα πρέπει να εφαρμοστεί περισσότερες από μία φορές, πράγμα που σημαίνει οτι ένα μέρος της δουλειάς θα επαναληφθεί. Αντίθετα αυτό δεν συμβαίνει όταν χρησιμοποιούνται N-αδικές μέθοδοι, πράγμα που μεταφράζεται σε μικρότερη υπολογιστική πολυπλοκότητα αλλά και υψηλότερες απαιτήσεις σε υπολογιστικούς πόρους και συγκεκριμένα μνήμη. Στην εργασία αυτή προτιμήσαμε μια απλή και σχετικά συμφέρουσα υπολογιστικά, υλοποίηση καταλήγοντας σε μια δυαδική μέθοδο συγχώνευσης.



Σχήμα 4.1: Δύο ξένοι μεταξύ τους θησαυροί

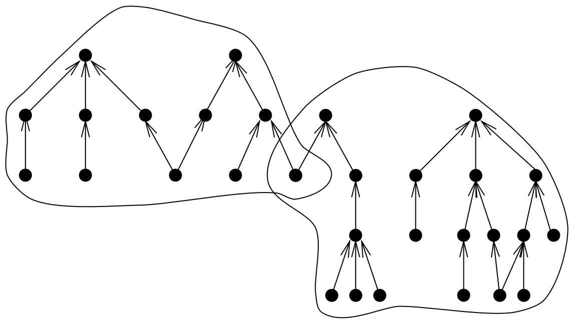
Η διαδικασία της συγχώνευσης, ξεκινά μ' ένα σύνολο ξένων μεταξύ τους ιεραρχιών οι οποίες προέρχονται απ' τους δύο συγχωνευόμενους θησαυρούς όπως παρουσιάζεται στο σχήμα 4.1. Η ιδέα στην οποία βασίζεται η μέθοδός μας είναι η αναγνώριση όρων οι οποίοι πιθανώς παριστάνουν μια κοινή έννοια στους συγχωνευόμενους θησαυρούς.

Τέτοιοι όροι —οι οποίοι θα ονομάζονται στο εξής *όμοιοι*— χρησιμοποιούνται ως κόμβοι άρθρωσης των ιεραρχιών στις οποίες ανήκουν, όπως φαίνεται στο σχήμα 4.2.



Σχήμα 4.2: Ομοιοι όροι χρησιμοποιούνται για την άρθρωση των ιεραρχιών

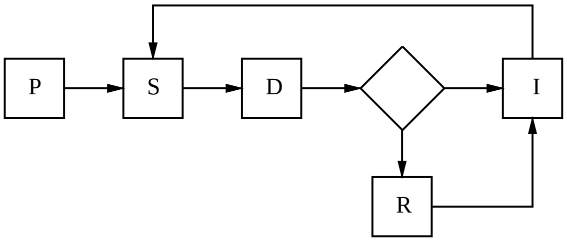
Η ενοποίηση δύο όρων έχει ως αποτέλεσμα την κατάσταση που απεικονίζεται στο σχήμα 4.3. Ενα θέμα που προκύπτει εδώ, σχετίζεται με την εκτέλεση των φάσεων συγχώνευσης: θα πρέπει οι φάσεις να εκτελούνται ακολουθιακά με την λογική σειρά με την οποία περιγράφηκαν στο Κεφάλαιο 2 ή μήπως μια διαφορετική προσέγγιση απαιτείται; Αν η σειρά με την οποία εκτελούνται οι ενοποιήσεις όρων δεν είναι σημαντική όπως π.χ., στην μέθοδο των Mili και Rada, τότε οι φάσεις συγχώνευσης μπορούν να εκτελεστούν ακολουθιακά.



Σχήμα 4.3: Αρθρωση των ιεραρχιών του σχήματος 4.2

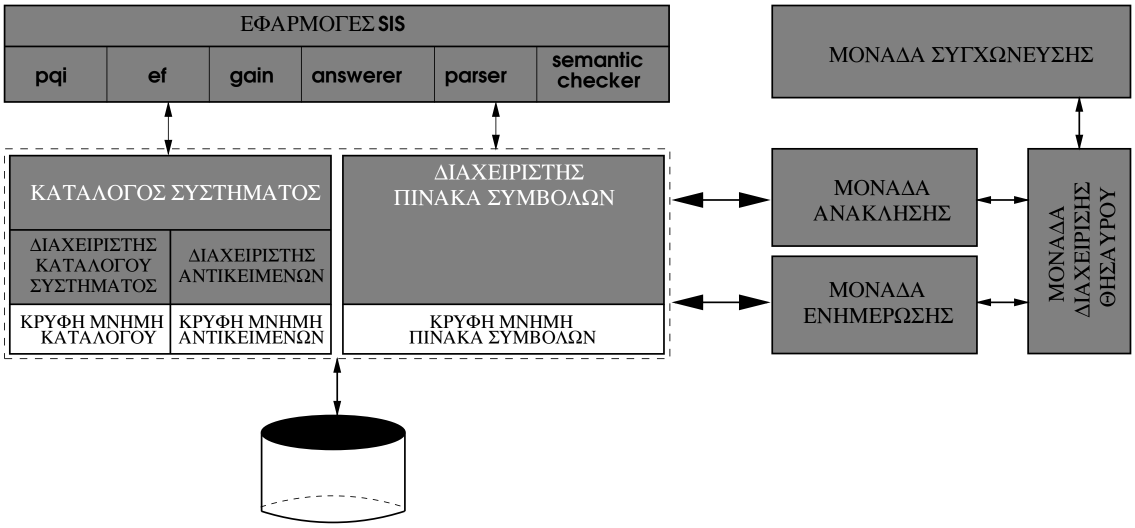
Είναι φανερό απ' τα σχήματα 4.1 και 4.3 οτι κάθε ενοποίηση όρων τροποποιεί την δομή των συγχωνευόμενων θησαυρών. Στην μέθοδό μας ωστόσο, χρησιμοποιούμε αυτή την δομή για να εντοπίσουμε όμοιους όρους. Ετσι η σειρά των ενοποιήσεων παίζει σημαντικό ρόλο, στην μέθοδό μας και οι ενοποιήσεις όρων εκτελούνται με βάση την τοπολογική διάταξη που εισάγει η ιεραρχία γενίκευσης, ξεκινώντας απ' τους γενικότερους όρους και προχωρώντας προς τους ειδικότερους. Κατά συνέπεια οι φάσεις της συγχώνευσης δεν εκτελούνται ακολουθιακά αλλά όπως παρουσιάζεται στο διάγραμμα ροής του σχήματος 4.4.

Κάθε ενοποίηση όρων φέρνει κατά μια έννοια, τις ιεραρχίες των συγχωνευόμενων θησαυρών "κοντύτερα" και αυτό μπορεί να έχει ως συνέπεια τον εντοπισμό όμοιων όρων που ενδεχόμενα δεν θα είχαν εντοπιστεί αν η ενοποίηση δεν είχε λάβει χώρα. Αυτό σε συνδυασμό με την από πάνω προς τα κάτω πολιτική συγχώνευσης οδηγεί στην υπόθεση



Σχήμα 4.4: Ροή εκτέλεσης της συγχώνευσης θησαυρών
P: προενοποίηση, S: εντοπισμός όμοιων όρων, D: Εντοπισμός συγκρούσεων, R: Επίλυση συγκρούσεων, I: Ενοποίηση.

οτι οι εκτιμήσεις ομοιότητας όρων θα πρέπει να βελτιώνονται καθώς η συγχώνευση προχωρά προς τα χαμηλότερα επίπεδα της ιεραρχίας και υποστήριξη από ανθρώπους θα χρειάζεται κυρίως στα αρχικά στάδια της διαδικασίας. Για το λόγο αυτό η υλοποίησή μας επιτρέπει δεσμικό, διαλογικό και ένα υβριδικό (αρχικά διαλογικό και στην συνέχεια δεσμικό) τρόπο λειτουργίας. Οπως αναφέραμε και στο Κεφάλαιο 2 θεωρούμε μια τέτοια ιδιότητα σημαντική διότι για πολύ μεγάλους θησαυρούς είναι πιθανό ο διαλογικός τρόπος λειτουργίας να είναι κουραστικός.



Σχήμα 4.5: Η αρχιτεκτονική της υλοποίησης της μεθόδου συγχώνευσης
Η μονάδα συγχώνευσης είναι ανεξάρτητη απ' το SIS. Η μονάδα διαχείρισης θησαυρού χρησιμοποιεί τις μεθόδους που παρέχονται απ' τις μονάδες ανάκλησης και ενημέρωσης και παρέχει μεθόδους ανεξάρτητες από το SIS στην μονάδα διαχείρισης θησαυρού.

4.2 Προενοποίηση

Κατά την διάρκεια της προενοποίησης σχηματίζουμε ένα θησαυρό

$$\theta = (T, P, \mathcal{F}, G, E, A)$$

απ' τους δύο συγχωνευόμενους θησαυρούς

$$\theta_1 = (T_1, P_1, \mathcal{F}_1, G_1, E_1, A_1) \text{ και } \theta_2 = (T_2, P_2, \mathcal{F}_2, G_2, E_2, A_2)$$

οι οποίοι ικανοποιούν την συνθήκη $T_1 \cap T_2 = \emptyset$, παίρνοντας την ένωση των αντίστοιχων πεδίων. Για για εξασφαλίσουμε οτι τα σύνολα όρων των θ_1 και θ_2 είναι ξένα μεταξύ τους, εισάγουμε ένα πρόθεμα (συνήθως το όνομα του θησαυρού) σε κάθε όρο π.χ., "AAT.SCULPTURE".

4.3 Εντοπισμός όμοιων όρων

Για τον εντοπισμό όμοιων όρων στηριζόμαστε τόσο στους ίδιους τους όρους εφαρμόζοντας λεκτικά κριτήρια, όσο και στις συσχετίσεις μεταξύ αυτών εφαρμόζοντας εννοιολογικά κριτήρια. Και στις δύο περιπτώσεις προσπαθούμε να ισορροπίσουμε μεταξύ της επιθυμητής απόδοσης και της χαμηλής υπολογιστικής πολυπλοκότητας.

4.3.1 Λεκτική ομοιότητα όρων

Χρησιμοποιούμε δύο μηχανισμούς τους οποίους συνδυάζουμε για να εντοπίσουμε λεκτικά όμοιους όρους. Αφαίρεση καταλήξεων και μια παραλλαγή υπογραφών όρων (term signatures). Η αφαίρεση καταλήξεων χρησιμοποιείται για να επιτρέψει φθηνή προσεγγιστική σύγκριση όρων ενώ οι υπογραφές όρων για την εξάλειψη του προβλήματος της σύνταξης όρων που περιγράφηκε στο Κεφάλαιο 2.

Αφαίρεση καταλήξεων

Ο αλγόριθμος που χρησιμοποιούμε οφείλεται στον Porter [Por80] και συνδυάζει ταχύτητα, αποτελεσματικότητα και απλότητα. Δεδομένου οτι αφαιρούμε μόνο καταλήξεις πληθυντικού αριθμού και ενεργητικής/παθητικής φωνής, έχουμε υλοποιήσει ένα μέρος του αλγορίθμου. Πριν παρουσιάσουμε τον αλγόριθμο, δίνουμε μερικούς απαραίτητους συμβολισμούς.

Θα συμβολίζουμε με c και v ένα σύμφωνο και ένα φωνήεν αντίστοιχα. Με l θα συμβολίζουμε ένα σύμφωνο ή φωνήεν. Συνεπώς μια λέξη μπορεί να πάρει την μορφή

$$C^*(VC)^m V^*$$

όπου,

$$\begin{aligned} C &= c^+ \\ V &= v^+ \end{aligned}$$

Το $m \in \boldsymbol{N}$, θα ονομάζεται το *μέτρο* της λέξης και χρησιμοποιείται για ν' αποτρέψει την αφαίρεση καταλήξεων που αφήνουν ένα πολύ μικρό θέμα όπως για παράδειγμα η αφαίρεση της κατάληξης "eed" απ' την λέξη "need". Σύμφωνα με τα παραπάνω, $^*l$ είναι μια οποιαδήποτε λέξη, *SS είναι μια λέξη που καταλήγει σε SS, $^*v^*$ είναι μια λέξη που περιέχει φωνήεν, ενώ $^*c$ και $^*v$ είναι λέξεις που καταλήγουν σε σύμφωνο και φωνήεν αντίστοιχα. Τέλος με e θα συμβολίζουμε την κενή λέξη.

Ο αλγόριθμός 4.1 είναι μια σειρά κανόνων της μορφής:

$$[\mathit{SuffixCondition}][,\mathit{PrefixCondition}] \longrightarrow \mathit{Suffix}$$

οι οποίοι για μια δεδομένη λέξη W ερμηνεύονται ως εξής: "Αν η κατάληξη της W ικανοποιεί την συνθήκη *SuffixCondition* και το πρόθεμα της W ικανοποιεί την συνθήκη *PrefixCondition*, τότε αφαίρεσε την κατάληξη της W και πρόσθεσε την κατάληξη *Suffix*".

Στην περίπτωση που περισσότεροι από ένας κανόνες μπορούν ν' εφαρμοστούν, τότε επιλέγεται εκείνος ο οποίος παρέχει το μεγαλύτερο ταίριασμα. Για παράδειγμα, αν η λέξη μας είναι "caresses", ο αλγόριθμος 4.1 θα επιλέξει να εφαρμόσει τον κανόνα A1 αντί του κανόνα A4.

ΑΛΓΟΡΙΘΜΟΣ 4.1 (stem)

A1	*SSES $\longrightarrow$ SS	(CARESSES $\longrightarrow$ CARESS)
A2	*IES $\longrightarrow$ I	(PONIES $\longrightarrow$ PONI)
A3	*SS $\longrightarrow$ SS	(USELESS $\longrightarrow$ USELESS)
A4	*S $\longrightarrow e$	(COMPUTERS $\longrightarrow$ COMPUTER)
B1	*EED, $(m > 0) \longrightarrow$ EE	(AGREED $\longrightarrow$ AGREE)
B2	*ED, $(*v*) \longrightarrow e$	(CONFLATED $\longrightarrow$ CONFLAT)
B3	*ING, $(*v*) \longrightarrow e$	(BALANCING $\longrightarrow$ BALANC)
C1	*AT $\longrightarrow$ ATE	(CONFLAT $\longrightarrow$ CONFLATE)
C2	*BL $\longrightarrow$ BLE	(ENABL $\longrightarrow$ ENABLE)
C3	*IZ $\longrightarrow$ IZE	(COMPUTERIZ $\longrightarrow$ COMPUTERIZE)
C4	$(*cc$ and not $(*\mathrm{L}$ or $*\,\mathrm{S}$ or $*\,\mathrm{Z})) \longrightarrow$ c	(SHOPP $\longrightarrow$ SHOP)
C5	$(m = 1)$ and $*\,cvc$ and not $(*\mathrm{W}$ or $*\,\mathrm{X}$ or $*\,\mathrm{Y})) \longrightarrow$ E	(FIL $\longrightarrow$ FILE)

Υπογραφές όρων

Ενας όρος t μπορεί να θεωρηθεί ως ένα διατεταγμένο σύνολο λέξεων w που είναι παραθέσεις συμβόλων μιας αλφαβήτου Σ. Ετσι κάθε λέξη $w \in t$ είναι μια παράθεση

$w = \sigma_1, \sigma_2, \cdots, \sigma_{\mathrm{length}(w)}$. Η υπογραφή ενός όρου t, signature(t) δημιουργείται απ' τον αλγόριθμο 4.2.

ΑΛΓΟΡΙΘΜΟΣ 4.2 (signature)

input

t : Ενας όρος που παριστάνεται ως ένα σύνολο λέξεων

$\mathcal{S}$: Ενα σύνολο τετριμμένων λέξεων (αφαιρούνται από κάθε όρο)

output

s : Η υπογραφή του t

begin

$V \leftarrow \emptyset$

foreach λέξη $w \in t$ **do**

if $w \notin \mathcal{S}$ **then** $V \leftarrow V \cup \mathrm{stem}(w)$

end

sort(V)

foreach ρίζα $w \in V$ **do**

foreach $j \in [1, \mathrm{length}(w)]$ **do** $s \leftarrow s w_j$ // παράθεση

end

return s

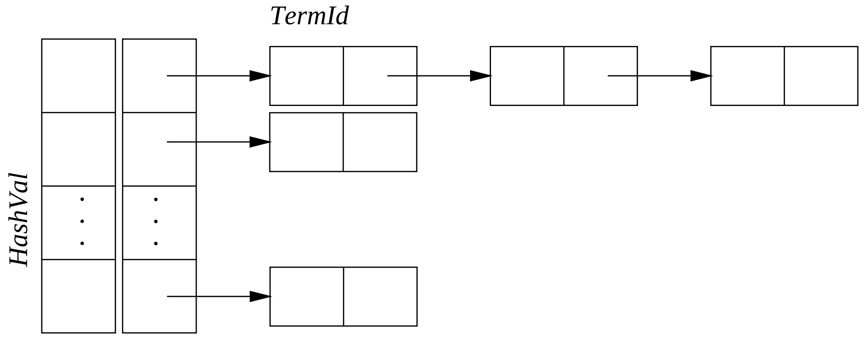
end

Σύμφωνα με τον αλγόριθμο 4.2, μια υπογραφή για τον όρο t, δημιουργείται με την αφαίρεση τετριμμένων λέξεων απ' αυτόν, την αφαίρεση καταλήξεων απ' τις υπόλοιπες λέξεις, την αλφαβητική ταξινόμηση των ριζών που προκύπτουν και τέλος, την παράθεση των χαρακτήρων κάθε λέξης στην ταξινομημένη ακολουθία ριζών. Ετσι εξαλείφεται το πρόβλημα της σύνταξης των όρων αφού αν δύο όροι περιέχουν τις ίδιες λέξεις, τότε θα έχουν την ίδια υπογραφή ανεξάρτητα απ' την σειρά των λέξεων σε καθένα απ' αυτούς.

Κατασκευάζουμε ένα ευρετήριο όρων για έναν απ' τους δύο θησαυρούς. Κατά την φάση εντοπισμού όμοιων όρων, μπορούμε γρήγορα να εντοπίζουμε τους όμοιους όρους αναζητώντας τους στο ευρετήριο όρων. Για εξοικονόμηση μνήμης αντί ν' αποθηκεύουμε τις υπογραφές των όρων στο ευρετήριο, υπολογίζουμε μια τιμή κατακερματισμού $h(s)$ για κάθε υπογραφή s με βάση τον τύπο

$$h(s) = \sum_{i=1}^{\mathrm{length}(s)} i \cdot \mathrm{ord}(s_i) \tag{4.1}$$

και αποθηκεύουμε για την τιμή αυτή, το σύνολο των όρων που η υπογραφή τους έχει την συγκεκριμένη τιμή κατακερματισμού. Ετσι, η δομή του ευρετηρίου είναι όπως παρουσιάζεται στο σχήμα 4.6. Ενα τέτοιο ευρετήριο υλοποιείται εύκολα ως πίνακας κατακερματισμού με ξεχωριστή αλυσίδωση (separete chaining)[1].



Σχήμα 4.6: Η δομή του ευρετηρίου όρων
Για κάθε όρο t αποθηκεύεται στον πίνακα *HashVal* η τιμή κατακερματισμού, $k = h(\mathrm{signature}(t))$ και το σύνολο $\{\#z : h(\mathrm{signature}(z)) = k\}$.

Είναι φανερό οτι η εξοικονόμηση μνήμης που επιτυγχάνεται με την χρήση τιμών κατακερματισμού για τις υπογραφές όρων, αντισταθμίζεται απ' την επιπλέον επεξεργασία που απαιτείται για τον υπολογισμό και την ανάκλησή τους, αφού τώρα ενδέχεται δύο όροι με διαφορετικές υπογραφές να έχουν κοινή τιμή κατακερματισμού. Επομένως θα πρέπει κατά τον εντοπισμό όμοιων όρων, αφού ανακληθούν οι όροι οι οποίοι έχουν την ίδια τιμή κατακερματισμού υπογραφής, να ελεγχθεί ποιοί απ' τους έχουν πράγματι την ίδια υπογραφή. Ετσι δύο λεκτικά όμοιοι όροι εντοπίζονται απ' τον αλγόριθμο 4.3.

ΑΛΓΟΡΙΘΜΟΣ 4.3 (search)

input
  x : Δόκιμος όρος
  H : Ευρετήριο υπογραφών
output
  W: $\{\#y : \mathrm{signature}(y) = \mathrm{signature}(x)\}$
var
  V : Σύνολο αναγνωριστών όρων
begin
  $W \leftarrow \emptyset$
  $s \leftarrow \mathrm{signature}(x)$
  $V \leftarrow \mathrm{lookup}(x, H)$
  foreach $\#y \in V$ **do**
    load($\#y$) // Φόρτωσε τον y απ' την βάση

[1] Βλπ. Παράρτημα II

    if signature(y) $= s$ **then** $W \leftarrow W \cup \{\#y\}$
  end
  return W
end

4.3.2 Συσχετίσεις ισοδυναμίας

Εκτός απ' την εφαρμογή λεκτικών κριτηρίων ομοιότητας, αξιοποιούμε τις συσχετίσεις ισοδυναμίας που περιέχουν οι συγχωνευόμενοι θησαυροί. Συγκεκριμένα όταν θέλουμε να εντοπίσουμε δόκιμους όρους οι οποίοι είναι όμοιοι με ένα δεδομένο όρο x, προσπαθούμε να εντοπίσουμε δόκιμους όρους οι οποίοι είτε είναι λεκτικά όμοιοι με τον x, ή συνδέονται μέσω συσχετίσεων ισοδυναμίας με αδόκιμους όρους οι οποίοι μοιάζουν λεκτικά με τον x[2]. Ο αλγόριθμος 4.4 εντοπίζει όμοιους όρους με βάση λεκτική ομοιότητα και συσχετίσεις ισοδυναμίας.

ΑΛΓΟΡΙΘΜΟΣ 4.4 (similar)
input
  x : Δόκιμος όρος
output
  W: Το σύνολο των όμοιων όρων με τον x (similar(x))
var
  V : Σύνολο αναγνωριστών όρων
begin
  $V \leftarrow \emptyset$; $W \leftarrow \emptyset$
  foreach $u : u \in (T - P) \cup \{x\}$ and $u.U \ni \#x$ **do**
    $V \leftarrow V \cup \text{similar}(u)$
  foreach $\#t \in V$ **do**
    if $t \notin T$ **then**
      load($\#t$)
      foreach $\#y \in t.U$and $t.r = 0$ **do** $W \leftarrow W \cup \{(\#y)\}$
    else
      $W \leftarrow W \cup \{\#t\}$
    end
  end
  return W
end

[2] Είναι φανερό οτι ενδιαφερόμαστε να εντοπίσουμε όμοιους δόκιμους όρους. Αυτό γίνεται επειδή οι δόκιμοι όροι είναι κόμβοι στις ιεραρχίες γενίκευσης και επομένως μπορούν να χρησιμοποιηθούν ως κόμβοι άρθρωσής τους, ενώ οι δόκιμοι όχι.

4.3.3 Εννοιολογική απόσταση όρων

Η προσέγγιση για τον εντοπισμό ομοιοτήτων που έχουμε περιγράψει μέχρι τώρα, έχει μερικά μειονεκτήματα, που έχουν ως συνέπεια να μην αποδίδει ικανοποιητικά σε ορισμένες περιπτώσεις. Συγκεκριμένα:

- Οροι που σημαίνουν διαφορετικές έννοιες μπορούν να έχουν την ίδια υπογραφή, όπως για παράδειγμα οι όροι "Management Information Systems" και "Management of Information Systems", με αποτέλεσμα την μείωση του βαθμού ακρίβειας.
- Οροι που σημαίνουν την ίδια έννοια μπορεί να έχουν διαφορετικές υπογραφές όπως για παράδειγμα οι όροι "Elevations" και "Lifts" με αποτέλεσμα την μείωση του βαθμού ανάκλησης.
- Σε αρκετούς θησαυρούς όπως για παράδειγμα στο θησαυρό CRCS (Computing Rewiews Classification System) της ACM, οι συσχετίσεις ισοδυναμίας είναι σχετικά σπάνιες με αποτέλεσμα μικρή συνεισφορά των συσχετίσεων ισοδυναμίας στον εντοπισμό όμοιων όρων.

Για ν' αντιμετωπίσουμε τις παραπάνω περιπτώσεις, ορίζουμε μια συνάρτηση εννοιολογικής απόστασης μεταξύ δόκιμων όρων βασιζόμενοι στις συσχετίσεις που ορίζουν. Η ιδέα διαισθητικά βασίζεται στην εξής παρατήρηση: αν η απόσταση δύο δόκιμων όρων στην ιεραρχία είναι "αρκετά μεγάλη", οι όροι μάλλον σημαίνουν διαφορετικές έννοιες, ενώ στην περίπτωση που η απόσταση είναι "αρκετά μικρή", μάλλον πρόκειται για όρους που σημαίνουν την ίδια έννοια. Στην υπόθεση αυτή, μπορεί ν' αντιταχθεί το επιχείρημα οτι πολλοί όροι μπορεί να είναι στην ιεραρχία χωρίς να σημαίνουν την ίδια έννοια. Για παράδειγμα οι όροι, "AXONOMETRIC DRAWINGS" και "ORTHOGRAPHIC DRAWINGS" αν και είναι πολύ κοντά (είναι ειδικότεροι όροι του όρου "SCALE DRAWINGS") στον θησαυρό AAT, δεν σημαίνουν την ίδια έννοια. Η αλήθεια είναι πράγματι, οτι η μικρή απόσταση δύο όρων δεν συνεπάγεται κατ' ανάγκη την ταυτοσημία ή την σημασιολογική τους συγγένεια. Ωστόσο όταν μια μικρή απόσταση συνδυάζεται με λεκτική ομοιότητα, τότε αναμφισβήτητα αποτελεί σοβαρότατη ένδειξη σημασιολογικής συγγένειας.

Μοντέλα απόστασης ή ομοιότητας όρων σε θησαυρούς έχουν προταθεί από αρκετούς ερευνητές: τον Paice στο [Pai91], τους Mili και Rada στο [MR88] και [RM89], τον Chen και άλλους στο [CLBD93] και τον Batini και άλλους στο [BL84]. Οι τρεις πρώτες περιπτώσεις βασίζονται στην θεωρία διαδιδόμενης ενεργοποίησης (spreading activation) κόμβων σε σημασιολογικά δίκτυα, η οποία έχει τις ρίζες της στην Ψυχολογία, ενώ η τελευταία είναι πιο ειδική (ad hoc) προς το πεδίο προσέγγιση. Η δική μας προσέγγιση είναι επηρεασμένη κυρίως απ' το μοντέλο αναλογικής ομοιότητας του Σπανουδάκη, [Spa94], [SC96] αλλά και την εργασία του Tversky [Tve77].

Συναρτήσεις απόλυτης απόστασης

ΟΡΙΣΜΟΣ 4.1

Η *απόλυτη απόσταση ταύτισης* δόκιμων όρων δίνεται από την συνάρτηση

$$D_I : P \times P \longrightarrow \{0, 1\}$$

που ορίζεται ως

$$D_I(x, y) = \begin{cases} 0 & \text{αν } \#x = \#y \\ 1 & \text{αν } \#x \neq \#y \end{cases} \tag{4.2}$$

ΟΡΙΣΜΟΣ 4.2

Ο *βαθμός ειδίκευσης* του δόκιμου όρου x, $L(x)$ δίνεται απ' τη συνάρτηση

$$L : P \longrightarrow \boldsymbol{N}$$

η οποία ορίζεται αναδρομικά ως

$$L(x) = \begin{cases} 1 & \text{αν } x.\mathrm{B} = \emptyset \\ 1 + \max\{L(y) : \#y \in x.B\} & \text{διαφορετικά} \end{cases} \tag{4.3}$$

ΟΡΙΣΜΟΣ 4.3

Η *απόλυτη απόσταση γενίκευσης* δόκιμων όρων δίνεται από την συνάρτηση

$$D_G : P \times P \longrightarrow \boldsymbol{R}$$

που ορίζεται ως

$$D_G(x, y) = \sum_{z \in G^+(x) \div G^+(y)} \frac{1}{L(z)} \tag{4.4}$$

Σύμφωνα με τον παραπάνω ορισμό, η απόσταση γενίκευσης δύο όρων εξαρτάται αφενός απ' το σύνολο των διαφορετικών γενικότερων όρων τους και αφετέρου απ' το βαθμό ειδίκευσης αυτών, ο οποίος θεωρείται ως μια ένδειξη της σπουδαιότητάς τους. Συγκεκριμένα θεωρούμε οτι οι γενικοί όροι είναι πιο σημαντικοί διότι συγκεντρώνουν κοινά (αλλά άδηλα) χαρακτηριστικά των ειδικότερων όρων τους, ενώ αντίθετα οι ειδικοί όροι είναι λιγότερο σημαντικοί. Κατά συνέπεια όσο πιο χαμηλά στην ιεραρχία βρίσκεται ένας όρος ο οποίος ανήκει στην διαφορά των συνόλων γενικότερων όρων των όρων x και y τόσο λιγότερο συνεισφέρει στην απόστασή τους.

ΟΡΙΣΜΟΣ 4.4

Η *απόλυτη απόσταση ταξινόμησης* δόκιμων όρων δίνεται από την συνάρτηση

$$D_C : P \times P \longrightarrow \boldsymbol{R}$$

που ορίζεται ως

$$D_C(x, y) = |x.F \div y.F| \tag{4.5}$$

Η απόσταση ταξινόμησης δύο δόκιμων όρων είναι το πλήθος των διαφορετικών κατηγοριών τους.

Συνάθροιση συναρτήσεων απόστασης

Οι συναρτήσεις απόλυτης απόστασης υπολογίζουν την απόσταση δύο δόκιμων όρων βασιζόμενες σε διαφορετικά κριτήρια η κάθε μία. Είναι επομένως υπό αυτήν έννοια, *συναρτήσεις μερικής απόστασης*. Προκειμένου να έχουμε μια *ολική απόσταση* θα πρέπει να τις συναθροίσουμε σε μια ολική συνάρτηση απόστασης όρων.

ΟΡΙΣΜΟΣ 4.5

Η *απόλυτη ολική απόσταση* δύο δόκιμων όρων δίνεται από την συνάρτηση

$$\Delta : P \times P \longrightarrow \boldsymbol{R}$$

που ορίζεται ως

$$\Delta(x, y) = w_I D_I(x, y) + w_G D_C(x, y) + w_C D_G(x, y) \tag{4.6}$$

όπου

$$w_I, w_C, w_G \in [0, +\infty]$$

Οι συντελεστές στάθμισης w_I, w_C και w_G χρησιμοποιούνται για ν' αποδώσουν ένα ειδικό βάρος σε κάθε μία απ' τις αντίστοιχες μερικές συναρτήσεις D_I, D_C, D_G. Αυτό γίνεται για περισσότερη ευελιξία στην ρύθμιση της συμπεριφοράς της Δ σε συγκεκριμένες περιπτώσεις. Για παράδειγμα, δεν έχει κανένα λόγο η χρήση της D_C στην συγχώνευση θησαυρών που δεν υποστηρίζουν κατηγορίες όρων. Επίσης για ένα θησαυρό μια μερική συνάρτηση μπορεί να θεωρηθεί περισσότερο σημαντική απ' τις υπόλοιπες.

Κανονικοποίηση συναρτήσεων απόστασης

Μια επιθυμητή ιδιότητα μιας συνάρτηση απόστασης είναι να λαμβάνει τιμές σ' ένα κλειστό διάστημα τιμών, διότι έτσι μπορούμε πιο εύκολα να χαρακτηρίσουμε με ασαφή (fuzzy) τρόπο μια συγκεκριμένη τιμή της ως "μικρή", "μεγάλη", "πολύ μεγάλη" κλπ. Για το λόγο αυτό μετασχηματίζουμε την Δ έτσι ώστε να λαμβάνει τιμές στο διάστημα $[0, 1]$.

ΟΡΙΣΜΟΣ 4.6

Η *κανονικοποιημένη ολική απόσταση* δύο όρων δίνεται από την συνάρτηση

$$\delta : P \times P \times (0, +\infty] \longrightarrow [0, 1]$$

που ορίζεται ως

$$\delta(x, y, \omega) = 1 - e^{-\omega\Delta(x,y)} \tag{4.7}$$

ΘΕΩΡΗΜΑ 4.1 *Η κανονικοποιημένη απόσταση όρων σ' ένα θησαυρό θείναι μια ψευδομετρική στο σύνολο $\theta.P$.*

Ενα επιθυμητό χαρακτηριστικό μιας κανονικοποιημένης απόστασης είναι να παίρνει την τιμή 0.5 όταν η αντίστοιχη απόλυτη απόσταση βρίσκεται στην μέση τιμή της. Για το λόγο αυτό εισάγεται η παράμετρος ω η οποία χρησιμοποιείται για την εμπειρική ρύθμιση της συμπεριφοράς της δ. Στο σχήμα 4.7 δίνεται η γραφική παράσταση της δ για διάφορες τιμές της παραμέτρου ω.

Χρησιμοποιώντας την δ μπορούμε να αποφασίσουμε αν δύο όροι αποδίδουν την ίδια έννοια, βασιζόμενοι σε εννοιολογικά κριτήρια και όχι μόνο σε λεκτικά. Εχοντας ένα όρο x και αναζητώντας τον πιο κατάλληλο όρο y για ενοποίηση, έχουμε δύο επιλογές:

- Να εντοπίσουμε χρησιμοποιώντας τον αλγόριθμο 4.4 το σύνολο λεκτικά όμοιων όρων similar(x), να επιλέξουμε το ζεύγος

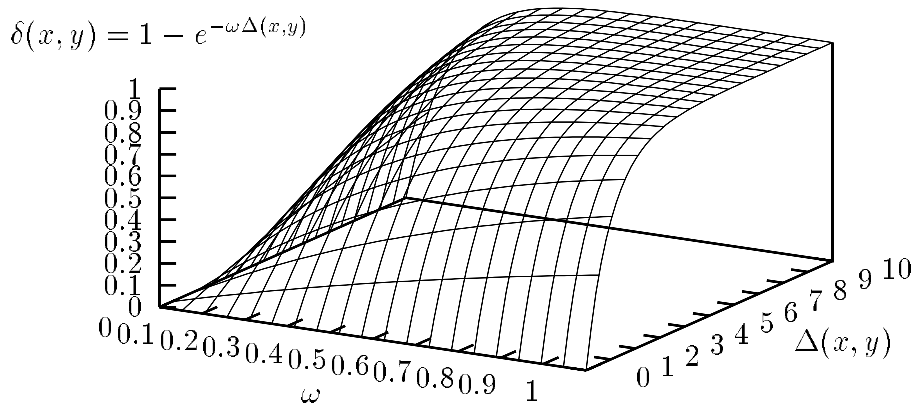
$$(x, y) : \delta(x, y) = \min_{z \,\in\, \mathrm{similar}(x)} \{\delta(x, z)\}$$

  και να προχωρήσουμε σε ενοποίησή του.

- Να παρουσιάσουμε στον επόπτη, όλους τους όρους μαζί με τις αντίστοιχες εννοιολογικές αποστάσεις με τον x και να ζητήσουμε απ' αυτόν την λήψη της σχετικής απόφασης

Στην μέθοδό μας και οι δύο επιλογές είναι διαθέσιμες μέσω του δεσμικού και διαλογικού τρόπου συγχώνευσης αντίστοιχα.

[3]Η απόδειξη του θεωρήματος 4.1, δίνεται στο Παράρτημα ΙΙΙ.



Σχήμα 4.7: Γραφική παράσταση της κανονικοποιημένης απόστασης όρων

Πολυπλοκότητα υπολογισμού της εννοιολογικής απόστασης όρων

Η απόλυτη απόσταση ταύτισης υπολογίζεται σε σταθερό χρόνο. Ο υπολογισμός της απόλυτης απόστασης γενίκευσης δύο όρων απαιτεί τον υπολογισμό της συμμετρικής διαφοράς δύο συνόλων. Αν ϕ_1 και ϕ_2 είναι οι αντίστοιχοι πληθάριθμοι, τότε ο υπολογισμός της D_C απαιτεί χρόνο

$$\chi(D_C) = O(\phi_1 \log \phi_2 + \phi_2 \log \phi_1) \tag{4.8}$$

Είναι σαφές ότι το κρίσιμο σημείο στην εκτίμηση του χρόνου που απαιτεί ο υπολογισμός της D_I είναι το μέγεθος του συνόλου κατηγοριών στις οποίες ανήκει ένας όρος. Το πλήθος των κατηγοριών όρων σ' ένα θησαυρό είναι κατά κανόνα μικρό και συνήθως ένας όρος ανήκει σε μία ή δύο κατηγορίες [Sve89], [ISO86].

Ο υπολογισμός της D_G για δύο όρους x, y, είναι πιο πολύπλοκος. Αρχικά θα πρέπει να υπολογίσουμε τα σύνολα $G^+(x)$ και $G^+(y)$, στην συνέχεια να πάρουμε την συμμετρική τους διαφορά και τέλος να να υπολογίσουμε το βαθμό ειδίκευσης κάθε όρου που ανήκει στην διαφορά.

Εστω οτι ϵ και β είναι το πλήθος των ακμών του γράφου ιεραρχικών συσχετίσεων με ρίζα τον x και το πλήθος των όρων στο σύνολο $G^+(x)$ αντίστοιχα. Ο αλγόριθμος 4.5 ο οποίος υπολογίζει το σύνολο $G^+(x)$ του όρου x, επεξεργάζεται κάθε ακμή ακριβώς μία φορά —γραμμές (6)-(8) και (13)-(15)— εκτελώντας το πολύ μία αναζήτηση και εισαγωγή στο σύνολο V το οποίο δεν περιέχει πάνω από β στοιχεία. Συνεπώς προκύπτει οτι για την εκτέλεση του αλγόριθμου 4.5 απαιτείται χρόνος

$$\chi(G^+) = O(\epsilon \log \beta) \tag{4.9}$$

ΑΛΓΟΡΙΘΜΟΣ 4.5 (G^+)

```
   input    x
   output   G+(x)
   begin
      V ← ∅; load(#x)
(6)   foreach #y ∈ x.B do
(7)      V ← V ∪ {#y}
(8)      append(#y, Q)
      end
      while Q ≠ ∅ do
         #y ← head(Q)
         load(#y)
(13)     foreach #z ∈ y.B and #z ∉ V do
(14)           V ← V ∪ {#z}
(15)           append(#z, Q)
         end
      end
      return V
   end
```

Ο αλγόριθμος 4.6 ο οποίος υπολογίζει το βαθμό ειδίκευσης κάθε όρου που ανήκει σ' ένα σύνολο V, επεξεργάζεται κάθε μία απ' τις ϵ ακμές του γράφου ιεραρχικών συσχετίσεων με ρίζα κάποιο $x \in V$, μία φορά ακριβώς —γραμμές (13)-(16). Αν $\beta = |V|$ τότε ο απαιτούμενος χρόνος για τον τερματισμό του αλγορίθμου είναι

$$\chi(L) = O(\epsilon + \beta) \tag{4.10}$$

Συνοψίζοντας τις εξισώσεις (4.8), (4.9) και (4.10), καταλήγουμε οτι για τον υπολογισμό της απόλυτης απόστασης γενίκευσης δύο όρων x, y με $|G^+(x)| = \beta_1$ και $|G^+(x)| = \beta_2$ όταν οι γράφοι ιεραρχικών συσχετίσεων των x και y έχουν ϵ_1 και ϵ_2 ακμές αντίστοιχα, απαιτείται χρόνος

$$\chi(D_G) = O((\epsilon_1 \log \beta_1 + \epsilon_2 \log \beta_2) + (\beta_1 \log \beta_2 + \beta_2 \log \beta_1) + (\epsilon_1 + \beta_1 + \epsilon_2 + \beta_2)).$$

Δεδομένου ότι

$$\epsilon_1 \geq \beta_1 - 1 \text{ και } \epsilon_2 \geq \beta_2 - 1,$$

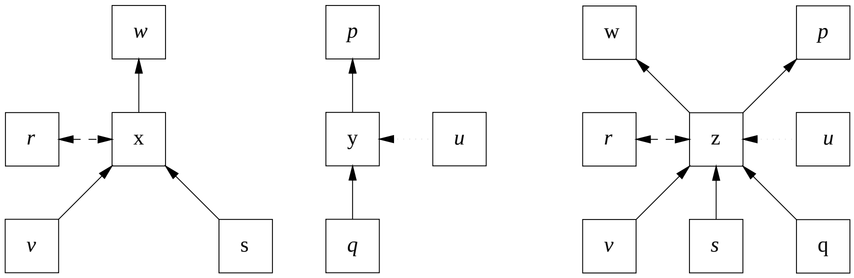
αν κρατήσουμε μόνο τους σημαντικότερους όρους θα έχουμε τελικά

$$\chi(D_G) = O(\epsilon_1 \log \beta_1 + \epsilon_2 \log \beta_2) \tag{4.11}$$

ΑΛΓΟΡΙΘΜΟΣ 4.6 (L)

```
input    V
output   L(x) ∀#x ∈ V
begin
  foreach #x ∈ V do
    load(#x)
    outdegree[x]←|x.B|
    spzdegree[x]←1
    if outdegree[x] = 0 then insert(#x, Q)
  end
  while Q ≠ ∅ do
    #x←head(Q)
    foreach #y : #x ∈ y.B do
      outdegree[y]←outdegree[y] − 1
      spzdegree[y]← max{spzdegree[y], 1 + spzdegree[x]}
      if outdegree[y] = 0 then insert(#y, Q)
    end
  end
end
```

Δεδομένου οτι η συνάθροιση των απόλυτων αποστάσεων και η κανονικοποίηση απαιτούν σταθερό χρόνο υπολογισμού, ο σημαντικότερος παράγοντας στην πολυπλοκότητα υπολογισμού της εννοιολογικής απόστασης εισάγεται απ' την εξίσωση (4.11).



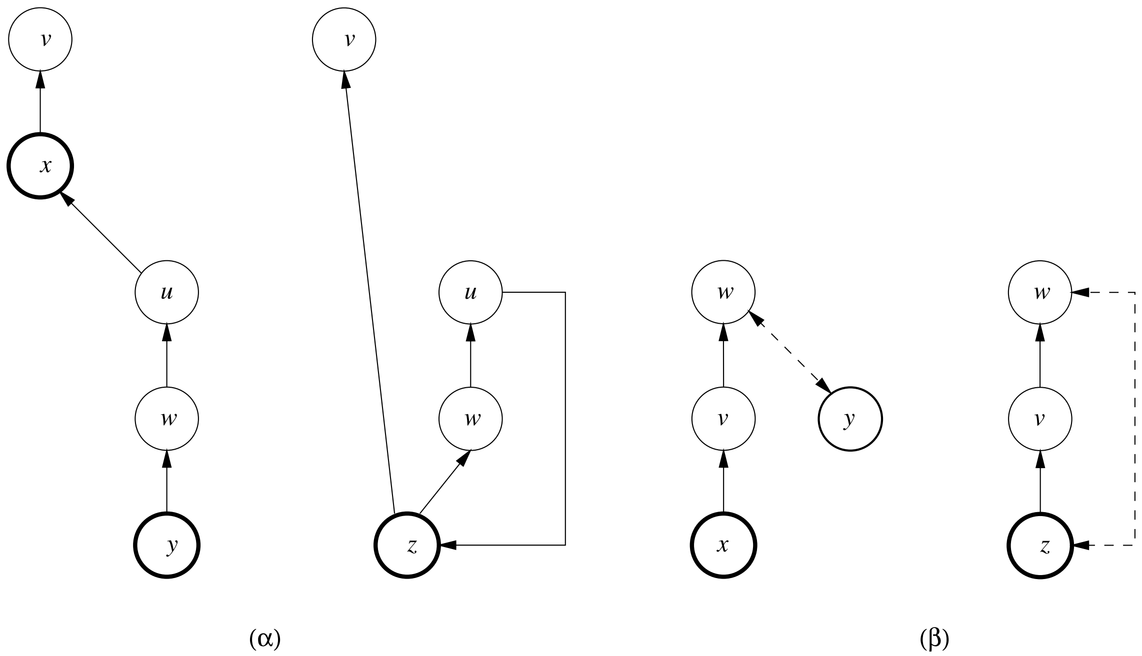
(α) (β)

Σχήμα 4.8: Ενοποίηση όρων

Οι όροι x και y στο σχήμα (α) ενοποιούνται παράγοντας τον όρο z του σχήματος (β).

4.4 Ενοποίηση όρων και εντοπισμός συγκρούσεων

Η ενοποίηση όρων υλοποιείται (αλγόριθμος 4.7) με τρόπο παρόμοιο μ' αυτόν της εργασίας [MR88] όπως φαίνεται και στο σχήμα 4.8. Ο υπολογισμός του συνόλου των άμεσα γενικότερων όρων ($z.B$) γίνεται έτσι ώστε να εξασφαλίζεται η ισχύς της συνθήκης (3.2)[4]. Δεν συμβαίνει ωστόσο το ίδιο και με τις συνθήκες (3.1) και (3.7). Είναι πιθανό οι συσχετίσεις δύο όρων που πρόκειται να ενοποιηθούν να παραβιάζουν τις συνθήκες αυτές σε περίπτωση που εφαρμοστεί ο αλγόριθμος 4.7 όπως παρουσιάζεται στο σχήματα 4.9α και 4.9β.



Σχήμα 4.9: Παραβίαση συνθηκών ακαιρεότητας έπειτα από ενοποιήσεις όρων. Οι έντονοι κύκλοι υποδεικνύουν τους όρους που πρόκειται να ενοποιηθούν. Στ' αριστερά κάθε σχήματος δίνεται η κατάσταση πριν την ενοποίηση, ενώ στα δεξιά η κατάσταση μετά την ενοποίηση. (α) Παραβιάζεται η συνθήκη (3.1). (β) Παραβιάζεται η συνθήκη (3.7).

ΑΛΓΟΡΙΘΜΟΣ 4.7 (integrate)

input

$\#x, \#y$: Οι αναγνωριστές των όρων που ενοποιούνται

z : Ο νέος όρος που θα δημιουργηθεί

output

$\#z$: Ο αναγνωριστής του όρου που προκύπτει απ' την ενοποίηση

begin

$\#z \leftarrow$ create(z)

[4] Η απόδειξη δίνεται στο Παράρτημα III.

$z.F \leftarrow x.F \cup y.F$
$z.A \leftarrow x.A \cup y.A$
$z.B \leftarrow (x.B - (G^+(y) - y.B)) \cup (y.B - (G^+(x) - x.B))$
foreach $u : \#v \in u.U$ and $\#v \in \{\#x, \#y\}$ **do** $u.U \leftarrow (u.U - \{\#v\}) \cup \{\#z\}$
return $\#z$
end

Για το λόγο αυτό, πρίν από κάθε ενοποίηση όρων πρέπει να ελέγξουμε αν κάποια απ' τις (3.1), (3.2) παραβιάζεται. Σε μια τέτοια περίπτωση θα πρέπει είτε η συγκεκριμένη ενοποίηση να απορριφθεί ή να επιλυθεί η σύγκρουση και να επιχειρηθεί ξανά η ενοποίηση. Οι Mili και Rada στο άρθρο [MR88] υιοθετούν μια προκαθορισμένη πολιτική επίλυσης συγκρούσεων όπως είδαμε στο Κεφάλαιο 2. Η δική μας προσέγγιση είναι να μην επιχειρούμε επίλυση συγκρούσεων. Ετσι σε περιπτώσεις που κάτι τέτοιο είναι απαραίτητο, η σύγκρουση θα πρέπει να επιλυθεί από τον επόπτη της διαδικασίας. Η επιλογή αυτή απορρέει απ' την πεποίθησή μας πως απ' την στιγμή που οι συσχετίσεις αντανακλούν συγκεκριμένες σχεδιαστικές επιλογές, δεν θα πρέπει να αναδιαμορφώνονται με προκαθορισμένο τρόπο.

Παραβίαση της αντισυμμετρίας των ιεραρχικών συσχετίσεων. Εστω οτι x, y είναι δύο δόκιμοι όροι οι οποίοι πρόκειται να συγχωνευθούν και $L(x) < L(y)$. Η συνθήκη (3.1) παραβιάζεται αν και μόνο αν υπάρχει ιεραρχικό μονοπάτι $y \rightsquigarrow x$. Αν ένα τέτοιο μονοπάτι υπάρχει,το μήκος του προφανώς δεν μπορεί να υπερβαίνει την τιμή $k = L(y) - L(x)$, διότι τότε ο βαθμός ειδίκευσης του y θα ήταν μεγαλύτερος από $L(y)$. Αν με $G^k(y)$ συμβολίσουμε το σύνολο των όρων v για τους οποίους υπάρχει ένα ιεραρχικό μονοπάτι $y \rightsquigarrow v$ μήκους το πολύ k, τότε η ενοποίηση των x και y παραβιάζει την (3.1) αν και μόνο αν $x \in G^k(y)$. Ο έλεγχος παραβίασης της (3.1) είναι πιο οικονομικός αν χρησιμοποιήσουμε το σύνολο G^k αντί του G^+, αφού $G^k(x) \subseteq G^+(x)$ για κάθε x, απ' την στιγμή που

$$G^+(x) = \bigcup_{k=1}^{\infty} G^k(x)$$

ΑΛΓΟΡΙΘΜΟΣ 4.8 (G^k)
input x, k
output $G^k(x)$
begin
$V \leftarrow \emptyset$
foreach $\#y \in x.B$ **do**
$V \leftarrow V \cup \{\#y\}$
append$((\#y, 1), Q)$

```
  end
  while Q ≠ ∅ do
    (#y, m) ← head(Q)
    if m < k then
      load(#y)
      foreach #z ∈ y.B and #z ∉ V do
        V ← V ∪ {#z}
        append((#z, m + 1), Q)
      end
    end
  end
  return V
end
```

Παραβίαση της διαζευξιμότητας των σχέσεων. Αν οι δόκιμοι όροι x, y πρόκειται να συγχωνευθούν, η συνθήκη (3.7) παραβιάζεται αν και μόνο αν ισχύει τουλάχιστον μία απ' τις σχέσεις:

$$G^+(x) \cap y.A \neq \emptyset \tag{4.12}$$
$$G^+(y) \cap x.A \neq \emptyset \tag{4.13}$$
$$S^+(x) \cap y.A \neq \emptyset \tag{4.14}$$
$$S^+(y) \cap x.A \neq \emptyset \tag{4.15}$$

Ο τρόπος ελέγχου των (4.12) και (4.13) είναι προφανής. Ωστόσο ο έλεγχος των (4.12) και (4.15) μπορεί να παρουσιάζει προβλήματα αν τα σύνολα $S^+(x)$ και $S^+(y)$ είναι πολύ μεγάλα, πράγμα που μπορεί να συμβεί αν οι x και y είναι αρκετά υψηλά στην ιεραρχία γενίκευσης. Σε μια τέτοια περίπτωση μπορούμε να εκμεταλλευτούμε το γεγονός οτι αν π.χ., η (4.14) ισχύει, τότε θα υπάρχει ένα μονοπάτι γενίκευσης $w \rightsquigarrow x, w \in y.A$. Μπορούμε επομένως να υπολογίσουμε το $G^{L(w)-L(x)}(w)$ αντί του $S^+(x)$.

4.5 Ο αλγόριθμος συγχώνευσης θησαυρών

Μέχρι τώρα έχουμε περιγράψει πως υλοποιείται κάθε φάση συγχώνευσης ξεχωριστά. Ο αλγόριθμος 4.9 συνοψίζει όλα τα προηγούμενα και δίνει μια ολική περιγραφή της μεθόδου συγχώνευσης.

ΑΛΓΟΡΙΘΜΟΣ 4.9 (ThesauriMerging)

var

  Q_1, Q_2 : Ουρές διάσχισης
  $\mathcal{W}$: Ουρά αναμονής όρων (υλοποιεί την τοπολογική διάσχιση)
  $\mathcal{M}$: Ουρά προταιρεότητας όρων για συγχώνευση
  L : Επίπεδο ιεραρχίας
  M_T : Κατώφλι συγχώνευσης

begin

  foreach $\#x : x \in \theta.T$ and $x.B = \emptyset$ **do** append$(\#x, Q_1)$
  L$\leftarrow$ 1; $Q_2 \leftarrow \emptyset$
  while $Q_1 \neq \emptyset$ **do**
    while $Q_1 \neq \emptyset$ **do**
      $\#x \leftarrow$pop(Q_1)
      if $(\#x, k) \in \mathcal{W}$ **then**
        if $k - 1 > 0$ **then**
          update$((\#x, k), (\#x, k-1), \mathcal{W})$
          continue
        else
          remove$((\#x, k), \mathcal{W})$
        end
      else
        if $|x.B| > 1$ **then** insert$((\#x, |x.B| - 1), \mathcal{W})$
      end
      $V \leftarrow$ similar(x)
      if $V \neq \emptyset$ **then**
        find $y \in V : \delta(x, y) = \min\{\delta(x, i) : \#i \in V\}$
        if $\delta(x, y) < M_T$ **then**
          insert$((\#x, \#y), \max\{L(x), L(y)\}), \mathcal{M})$
      else
        foreach $\#i \in x.B$ **do** append$(\#i, Q_2)$
      end
    end
    loop
      $(\#x, \#y, l) \leftarrow$getmin$(\mathcal{M})$
      while not empty$(\mathcal{M})$ **do**
        $\#z \leftarrow$integrate$(\#x, \#y)$
        διέγραψε όλες τις εμφανίσεις των x, y απ' τις Q_1, Q_2
        remove$((\#x, \#y), Q)$
        foreach $u : u.B \ni \#z$ **do** append$(Q_2, \#u)$

$(\#x, \#y, l) \leftarrow \mathrm{getmin}(\mathcal{M})$
end
$Q_1 \leftarrow Q_2$; $Q_2 \leftarrow \emptyset$
$L \leftarrow L + 1$
until $Q_1 \neq \emptyset$ or $M = \emptyset$
end
end

ΚΕΦΑΛΑΙΟ 5

ΧΡΗΣΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1 Πειραματική χρήση

Προκειμένου ν' αποδειχθεί η αποτελεσματικότητα της μεθόδου συγχώνευσης που προτείνουμε θα πρέπει να γίνουν εκτενή πειράματα συγχώνευσης θησαυρών. Στα πλαίσια αυτής της εργασίας θα περιοριστούμε στην παρουσίαση ενός πειράματος συγχώνευσης μεταξύ του θησαυρού "Computing Reviews Classification System" — CRCS της ACM και ενός τμήματος απ' τον θησαυρό "Library of Congress Subject Headings" — LCSH της βιβλιοθήκης του Κονγκρέσσου των ΗΠΑ το οποίο περιλαμβάνει όρους σχετικούς με την Επιστήμη Υπολογιστών[1]. Ο πρώτος θησαυρός χρησιμοποιείται εκτενέστατα για τον ευρετηριασμό και την αναζήτηση άρθρων που δημοσιεύονται στα περιοδικά της ACM, ενώ ο δεύτερος είναι ο μεγαλύτερος θησαυρός σε χρήση σήμερα περιλαμβάνοντας 500.000 όρους περίπου. Στον πίνακα 5.1 συγκεντρώνονται τα βασικά χαρακτηριστικά των δύο αυτών θησαυρών.

Μέσω της πρώτης αυτής απόπειρας συγχώνευσης με πραγματικούς θησαυρούς στοχεύαμε στα παρακάτω:

1. Να εντοπίσουμε ατέλειες και σφάλματα της μεθόδου τα οποία ενδεχομένως δεν είχαν εντοπιστεί στα πειράματα με μικρούς και τεχνητούς θησαυρούς που έγιναν κατά την διάρκεια της υλοποίησης και του ελέγχου.

2. Να πάρουμε ένα πρώτο δείγμα της αποτελεσματικότητας και της απόδοσης της μεθόδου και πιο συγκεκριμένα:

   - Να διαπιστώσουμε αν και κατά πόσο είναι αποτελεσματικός ο μηχανισμός

[1] Στο εξής θ' αναφέρουμε αυτό το υποσύνολο του LCSH ως LCSH/CS.

εντοπισμού όρων που ενδεχομένως αποδίδουν μια κοινή έννοια, χρησιμοποιώντας λεκτικά κριτήρια.

- Να διαπιστώσουμε αν και κατά πόσο βελτιώνει τον εντοπισμό όμοιων όρων η χρήση συσχετίσεων ισοδυναμίας.
- Να εξακριβώσουμε κατά πόσο μπορεί η εννοιολογική απόσταση δ να εντοπίσει εσφαλμένες ομοιότητες μεταξύ όρων και αν αυτό συμβαίνει κατά πόσο αυτή η ικανότητα βελτιώνεται καθώς η συγχώνευση προχωρά.

ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ	CRCS	LCSH/CS	**ΣΥΝΟΛΑ**
ΟΡΟΙ	992	5274	6266
ΔΟΚΙΜΟΙ ΟΡΟΙ	967	3054	4021
ΑΔΟΚΙΜΟΙ ΟΡΟΙ	25	2220	2245
ΙΕΡΑΡΧΙΚΕΣ ΣΥΣΧΕΤΙΣΕΙΣ	1056	2774	3838
ΣΥΣΧΕΤΙΣΕΙΣ ΙΣΟΔΥΝΑΜΙΑΣ	25	2220	2245
ΣΥΣΧΕΤΙΣΕΙΣ ΣΥΝΑΦΕΙΑΣ	82	171	253
ΒΑΘΟΣ ΙΕΡΑΡΧΙΑΣ	6	11	
ΠΟΛΥΕΡΑΡΧΙΚΟΤΗΤΑ	ΝΑΙ	ΝΑΙ	

Πίνακας 5.1: Χαρακτηριστικά των θησαυρών CRCS και LCSH/CS

5.1.1 Λεκτική ομοιότητα όρων

Προκειμένου ν' αξιολογήσουμε την απόδοση του εντοπισμού όμοιων όρων με λεκτικά κριτήρια εκτελέσαμε τον αλγόριθμο 4.4 για κάθε όρο του θησαυρού CRCS αναζητώντας λεκτικά όμοιους όρους στο ευρετήριο όρων. Στην συνέχεια επαναλάβαμε το ίδιο για τον θησαυρό LCSH/CS. Τ' αντίστοιχα αποτελέσματα δίνονται στον πίνακα 5.2.

	CRCS	LCSH/CS
ΑΝΑΖΗΤΗΣΕΙΣ ΟΡΩΝ	967	3053
ΕΠΙΤΥΧΕΙΣ ΑΝΑΖΗΤΗΣΕΙΣ	130	127
ΠΟΣΟΣΤΟ ΕΠΙΤΥΧΙΑΣ	0.134	0.04
ΜΕΣΟ ΜΕΓΕΘΟΣ ΑΠΑΝΤΗΣΗΣ	1.377	1.35

Πίνακας 5.2: Αποτελέσματα εντοπισμού λεκτικά όμοιων όρων

Παρατηρούμε απ' τον πίνακα 5.2 οτι ο αριθμός επιτυχών αναζητήσεων για όρους του CRCS είναι μεγαλύτερος κατά 3 του αντίστοιχου αριθμού για τον LCSH/CS. Αυτό οφείλεται στο γεγονός οτι ο CRCS περιέχει τους όρους "Text Processing", "User Interfaces" και "Image Processing" δύο φορές τον καθένα. Κατά τα άλλα το σύνολο των απαντήσεων που λάβαμε ήταν το ίδιο σε κάθε περίπτωση. Τα παραπάνω ωστόσο δεν δίνουν καμιά πληροφορία για το πόσο καλά αποδίδει ο εντοπισμός ομοιοτήτων με λεκτικά κριτήρια. Για να έχουμε μια τέτοια πληροφορία θα πρέπει να γνωρίζουμε:

1. για κάθε ζεύγος όμοιων όρων που εντοπίστηκε αν πράγματι αυτό περιλαμβάνει ταυτόσημους όρους,

2. πόσα ζεύγη ταυτόσημων όρων δεν εντοπίστηκαν ενώ θα έπρεπε.

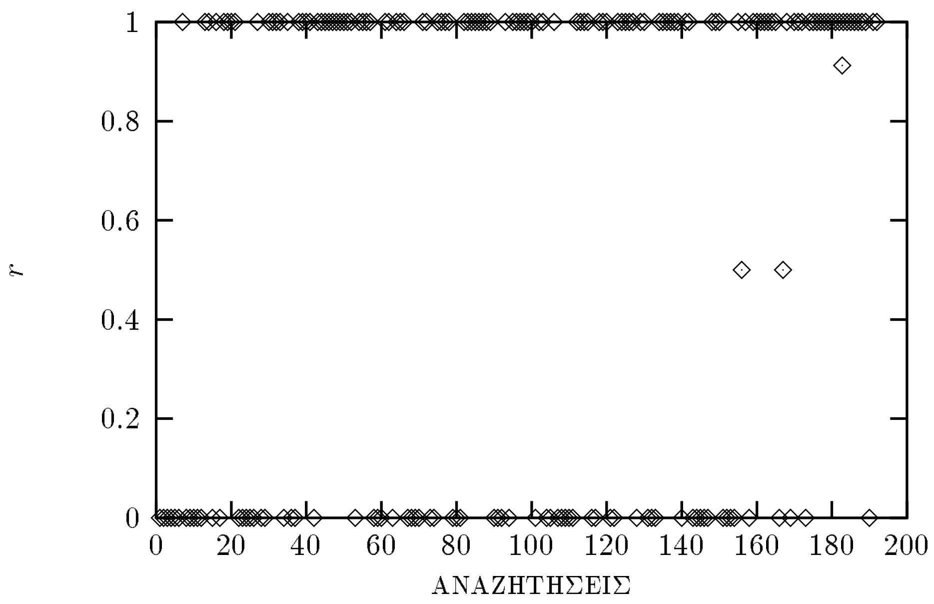
Συγκρίναμε το σύνολο των αναζητήσεων και των αντίστοιχων αποτελεσμάτων με τα περιεχόμενα της βάσης δεδομένων. Υπολογίσαμε το μέσο βαθμό ανάκλησης

$$\bar{r} = \frac{1}{n} \sum \frac{R_R}{R_R + R_N}$$

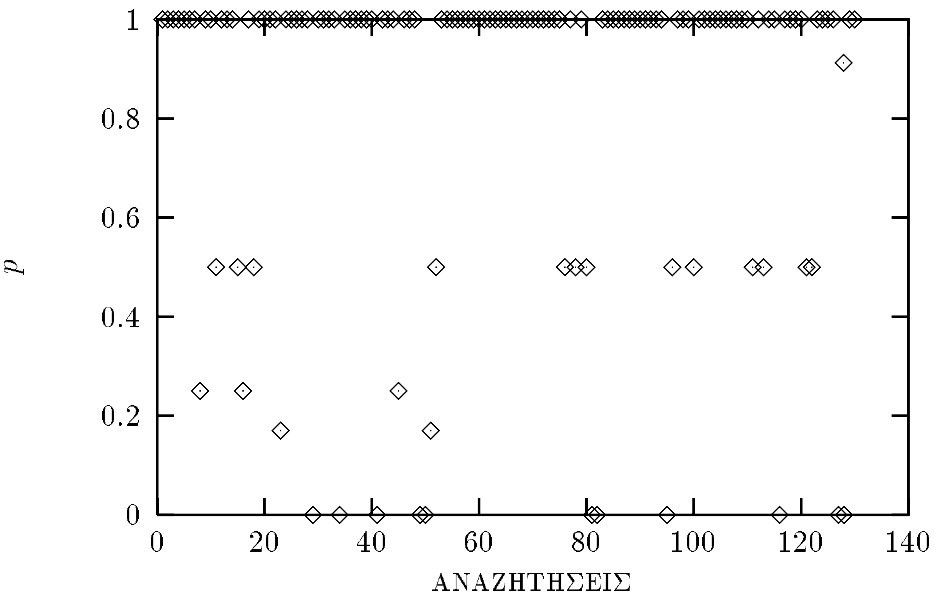
και το μέσο βαθμό ακρίβειας

$$\bar{p} = \frac{1}{n} \sum \frac{R_R}{T_R}$$

όπου n είναι ο αριθμός των αναζητήσεων που έγιναν, R_R είναι ο αριθμός των ζευγών ταυτόσημων όρων, R_N ο αριθμός των ζευγών μη-ταυτόσημων όρων και T_R ο συνολικός αριθμός ζευγών όρων που ανακλήθηκαν σε κάθε ερώτηση, αντίστοιχα. Ο βαθμός ανάκλησης σε 192 αναζητήσεις με $R_R + R_N > 0$ ήταν 61.5%, ενώ ο βαθμός ακρίβειας σε 130 αναζητήσεις με $T_R > 0$ ήταν 83.5%. Στα σχήματα 5.1 και 5.2 δίνονται τα διαγράμματα διασποράς για το βαθμό ανάκλησης και ακρίβειας αντίστοιχα.



Σχήμα 5.1: Διασπορά του βαθμού ανάκλησης στον εντοπισμό όμοιων όρων



Σχήμα 5.2: Διασπορά του βαθμού ακρίβειας στον εντοπισμό όμοιων όρων

Για να αξιολογήσουμε την συμβολή των συσχετίσεων ισοδυναμίας στον εντοπισμό όμοιων όρων, εκτελέσαμε την ίδια διαδικασία, χωρίς να χρησιμοποιούμε τις συσχετίσεις ισοδυναμίας. Τ' αποτελέσματα έδειξαν μια μείωση των επιτυχών αναζητήσεων κατά 23.5%.

Ας συνοψίσουμε τα βασικά συμπεράσματα που μπορούν να εξαχθούν απ' τα παραπάνω αποτελέσματα.

- Ο μέσος βαθμός ακρίβειας του εντοπισμού όμοιων όρων είναι πολύ υψηλός. Είναι ίσως ελαφρά υψηλότερος απ' οτι θ' ανέμενε κανείς γεγονός που οφείλεται κατά την γνώμη μας στο γεγονός οτι και οι δύο θησαυροί καλύπτουν το ίδιο πεδίο γνώσης και κατά συνέπεια είναι δεν υπήρχαν πολλά ομώνυμα. Σε τουλάχιστον δύο περιπτώσεις διαπιστώσαμε αποτυχία του αλγόριθμου αφαίρεσης καταλήξεων. Η πρώτη οφείλεται στον διπλό πληθυντικό του όρου "Thesaurus" ("Thesauri" και "Thesauruses"), ενώ η δεύτερη στην μετατροπή της κατάληξης "ies" σε "i". Ετσι οι όροι "Cache Memories" και "Cache Memory" θεωρήθηκαν διαφορετικοί.

- Ο μέσος βαθμός ανάκλησης του εντοπισμού όμοιων όρων είναι αρκετά καλός. Πιστεύουμε οτι θα ήταν ακόμα καλύτερος αν οι όροι του CRCS χρησιμοποιούσαν περισσότερο τον επιθετικό προσδιορισμό "Computer" π.χ., "Computer Graphics", Πράγμα που δεν γίνεται σαφώς λόγω της περιορισμένης εμβέλειας του συγκεκριμένου θησαυρού.

- Η χρήση των συσχετίσεων ισοδυναμίας κρίνεται απόλυτα επιτυχής παρά το γεγονός οτι επιβαρύνει ελαφρά την διαδικασία Χωρίς την χρήση τους ο μέσος βαθμός ανάκλησης θα είχε κυμανθεί σε επίπεδα κάτω του 50% ενώ η μικρή αύξηση του βαθμού ακρίβειας την οποία θα συνεπαγόταν δεν θα αντιστάθμιζε την κατάσταση αφού μας ενδιαφέρει κυρίως να βρούμε όσο πιο πολλούς πιθανά ταυτόσημους όρους.

- Αν και απέτυχε σε μερικές περιπτώσεις ο αλγόριθμος του Porter [Por80] αποδίδει αρκετά καλά. Θα μπορούσε ίσως να επεκταθεί κάπως αφού ο βαθμός ακρίβειας είναι αρκετά υψηλός, με προφανή στόχο την αύξηση του βαθμού ανάκλησης.

5.1.2 Εννοιολογική απόσταση

Η εννοιολογική απόσταση όρων χρησιμοποιήθηκε με σκοπό να ισχυροποιήσει ή να αποδυναμώσει εκτιμήσεις για ταυτοσημία όρων οι οποίες βασίζονται σε λεκτικά κριτήρια. Κατά την συγχώνευση των θησαυρών CRCS και LCSH/CS χρησιμοποιήσαμε μόνο την απόσταση γενίκευσης αφού και οι δύο θησαυροί δεν υποστηρίζουν κατηγορίες όρων με την έννοια που παρουσιάστηκαν στην εργασία αυτή. Η παράμετρος ω ρυθμίστηκε εμπειρικά στην τιμή 0.275 η οποία δίνει μια καμπύλη σχετικά σταθερή στις διαφορές στην ιεραρχία, πράγμα που έγινε διότι οι δύο θησαυροί παρουσίαζαν σημαντικές διαφορές στην ιεράρχηση των όρων.

Στις περισσότερες περιπτώσεις, ζεύγη ομώνυμων όρων είχαν μεγαλύτερη απόσταση από ζεύγη ταυτόσημων όρων που προτάθηκαν για ενοποίηση όπως μπορεί να παρατηρήσει κανείς απ' τον πίνακα 5.3, όπου για παράδειγμα η απόσταση μεταξύ του όρου Music και του όρου MUSIC (McGill University System for Interactive Computing) είναι εξαιρετικά υψηλή σε αντίθεση με την απόσταση των όρων Music και Music.

ΟΡΟΙ		
CRCS	LCSH/CS	ΑΠΟΣΤΑΣΗ
Music	Music	0.3380
Music	MUSIC	0.9886
Artificial Intelligence	Artificial intelligence	0.2404
Artificial Intelligence	AIS	0.9109
Computer-aided design	Computer-aided design	0.5618
Computer-aided design	CADS	0.6181
Computer-managed instruction	Computer managed instruction	0.5412
Computer-managed instruction	Computer assisted instruction	0.8248

Πίνακας 5.3: Δείγμα εννοιολογικών αποστάσεων όρων του CRCS και του LCSH/CS

Παρουσιάστηκαν ωστόσο και περιπτώσεις ταυτόσημων όρων με μεγάλη απόσταση καθώς και περιπτώσεις μη-ταυτόσημων όρων με σχετικά μικρή απόσταση όπως μπορεί να παρατηρήσει κανείς στον πίνακα 5.4.

Τέτοιες μεγάλες αποστάσεις παρατηρήθηκαν κυρίως στα υψηλά επίπεδα της ιεραρχίας όταν ακόμα δεν είχαν συνδεθεί σε ικανοποιητικό βαθμό οι δύο θησαυροί. Με την πρόοδο της συγχώνευσης όμοιοι όροι που εντοπίστηκαν κάτω από ενοποιημένους όρους παρουσίασαν όπως ήταν αναμενόμενο δραματική μείωση της απόστασης όπως αυτή μετρήθηκε πριν και κατά την διάρκεια της συγχώνευσης, όπως φαίνεται στον πίνακα 5.5.

ΟΡΟΙ		
CRCS	LCSH/CS	ΑΠΟΣΤΑΣΗ
Virtual memory	Virtual storage	0.9992
Calculator	Calculators	0.7353
Neural nets	Neural networks	0.7229
Markets	Marketing	0.3380

Πίνακας 5.4: Δείγμα εννοιολογικών αποστάσεων όρων του CRCS και του LCSH/CS

ΟΡΟΙ		ΑΠΟΣΤΑΣΗ	
CRCS	LCSH/CS	A	B
Engineering	Engineering	0.338	0.338
Software Engineering	Software engineering	0.423	0.212
Computer-aided software engineering	Computer-aided software engineering	0.600	0.054
Computer-aided design	Computer-aided design	0.562	0.000

Πίνακας 5.5: Δείγμα εννοιολογικών αποστάσεων όρων του CRCS και του LCSH/CS Στην στήλη A δίνονται οι αρχικές εννοιολογικές αποστάσεις ενώ στην στήλη B, οι εννοιολογικές αποστάσεις ακριβώς την στιγμή της ενοποίησης των αντίστοιχων όρων. Παρατηρούμε οτι η ενοποίηση του πρώτου ζεύγους προκάλεσε μεγάλη μείωση των αποστάσεων των υπόλοιπων ζευγών.

Συμπερασματικά μπορούμε να πούμε οτι παρά την περιορισμένης έκτασης πειραματική χρήση, η εννοιολογική απόσταση μπορεί να χρησιμοποιηθεί σαν εργαλείο καθοδήγησης του σχεδιαστή κατά τα πρώτα επίπεδα της ιεραρχίας ενώ στην περίπτωση που εντοπισμού όμοιων όρων κάτω από ενοποιημένους όρους, μπορεί να χρησιμοποιηθεί με μεγάλη εμπιστοσύνη για να καθοδηγήσει μηχανικά την διαδικασία της συγχώνευσης.

5.2 Επίλογος

5.2.1 Συνεισφορά

Στην εργασία αυτή, παρουσιάστηκε η σχεδίαση και η υλοποίηση με την χρήση του SIS μιας μεθόδου για την συγχώνευση μονόγλωσσων θησαυρών. Αν εξαιρέσει κανείς την εργασία των Mili και Rada, το θέμα δεν έχει απασχολήσει ιδιαίτερα τους ερευνητές στο χώρο της Επιστήμης Υπολογιστών και υπό αυτή την έννοια η παρούσα εργασία συνεισφέρει στην ερευνητική περιοχή. Ειδικότερα, σε σχέση με την εργασία των Mili και Rada, η παρούσα εργασία προσφέρει επιπλέον μια σαφή και αυστηρή χρήση της σημασιολογίας των συσχετίσεων ακολουθώντας τις κατευθυντήριες γραμμές που τίθενται απ' το πρότυπο ISO-2788 [ISO86] αν και αυτό συνεπάγεται μια υψηλότερη υπολογιστική πολυπλοκότητα. Ο μηχανισμός εντοπισμού όμοιων όρων που χρησιμοποιούμε λαμβάνει υπόψη του διαφορές στην σύνταξη των όρων και επιπλέον αξιοποιεί τις συσχετίσεις ισοδυναμίας. Τέλος εκτός από την λεκτική ομοιότητα, εισάγονται εννοιολογικά κριτήρια για την καθοδήγηση της διαδικασίας ή των σχεδιαστών.

5.2.2 Ανοικτά θέματα

Πειραματική αξιολόγηση. Η επίδοση της μεθόδου που προτείνουμε θα πρέπει ν' αναλυθεί μέσω εκτενών πειραμάτων. Μια τέτοια πειραματική αξιολόγηση θα πρέπει να περιλαμβάνει περιπτώσεις που θα ποικίλουν σε σχέση με:

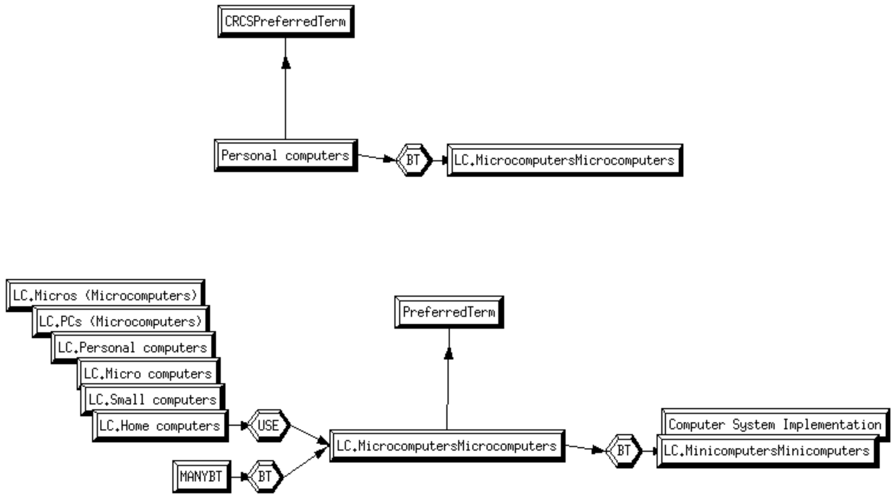
- το μέγεθος και την δομή των συγχωνευόμενων θησαυρών,
- την γνωστική περιοχή την οποία αυτοί καλύπτουν, και
- τους σκοπούς για τους οποίους γίνεται η συγχώνευση.

Επιπλέον, σημαντικό ενδιαφέρον παρουσιάζει η πειραματική αξιολόγηση της μεθόδου στην συγχώνευση θησαυρών που βρίσκονται σε ευρεία χρήση.

Υπολογιστική πολυπλοκότητα. Είναι ίσως σαφές πως μια θεωρητική μελέτη της πολυπλοκότητας της μεθόδου συγχώνευσης παρουσιάζει σημαντικές δυσκολίες. Αυτό οφείλεται στο γεγονός οτι οι γράφοι που σχηματίζονται απ' τις ιεραρχίες γενίκευσης των θησαυρών δεν παρουσιάζουν τυχαιότητα, αφού η ύπαρξη μιας συγκεκριμένης ακμής αποκλείει την ύπαρξη άλλων. Μια εναλλακτική προσέγγιση στο θέμα της πολυπλοκότητας είναι η εμπειρική μελέτη σε κλίμακα τέτοια ώστε να εξασφαλίζεται η ικανοποιητική αξιοπιστία των αποτελεσμάτων.

Ευρετήριο όρων σε πολύ μεγάλους θησαυρούς. Αν και στην τρέχουσα υλοποίηση 8 ως 12 Mb αρκούν για να κρατήσουν στην μνήμη περίπου 1.000.000 όρους, ίσως είναι επιθυμητό το ευρετήριο όρων να κατασκευάζεται στο δίσκο. Σε μια τέτοια περίπτωση θα πρέπει το ευρετήριο να κατασκευάζεται σ' ένα πέρασμα αφού πρώτα έχουν ταξινομηθεί οι υπογραφές των όρων ώστε ν' αποφεύγονται πολλές προσπελάσεις στο δίσκο. Για τον ίδιο επίσης λόγο, θα πρέπει να εγκαταληφθεί η ιδέα του διπλού κατακερματισμού των υπογραφών όπως γίνεται στην τρέχουσα υλοποίηση αφού πλέον το ζητούμενο είναι η ελαχιστοποίηση των προσπελάσεων και όχι η οικονομία σε μνήμη.

Εντοπισμός και διαγραφή πλεονασμών. Από την στιγμή που σε μερικές περιπτώσεις οι συσχετίσεις ισοδυναμίας έχουν σημασιολογία ιεραρχικής συσχέτισης (βλπ. Κεφάλαιο 2) είναι πολύ πιθανό με την προσθήκη νέων δόκιμων όρων στο θησαυρό μέσω της συγχώνευσης, τέτοιες συσχετίσεις είτε ν' αποτελούν πλεονασμό ή να είναι πλέον λανθασμένες. Ενα παράδειγμα δίνεται στο σχήμα 5.3 όπου η συγχώνευση των όρων “Microcomputers” και “LC.Microcomputers” καθιστά την συσχέτιση ισοδυναμίας μεταξύ του “LC.Personal computers” και του ενοποιημένου όρου “LC.MicrocomputersMicrocomputers” περιττή. Η ανακάλυψη τέτοιων πλεονασμών μπορεί να στηριχθεί στην ίδια φιλοσοφία στην οποία βασίζεται και η μέθοδος συγχώνευσης. Η επέκταση της ιδέας αυτής μπορεί να οδηγήσει στον σχεδιασμό ενός διαλογικού εργαλείου αναδόμησης θησαυρών που δημιουργούνται από συγχώνευση, το οποίο εντοπίζει περιοχές του θησαυρού οι οποίες πιθανόν να χρειάζονται αναδόμηση και προτείνει τροποποιήσεις στους σχεδιαστές.



Σχήμα 5.3: Πλεονάζουσα συσχέτιση ισοδυναμίας έπειτα από συγχώνευση.

ΠΑΡΑΡΤΗΜΑ Ι

Η ΓΛΩΣΣΑ Telos ΚΑΙ ΤΟ SIS

Η Telos [MBJK90], είναι μια γλώσσα παράστασης γνώσης η οποία εισάγει ένα οντοκεντρικό μοντέλο δεδομένων. Μια εκδοχή του δομικού της τμήματος έχει υλοποιηθεί στο Ινστιτούτο Επιστήμης Υπολογιστών του Ιδρύματος Τεχνολογίας και Ερευνας [DKT95] και υποστηρίζεται απ' το Σύστημα Σημασιολογικού Ευρετηριασμού (Semantic Index System — SIS). Εδώ θα σκιαγραφήσουμε τα βασικά χαρακτηριστικά τόσο της Telos όσο και του SIS.

Η γλώσσα Telos

Η Telos παρέχει αντικείμενα δύο τύπων: *άτομα (individuals)* που χρησιμοποιούνται για την παράσταση οντοτήτων και *γνωρίσματα (attributes)* που δηλώνουν συσχετίσεις μεταξύ αντικειμένων. Επιπλέον, η γλώσσα παρέχει τρεις *μηχανισμούς αφαίρεσης (abstractions)*: *ταξινόμηση (classification)*, *γενίκευση (generalization)* και *γνωρισματοδότηση (attribution)* με βάση τους οποίους τ' αντικείμενα οργανώνονται σε βάσεις.

Ο μηχανισμός ταξινόμησης διαμερίζει το σύνολο όλων των αντικειμένων μιας βάσης σε στάθμες αφαίρεσης που ονομάζονται *επίπεδα συγκεκριμενοποίησης (instantiation levels)*. Στο πιο χαμηλό επίπεδο βρίσκονται τα *ατομικά αντικείμενα (tokens)*, στο αμέσως επόμενο οι *κλάσεις* (σύνολα ατομικών αντικειμένων), στο αμέσως υψηλότερο, οι κλάσεις κλάσεων που ονομάζονται *μετακλάσεις* κοκ. Κάθε αντικείμενο που βρίσκεται σ' ένα δεδομένο επίπεδο συγκεκριμενοποίησης πρέπει να ταξινομείται ως παράδειγμα ενός τουλάχιστον αντικειμένου που βρίσκεται στο αμέσως υψηλότερο επίπεδο. Για παράδειγμα κάθε ατομικό αντικείμενο πρέπει ν' αποτελεί παράδειγμα μιας κλάσης και κάθε κλάση πρέπει να είναι παράδειγμα μιας μετακλάσης. Προκειμένου να είναι αυτό δυνατό παρέχεται ένα σύνολο προκαθορισμένων κλάσεων συστήματος που αποτελούν και τον αρχικό πληθυσμό κάθε βάσης.

Ο μηχανισμός γενίκευσης —ο οποίος υποστηρίζεται στο επίπεδο κλάσεων και τα ανώτερα αυτού— χρησιμοποιείται για να ορίσει σχέσεις υποσυνόλου μεταξύ κλάσεων του ίδιου επιπέδου συγκεκριμενοποίησης. Επιπλέον μέσω της γενίκευσης, μια κλάση αντικειμένων κληρονομεί όλα τα γνωρίσματα των που ορίζουν ή κληρονομούν οι υπερκλάσεις της.

Μέσω του μηχανισμού γνωρισματοδότησης, κάθε αντικείμενο έχει την δυνατότητα να ορίζει γνωρίσματα τα οποία παίρνουν τιμές σε κάποια κλάση: συστήματος, οριζόμενη από χρήστη ή πρωτογενή. Πρωτογενείς κλάσεις είναι οι κλάσεις `Telos_Integer`, `Telos_Real`, `Telos_String` και `Telos_Time`. Εφόσον τα γνωρίσματα είναι αντικείμενα, μπορούν να ταξινομηθούν σε κλάσεις, οι οποίες μπορεί ν' αποτελούν ειδικεύσεις άλλων κλάσεων και τέλος να ορίζουν δικά τους γνωρίσματα. Ας δούμε ένα παράδειγμα:

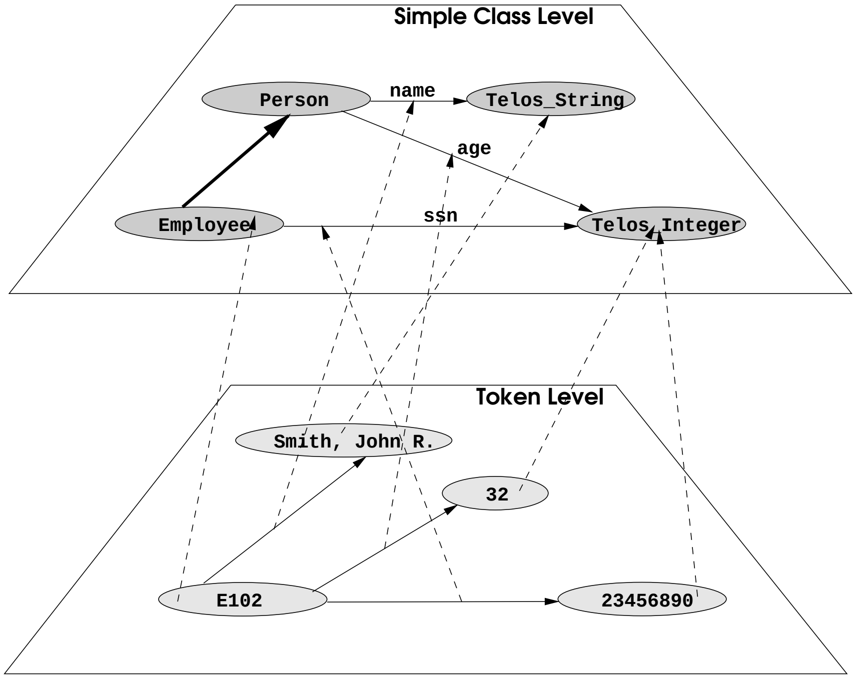
ΠΑΡΑΔΕΙΓΜΑ Ι.1

```
TELL Individual Person
in   S_Class
with attribute
       name : Person;
       age  : Telos_string;
end

TELL Individual Employee
in   S_Class
isA  Person
with attribute
       ssn  : Telos_Integer;
end

TELL Individual E102
in   Token, Employee
with name : "Smith, John R."
with age  : 32
with ssn  : 24456890
end
```

Οι προτάσεις Telos στο παράδειγμα Ι.1, δημιουργούν την βάση που απεικονίζεται στο σχήμα Ι.1.



Σχήμα I.1: Ενα παράδειγμα μιας βάσης Telos
Οι έντονες συμπαγείς ακμές δηλώνουν γενίκευση, οι διακεκομμένες ακμές δηλώνουν ταξινόμηση και οι συμπαγείς ακμές δηλώνουν γνωρισματοδότηση.

Το SIS

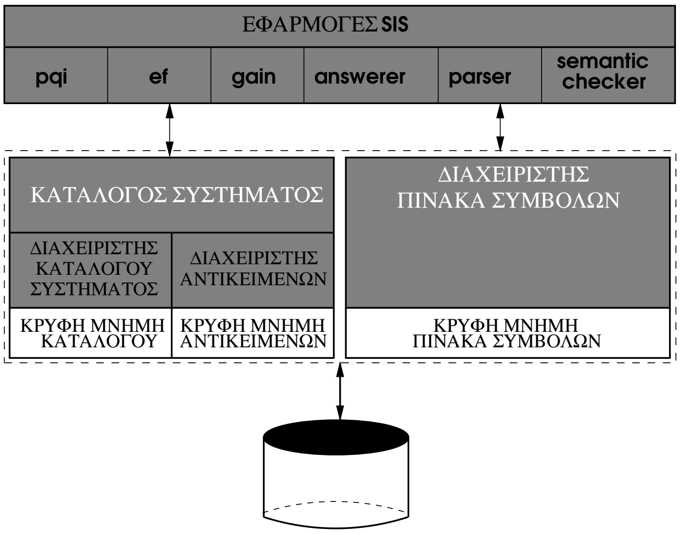
Το SIS είναι ένα *σύστημα διαχείρισης γνώσης (knowledge management system — KMS)* το οποίο ακολουθεί το μοντέλο δεδομένων της γλώσσας Telos. Το SIS αποτελείται από ένα σύνολο εφαρμογών και διεπαφών οι οποίες επικοινωνούν με κάθε βάση δεδομένων μέσω του μηχανισμού αποθήκευσης και διαχείρισης οντοτήτων [Γεω94], όπως φαίνεται στο σχήμα I.2. Οι σημαντικότερες απ' αυτές είναι:

- Ο συντακτικός αναλυτής της Telos είναι η εφαρμογή που χρησιμοποιείται για την κατασκευή μιας βάσης από μια σειρά προτάσεων Telos.

- Ο ελεγκτής σημασιολογίας (Semantic Checker). Μια διεπαφή που χρησιμοποιείται για την ενημέρωση και τον έλεγχο της σημασιολογικής ακεραιότητας μιας βάσης.

- Τα δελτία εισαγωγής δεδομένων (Data Entry Forms — ef). Μια εφαρμογή για την διαλογική ενημέρωση βάσεων.

- Η προγραμματιστική διεπαφή ερωτήσεων (Programatic Query Interface — pqi). Μια διεπαφή που παρέχει την δυνατότητα σε προγράμματα C και C++, να υποβάλλουν ερωτήσεις σε μια βάση.



Σχήμα Ι.2: Η αρχιτεκτονική του SIS

- Ο answerer. Μια διαλογική εφαρμογή για την επερώτηση βάσεων.
- Η διεπαφή γραφικής ανάλυσης Graphical Analysis INterface — gain. Ενα πλήρως προσαρμοζόμενο εργαλείο για την εξερεύνηση βάσεων Telos, με προχωρημένες δυνατότητες παράστασης γράφων. Συνήθως προσαρμόζεται ώστε να εκτελεί προκαθορισμένες ερωτήσεις χρηστών ή να συνδυάζεται με τα ef. Είναι η πιο υψηλού επιπέδου εφαρμογή του SIS.

ΠΑΡΑΡΤΗΜΑ II

ΔΟΜΕΣ ΔΕΔΟΜΕΝΩΝ

Για την υλοποίηση των δομών δεδομένων που χρησιμοποιούνται στην εργασία αυτή, χρησιμοποιούμε την βιβλιοθήκη LEDA (Library of Efficient Data structures and Algorithms) [MN94], [NU95]. Πρόκειται για μια συλλογή αφηρημένων πολυμορφικών, τύπων δεδομένων που έχει υλοποιηθεί σε C++. Κάθε τύπος δεδομένων είναι μια συλλογή στοιχείων που είτε προέρχονται από ένα σύνολο βασικών τύπων ή είναι άλλες δομές. Επειδή υπάρχει μια ιδιομορφία του σχετικού κώδικα, εδώ δίνουμε μια σύντομη αφηρημένη περιγραφή των πράξεων πάνω στις χρησιμοποιούμενες δομές και του απαιτούμενου χρόνου εκτέλεσής τους. Οι βασικές δομές που χρησιμοποιούμε είναι γραμμικές λίστες, σύνολα, πίνακες κατακερματισμού και ουρές προτεραιότητας. Στα επόμενα L, θα παριστάνει μια γραμμική λίστα, H, ένα πίνακα κατακερματισμού, Q, ουρές προτεραιότητας και A, B, C παριστάνουν σύνολα. Τέλος αν S είναι μια δομή, $|S|$ θα συμβολίζει το πλήθος των στοιχείων της.

Ενας πίνακας κατακερματισμού H, είναι μια συλλογή στοιχείων (x, i). Το x ονομάζεται *κλειδί* του στοιχείου και ανήκει σ' ένα σύνολο κλειδιών K, ενώ το i ονομάζεται *πληροφορία* του στοιχείου και ανήκει σ' ένα τύπο δεδομένων I.

Μια ουρά προτεραιότητας Q, είναι μια συλλογή στοιχείων (i, p). Το i ονομάζεται *πληροφορία* του στοιχείου και ανήκει σ' ένα τύπο δεδομένων I ενώ p είναι η *προτεραιότητα* του στοιχείου και προέρχεται από ένα διατεταγμένο σύνολο π.χ., το $\boldsymbol{N}$.

ΠΡΑΞΗ	ΠΕΡΙΓΡΑΦΗ	ΧΡΟΝΟΣ
	ΓΡΑΜΜΙΚΕΣ ΛΙΣΤΕΣ	
$x \leftarrow \text{head}(L)$	Εξαγωγή του στοιχείου στην κεφαλή της L	$O(1)$
$x \leftarrow \text{tail}(L)$	Εξαγωγή του στοιχείου στην ουρά της L	$O(1)$
$\text{insert}(L, x)$	Εισαγωγή του x στην κεφαλή της L	$O(1)$
$\text{append}(L, x)$	Εισαγωγή του x στην ουρά της L	$O(1)$
$L = \emptyset$	Ελεγχος κενής λίστας	$O(1)$
	ΣΥΝΟΛΑ	
$x \in A$	Αναζήτηση στοιχείου	$O(\log \|A\|)$
$\|S\|$	Πληθάριθμος	$O(1)$
$A = \emptyset$	Ελεγχος κενού σύνολου	$O(1)$
$A \leftarrow A - \{x\}$	Διαγραφή του x απ' το A	$O(\log \|A\|)$
$A \leftarrow A \cup \{x\}$	Εισαγωγή του x στο A	$O(\log \|A\|)$
$C \leftarrow A \cup B$	Ενωση	$O(\|A\| \times \log \|B\|)$
$C \leftarrow A \cap B$	Τομή	$O(\|A\| \times \log \|B\|)$
$C \leftarrow A - B$	Διαφορά	$O(\|A\| \times \log \|B\|)$
	ΠΙΝΑΚΕΣ ΚΑΤΑΚΕΡΜΑΤΙΣΜΟΥ	
$x \in H$	Ελεγχος ύπαρξης στοιχείου με κλειδί x	$O(1)$
$i \leftarrow \text{lookup}(x, H)$	Επιστρέφει την πληροφορία του στοιχείου με κλειδί x	$O(1)$
$\text{remove}(x, H)$	Διαγραφή του στοιχείου με κλειδί x απ' τον H	$O(1)$
$\text{insert}(x, i, H)$	Εισαγωγή του στοιχείου (x, i) στον H	$O(1)$
$\text{update}(x, i, H)$	Αλλάζει σε i την πληροφορία του στοιχείου με κλειδί x στον H	$O(1)$
	ΟΥΡΕΣ ΠΡΟΤΕΡΑΙΟΤΗΤΑΣ	
$\text{insert}(i, p, Q)$	Εισαγωγή του i στην Q με προτεραιότητα p	$O(1)$
$\text{remove}(i, Q)$	Διαγραφή του i απ' την Q	$O(1)$
$(x, p) \leftarrow \text{getmin}(Q)$	Εξαγωγή του στοιχείου με την μεγαλύτερη προτεραιότητα	$O(1)$

Πίνακας ΙΙ.1: Δομές δομές δεδομένων και η πολυπλοκότητα των πράξεων τους

ΠΑΡΑΡΤΗΜΑ III

ΜΑΘΗΜΑΤΙΚΟ ΠΑΡΑΡΤΗΜΑ

ΘΕΩΡΗΜΑ III.1 *Η δ είναι ψευδομετρική*[1].

Απόδειξη:

ΛΗΜΜΑ III.1 *Η D_I είναι ψευδομετρική.*

Απόδειξη:
Αφού $\mathrm{Range}(D_I) = \{0,1\}$ θα είναι

$$(x,y) \in P \times P \Rightarrow D_I(x,y) \geq 0 \tag{III.5}$$

Εξάλλου $x = y \Rightarrow \#x = \#y$ —απ' τον ορισμό 3.2— άρα

$$x \in P \Rightarrow D_I(x,x) = 0 \tag{III.6}$$

[1]Εστω S ένα σύνολο και d μια συνάρτηση απ' το $S \times S$ στο $\boldsymbol{R}$. Η d καλείται ψευδομετρική στο S, ανν ικανοποιεί τις παρακάτω σχέσεις:

$$\begin{aligned} (a,b) \in S \times S &\Rightarrow d(a,b) \geq 0 && \text{(III.1)} \\ a \in S &\Rightarrow d(a,a) = 0 && \text{(III.2)} \\ (a,b) \in S \times S &\Rightarrow d(a,b) = d(b,a) && \text{(III.3)} \\ (a,b,c) \in S \times S \times S &\Rightarrow d(a,b) + d(b,c) \geq d(a,c) && \text{(III.4)} \end{aligned}$$

Οι σχέσεις (III.1) και (III.2) ονομάζονται συνθήκες ελαχιστότητας, η (III.3) συνθήκη συμμετρίας και η (III.4) συνθήκη τριγωνικότητας ή τριγωνική ανισότητα. Αν επιπλέον των συνθηκών αυτών ικανοποιείται και η συνθήκη

$$(a,b) \in S \times S \text{ και } d(a,b) = 0 \Rightarrow a = b$$

τότε η d καλείται μετρική στο S.

Επίσης λόγω της συμμετρίας της σχέσης ισότητας θα είναι

$$(x,y) \in P \times P \Rightarrow D_I(x,y) = D_I(y,x) \tag{III.7}$$

Τέλος, αν $D_I(x,z) = 0$, τότε λόγω της ελαχιστότητας θα είναι $D_I(x,y) + D_I(y,z) \geq 0 = D_I(x,z)$. Αν αντίθετα $D_I(x,z) = 1$, τότε δεν μπορεί να είναι $D_I(x,y) + D_I(y,z) = 0$, διότι θα πρέπει να ισχύει $\#x = \#y$ και $\#y = \#z$ και κατά συνέπεια $\#x = \#z$ το οποίο είναι άτοπο. Αρα σε κάθε περίπτωση ισχύει

$$(x,y,z) \in P \times P \times P \Rightarrow D_I(x,y) + D_I(x,z) \geq D_I(x,z) \tag{III.8}$$

Από τις (III.5)–(III.8) προκύπτει οτι η D_I είναι ψευδομετρική. □

ΛΗΜΜΑ III.2 *Αν A, B, C είναι υποσύνολα του $\boldsymbol{N}$ και η συνάρτηση $f : 2^{\boldsymbol{N}} \longrightarrow [0, +\infty)$ ικανοποιεί τις σχέσεις*

$$A \supseteq B \quad \Rightarrow \quad f(A) \geq f(B) \tag{III.9}$$

$$f(A \cup B) \quad \leq \quad f(A) + f(B) \tag{III.10}$$

τότε:

$$f(A \div B) + f(B \div C) \geq f(A \div C) \tag{III.11}$$

<u>Απόδειξη:</u>
Είναι γνωστό απ' την θεωρία συνόλων[2] οτι

$$(A \div B) \cup (B \div C) \supseteq (A \div C) \tag{III.12}$$

Κατά συνέπεια

$$\begin{aligned} (III.12) &\overset{(III.9)}{\Longrightarrow} \quad f((A \div B) \cup (B \div C)) \geq f(A \div C) \\ &\overset{(III.10)}{\Longrightarrow} \quad f(A \div B) + f(B \div C) \geq f(A \div C) \end{aligned}$$

□

Η συνάρτηση που δίνει τον πληθάριθμο ενός συνόλου ικανοποιεί τις (III.9) και (III.10), αφού αντικαθιστώντας σ' αυτές την f λαμβάνουμε δύο γνωστές ταυτότητες της θεωρίας συνόλων:

$$A \supseteq B \quad \Rightarrow \quad |A| \geq |B|$$

$$|A \cup B| \quad \leq \quad |A| + |B|$$

Συνεπώς ισχύει οτι

$$|A \div B| + |B \div C| \geq |A \div C| \tag{III.13}$$

[2] Βλπ. [RW85]

Ομοίως, η συνάρτηση

$$\sum_{a \in A} 1/g(a), \ g : A \longrightarrow \boldsymbol{N} - \{0\}$$

ικανοποιεί τόσο την (III.9) αφού η g παίρνει μόνο θετικές τιμές, όσο και την (III.10) αφού

$$\begin{aligned} \sum_{a \in A \cup B} 1/g(a) &= \sum_{a \in A} 1/g(a) + \sum_{a \in B} 1/g(a) - \sum_{a \in A \cap B} 1/g(a) \\ &\leq \sum_{a \in A} 1/g(a) + \sum_{a \in B} 1/g(a) \end{aligned}$$

και κατά συνέπεια είναι

$$\sum_{a \in A \div B} 1/g(a) + \sum_{a \in B \div C} 1/g(a) \geq \sum_{a \in A \div C} 1/g(a) \tag{III.14}$$

ΛΗΜΜΑ III.3 *Η D_C είναι ψευδομετρική.*

Απόδειξη:
Η D_C —ως πληθάριθμος συνόλου— παίρνει μη-αρνητικές τιμές. Συνεπώς

$$(x, y) \in P \times P \Rightarrow D_C(x, y) \geq 0 \tag{III.15}$$

Επιπλέον, $x.F \div x.F = \emptyset \Rightarrow |x.F \div x.F| = 0$ και κατά συνέπεια

$$x \in P \Rightarrow D_C(x, x) = 0 \tag{III.16}$$

Εξάλλου αφού $x.F \div y.F = y.F \div x.F$ θα είναι

$$(x, y) \in P \times P \Rightarrow D_C(x, y) = D_C(y, x) \tag{III.17}$$

Τέλος θέτοντας στην (III.13), $A = x.F, B = y.F, C = z.F$ παίρνουμε

$$|x.F \div y.F| + |y.F \div z.F| \geq |x.F \div z.F|$$

και κατά συνέπεια

$$(x, y, z) \in P \times P \times P \Rightarrow D_C(x, y) + D_C(x, z) \geq D_C(x, z) \tag{III.18}$$

Από τις (III.15)–(III.18) προκύπτει οτι η D_C είναι ψευδομετρική. □

ΛΗΜΜΑ III.4 *Η D_G είναι ψευδομετρική.*

Απόδειξη:
Κατ' αναλογία με την απόδειξη του λήμματος III.3 προκύπτουν οι σχέσεις:

$$\begin{aligned} (x, y) \in P \times P &\Rightarrow D_G(x, y) \geq 0 && \text{(III.19)} \\ x \in P &\Rightarrow D_G(x, x) = 0 && \text{(III.20)} \\ (x, y) \in P \times P &\Rightarrow D_G(x, y) = D_G(y, x) && \text{(III.21)} \end{aligned}$$

Αντικαθιστώντας τέλος στην (III.14), $A = x.F, B = y.F, C = z.F$ και $g = L$, παίρνουμε

$$\sum_{\#i \in x.B \div y.B} 1/L(i) + \sum_{\#i \in y.B \div z.B} 1/L(i) \geq \sum_{\#i \in x.B \div z.B} 1/L(i)$$

Κατά συνέπεια:

$$(x, y, z) \in P \times P \times P \Rightarrow D_G(x, y) + D_G(x, z) \geq D_G(x, z) \tag{III.22}$$

Απ' τις (III.19)–(III.22) προκύπτει οτι η D_G είναι ψευδομετρική. □

ΛΗΜΜΑ III.5 *Αν $w_1, w_2, \cdots, w_k$ είναι θετικοί πραγματικοί αριθμοί και $D_1, D_3, \cdots, D_k$ είναι ψευδομετρικές στο σύνολο S, τότε η συνάρτηση $D : S \times S \longrightarrow \boldsymbol{R}$ που ορίζεται ως*

$$D(a, b) = \sum_{i=1}^{k} w_i D_i(a, b)$$

είναι επίσης ψευδομετρική.

Απόδειξη:
Αφού $w_i \geq 0,\ i = 1, 2, \cdots, k$ και $(a, b) \in S \times S \Rightarrow D_i(a, b) \geq 0,\ i = 1, 2, \cdots, k$, θα είναι

$$\sum_{i=1}^{k} w_i D_i \geq 0$$

κατά συνέπεια,

$$(a, b) \in S \times S \Rightarrow D(a, b) \geq 0 \tag{III.23}$$

Εξάλλου, αφού $a \in S \Rightarrow D_i(a, a) = 0,\ i = 1, 2, \cdots, k$, θα είναι

$$a \in S \Rightarrow D(a, a) = 0 \tag{III.24}$$

Επίσης, έχουμε

$$D(a, b) = \sum_{i=1}^{k} w_i D_i(a, b) = \sum_{i=1}^{k} w_i D_i(b, a)$$

δηλαδή:

$$(a, b) \in S \times S \Rightarrow D(a, b) = D(b, a) \tag{III.25}$$

Τέλος λόγω της τριγωνικότητας των D_i έχουμε:

$$\begin{array}{ccccc} w_1 D_1(a, b) & + & w_1 D_1(b, c) & \geq & w_1 D_1(a, c) \\ w_2 D_2(a, b) & + & w_2 D_2(b, c) & \geq & w_2 D_2(a, c) \\ \vdots & + & \vdots & \geq & \vdots \\ w_k D_k(a, b) & + & w_k D_k(b, c) & \geq & w_k D_k(a, c) \end{array} \tag{III.26}$$

Προσθέτοντας κατά μέλη τις ανιοότητες (III.26) παίρνουμε

$$\sum_{i=1}^{k} w_i D_i(a,b) + \sum_{i=1}^{k} w_i D_i(b,c) \geq \sum_{i=1}^{k} w_i D_i(a,c)$$

και κατά συνέπεια προκύπτει οτι

$$(a,b,c) \in S \times S \times S \Rightarrow D(a,b) + D(b,c) \geq D(a,c) \tag{III.27}$$

Από τις σχέσεις (III.23)–(III.27) αποδεικνύεται οτι η D είναι μια ψευδομετρική στο S.
□

Εφαρμόζοντας αυτό το αποτέλεσμα για $k = 3, w_1 = w_I, w_2 = w_C, w_3 = w_G, D_1 = D_I, D_2 = D_C, D_3 = D_G$ και $S = P$, προκύπτει οτι η συνάρτηση συνάθροισης

$$\Delta = w_I D_I + w_C D_C + w_G D_G$$

είναι μια ψευδομετρική στο σύνολο των δόκιμων όρων ενός θησαυρού.

Κατά συνέπεια έχουμε:

$$(x,y) \in P \times P \Rightarrow \Delta(x,y) \geq 0 \Rightarrow 1 - e^{-\omega\Delta(x,y)} \geq 0$$

ή

$$(x,y) \in P \times P \Rightarrow \delta(x,y) \geq 0 \tag{III.28}$$

Επίσης

$$x \in P \Rightarrow \Delta(x,x) = 0 \Rightarrow e^{-\omega\Delta(x,x)} = 1 \Rightarrow 1 - e^{-\omega\Delta(x,x)} = 0$$

άρα:

$$x \in P \Rightarrow \delta(x,x,\omega) = 0 \tag{III.29}$$

Εξάλλου αφού η Δ είναι συμμετρική ισχύει οτι

$$(x,y) \in P \times P \Rightarrow \Delta(x,y) = \Delta(y,x) \Rightarrow 1 - e^{-\omega\Delta(x,y)} = 1 - e^{-\omega\Delta(y,x)}$$

και επομένως

$$(x,y) \in P \times P \Rightarrow \delta(x,y,\omega) = \delta(y,x,\omega) \tag{III.30}$$

Τέλος αφού η Δ είναι τριγωνική, για $(x,y,z) \in P \times P \times P$ και $\omega \in (0,+\infty)$ θα είναι

$$\begin{array}{llllll}
\Delta(x,y) & + & \Delta(y,z) & \geq & \Delta(x,z) & \Rightarrow \\
-\omega\Delta(x,y) & - & \omega\Delta(y,z) & \leq & -\omega\Delta(x,z) & \Rightarrow \\
e^{-\omega\Delta(x,y)-\omega\Delta(y,z)} & & & \leq & e^{-\omega\Delta(x,z)} & \Rightarrow \\
e^{-\omega\Delta(x,y)} & + & e^{-\omega\Delta(y,z)} & \leq & e^{-\omega\Delta(x,z)} & \Rightarrow \\
-e^{-\omega\Delta(x,y)} & - & e^{-\omega\Delta(y,z)} & \geq & -e^{-\omega\Delta(x,z)} & \Rightarrow \\
1 - e^{-\omega\Delta(x,y)} & + & 1 - e^{-\omega\Delta(y,z)} & \geq & 1 - e^{-\omega\Delta(x,z)} &
\end{array}$$

ή

$$(x,y,z) \in P \times P \times P \Rightarrow \delta(x,y,\omega) + \delta(y,z,\omega) \geq \delta(x,z,\omega) \tag{III.31}$$

Από τις (III.28)–(III.31) αποδεικνύεται ότι η κανονικοποιημένη απόσταση όρων είναι μια ψευδομετρική στο σύνολο των δόκιμων όρων ενός θησαυρού. □

ΘΕΩΡΗΜΑ III.2 *Εστω θ ένας θησαυρός και $x, y \in \theta.P$. Εστω επίσης θ' ο θησαυρός που προκύπτει απ' την ενοποίηση των όρων x, y και την παραγωγή του όρου $z \in \theta'.P$ απ' τον αλγόριθμο 4.7. Αν $\theta \models \boldsymbol{M}$ τότε και $\theta' \models \boldsymbol{M}$.*

<u>Απόδειξη:</u>
Εστω οτι $\theta \models \boldsymbol{M}$ και $\theta' \not\models \boldsymbol{M}$. Τότε

$$\exists\ \#z_1, \#z_2 \in z.B : z_1 \rightsquigarrow z_2 \text{ ή } z_2 \rightsquigarrow z_1 \tag{III.32}$$

Ας υποθέσουμε οτι

$$z_1 \rightsquigarrow z_2 \tag{III.33}$$

Τότε:

- Αν $\#z_1, \#z_2 \in x.B$, η (III.32) δεν μπορεί να ισχύει αφού $\theta \models \boldsymbol{M}$.
- Αν $\#z_1, \#z_2 \in y.B$, η (III.32) δεν μπορεί να ισχύει αφού $\theta \models \boldsymbol{M}$.
- Αν $z_1 \in x.B$ και $z_2 \in y.B$, τότε $z_2 \in G^+(y)$ και κατά συνέπεια

  $$z_2 \notin x.B - (G^+(y) - y.B) \tag{III.34}$$

  Εξάλλου $z_1 \rightsquigarrow z_2 \Rightarrow z_2 \in G^+(x)$ και αφού $z_2 \notin x.B$ θα είναι

  $$z_2 \notin y.B - (G^+(x) - x.B) \tag{III.35}$$

  Συνδυάζοντας τις (III.34), (III.35) και τον αλγόριθμο 4.7 προκύπτει οτι $z_2 \notin z.B$. Ατοπο αφού δεχθήκαμε οτι $z_2 \in z.B$.
- Ομοίως αποδεικνύεται οτι αν $z_1 \in y.B$ και $z_2 \in x.B$, θα είναι $z_1 \notin z.B$ κάτι επίσης άτοπο.

Κατά συνέπεια θα πρέπει να δεχθούμε οτι η (III.33) δεν ισχύει και επομένως

$$z_1 \not\rightsquigarrow z_2 \tag{III.36}$$

Ομοια επίσης αποδυκνύεται οτι

$$z_2 \not\rightsquigarrow z_1 \tag{III.37}$$

Από τις (III.36) και (III.37) προκύπτει το συμπέρασμα ότι $\theta' \models \boldsymbol{M}$. □

ΠΑΡΑΡΤΗΜΑ IV

ΓΛΩΣΣΑΡΙ

Α

Αδόκιμος όρος	Non-Preferred term
Ακυκλικός κατευθυνόμενος γράφος	Directed acyclic graph
Αμφιμονοσήμαντη αντιστοιχία	Bijection
Ανάκληση πληροφορίας	Information retrieval
Αναγνωριστής όρου	Term identifier
Αναδιάρθρωση	Conformation
Ανεστραμένη σύνταξη (όρου)	Reverse syntax
Αντικείμενο	Object
Απόσταση γενίκευσης	Generalization distance
Απόσταση ταξινόμησης	Classification distance
Απόσταση ταύτισης	Identification distance
Ασαφής	Fuzzy
Ατομικό αντικείμενο	Token
Αφαίρεση καταλήξεων	Suffix stripping

Β

Βαθμός ακρίβειας	Precision
Βαθμός ανάκλησης	Recall
Βαθμός ειδίκευσης	Specialization level

Γ

Γενίκευση	Generalization
Γενικότερος όρος	Broader term
Γνωρισματοδότηση	Attribution
Γνώρισμα	Attribute

Δ

Δεσμικός	Batch
Διαλογικός	Interactive
Διάταξη	Partial order
Διαφορά	Dissimilarity
Δόκιμος όρος	Preferred term

Ε

Ειδίκευση	Specialization
Ειδικότερος όρος	Narrower term
Ελεγχόμενο λεξιλόγιο	Controlled vocabulary
Εννοια	Concept
Εννοιολογική απόσταση	Conceptual distance
Εννοιολογικό σχήμα	Conceptual schema
Ενοποίηση	Integration
Ενταση (της ομοιότητας)	Aptness (of similarity)
Επίπεδο συγκεκριμενοποίησης	Instantiation level
Ερώτηση	Query
Ευρετήριο	Index
Ευρετηριασμός	Indexing

Θ

Θησαυρός	Thesaurus

Ι

Ιεραρχική συσχέτιση γενίκευσης	Generic hierarchical relationship
Ιεραρχική συσχέτιση μέρους όλου	Part-whole hierarchical relationship
Ιεραρχική συσχέτιση παραδείγματος	Instance-Of hierarchical relationship
Ιεραρχικό μονοπάτι	Hierarchical path

Κ

Κατάληξη ή επίθεμα	Suffix
Κατηγορία όρων	Facet
Κλειστό περίβλημα	Transitive closure

Μ

Μετρική	Metric
Μέτρο διαφοράς	Dissimilarity measure
Μέτρο ομοιότητας	Similarity measure
Μετακλάση	Metaclass
Μηχανισμός αφαίρεσης	Abstraction
Μοντέλο δεδομένων	Data model
Μονόγλωσσος θησαυρός	Monolingual thesaurus

Ο

Ολική διάταξη	Total order
Ομοιότητα	Similarity
Ορος	Term
Οψη	View

Π

Πεδίο γνώσης	Knowledge domain
Περιορισμός (ή συνθήκη) ακαιρεότητας	Integrity constraint
Πλειάδα	Tuple
Πολύγλωσσος θησαυρός	Multilingual thesaurus
Προενοποίηση	Pre-integration
Πρόθεμα	Prefix

Ρ

Ρίζα (όρου)	Stem

Σ

Σημασιολογία	Semantics
Συγχώνευση θησαυρών	Thesauri merging
Συνάθροιση	Aggregation
Συνάρτηση απόστασης	Distance function
Συναφής όρος	Related term
Συντακτική παραγοντοποίηση (όρου)	Syntactical factoring

Συσχέτιση	Relationship
Σχέση γενίκευσης	Generalization relation
Σχέση ειδίκευσης	Specialization relation
Σχέση ισοδυναμίας	Equivalence relation
Σχέση συνάφειας	Association relation
Σχέση	Relation
Σύγκρουση	Conflict
Σύνταξη φυσικής γλώσσας	Natural language syntax

Τ

Ταξινόμηση	Classification
Τοπολογική ταξινόμηση	Topological sorting

Υ

Υπερκλάση	Superclass
Υπογραφή όρου	Term signature
Υποκλάση	Subclass

Ψ

Ψευδοδιάταξη	Pseudo-order
Ψευδομετρική	Pseudo-metric

ΠΑΡΑΡΤΗΜΑ V

GLOSSARY

A

Abstraction	Μηχανισμός αφαίρεσης
Aggregation	Συνάθροιση
Aptness (of similarity)	Ενταση (της ομοιότητας)
Association relation	Σχέση συνάφειας
Attribute	Γνώρισμα
Attribution	Γνωρισματοδότηση

B

Batch	Δεσμικός
Bijection	Αμφιμονοσήμαντη αντιστοιχία
Broader term	Γενικότερος όρος

C

Classification	Ταξινόμηση
Classification distance	Απόσταση ταξινόμησης
Concept	Εννοια
Conceptual distance	Εννοιολογική απόσταση
Conceptual schema	Εννοιολογικό σχήμα
Conflict	Σύγκρουση
Conformation	Αναδιάρθρωση
Controlled vocabulary	Ελεγχόμενο λεξιλόγιο

D

Data model	Μοντέλο δεδομένων
Directed acyclic graph	Ακυκλικός κατευθυνόμενος γράφος
Dissimilarity	Διαφορά
Dissimilarity measure	Μέτρο διαφοράς
Distance function	Συνάρτηση απόστασης

E

Equivalence relation	Σχέση ισοδυναμίας

F

Facet	Κατηγορία όρων
Fuzzy	Ασαφής

G

Generalization	Γενίκευση
Generalization distance	Απόσταση γενίκευσης
Generalization relation	Σχέση γενίκευσης
Generic hierarchical relationship	Ιεραρχική συσχέτιση γενίκευσης

H

Hierarchical path	Ιεραρχικό μονοπάτι

I

Identification distance	Απόσταση ταύτισης
Index	Ευρετήριο
Indexing	Ευρετηριασμός
Information retrieval	Ανάκληση πληροφορίας
Instance-Of hierarchical relationship	Ιεραρχική συσχέτιση παραδείγματος
Instantiation level	Επίπεδο Συγκεκριμενοποίησης
Integration	Ενοποίηση
Integrity constraint	Περιορισμός (ή συνθήκη) ακαιρεότητας
Interactive	Διαλογικός

K

Knowledge domain	Πεδίο γνώσης

M

Metaclass	Μετακλάση
Metric	Μετρική
Monolingual thesaurus	Μονόγλωσσος θησαυρός
Multilingual thesaurus	Πολύγλωσσος θησαυρός

N

Narrower term	Ειδικότερος όρος
Natural language syntax	Σύνταξη φυσικής γλώσσας
Non-Preferred term	Αδόκιμος όρος

O

Object	Αντικείμενο

P

Part-whole hierarchical relationship	Ιεραρχική συσχέτιση μέρους όλου
Partial order	Διάταξη
Pre-integration	Προενοποίηση
Precision	Βαθμός ακρίβειας
Preferred term	Δόκιμος όρος
Prefix	Πρόθεμα
Pseudo-metric	Ψευδομετρική
Pseudo-order	Ψευδοδιάταξη

Q

Query	Ερώτηση

R

Recall	Βαθμός ανάκλησης
Related term	Συναφής όρος
Relation	Σχέση
Relationship	Συσχέτιση
Reverse syntax	Ανεστραμένη σύνταξη (όρου)

S

Semantics	Σημασιολογία

Similarity	Ομοιότητα
Similarity measure	Μέτρο ομοιότητας
Specialization	Ειδίκευση
Specialization level	Βαθμός ειδίκευσης
Specialization relation	Σχέση ειδίκευσης
Stem	Ρίζα (όρου)
Subclass	Υποκλάση
Suffix	Κατάληξη ή επίθεμα
Suffix stripping	Αφαίρεση καταλήξεων
Superclass	Υπερκλάση
Syntactical factoring	Συντακτική παραγοντοποίηση (όρου)

T

Term	Ορος
Term identifier	Αναγνωριστής όρου
Term signature	Υπογραφή όρου
Thesauri merging	Συγχώνευση θησαυρών
Thesaurus	Θησαυρός
Token	Ατομικό αντικείμενο
Topological sorting	Τοπολογική ταξινόμηση
Total order	Ολική διάταξη
Transitive closure	Κλειστό περίβλημα
Tuple	Πλειάδα

V

View	Οψη

ΒΙΒΛΙΟΓΡΑΦΙΑ

[Ντα93] Γ. Νταντουρής. Βιβλιοθήκη Στοιχειωδών Ερωτηματικών Συναρτήσεων και Επεξεργασία Ερωτήσεων για την Γλώσσα telos. Μεταπτυχιακή Εργασία, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης, 1993.

[Γεω94] Γ. Γεωργιαννάκης. Ο Μηχανισμός Αποθήκευσης και Διαχείρησης Οντοτήτων για την Γλώσσα Παράστασης Γνώσης Telos. Μεταπτυχιακή Εργασία, Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης, 1994.

[Baa88] S. Baase. *Computer Algorithms: Introduction to Design and Analysis.* Addison-Wesley, 2nd edition, 1988.

[BCDA+96] C. Batini, S. Castano, V. De Antonellis, M. G. Fugini, and B. Percini. Analysis of an Inventory of Information Systems in the Public Administration. *Requirements Engineering*, 1(1):47–62, 1996.

[BL84] C. Batini and M. Lenzerini. A Methodology for Data Schema Integration in the Entity Relationship Model. *IEEE Transactions on Software Engineering*, SE-10(6):650–664, 1984.

[BL86] C. Batini and M. Lenzerini. A Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys*, 18(4):323–364, 1986.

[Bra83] R. J. Brachman. What IS-A is and isn't: An Analysis of Taxonomic Links in Semantic Networks. *IEEE Computer*, pages 30–36, October 1983.

[CD95] P. Constantopoulos and M. Doerr. Component Classification in the Software Information Base. In Nierstrasz O. and Tsichritzis D., editors, *Object-Oriented Software Composition.* Prentice Hall, 1995.

[Chr96] T. Chrysos. Design and Implementation of a Thesaurus Browsing System and Connection with a Bibliographic System. Diploma thesis, Department of Computer Science, University of Crete, October 1996. http://www.csd.uch.gr/chrysos/diploma.

[CLBD93] H. Chen, K. Lynch, K. Basu, and T. Dorbin. Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval. *IEEE Expert*, April 1993.

[Cro90] C. J. Crouch. An Approach to the Automatic Construction of Global Thesauri. *Information Processing and Management*, 26(5):629–640, 1990.

[DH84] U. Dayal and H. Hwang. View Definition and Generalization for Database Integration in a Multidatabase System. *IEEE Transactions on Software Engineering*, SE-10(6):628–644, 1984.

[DKT95] M. Doerr, P. Klimathianakis, and M. Theodorakis. *SIS Data Entry Language User's Manual.* Institute of Computer Science, FOundation of Research and Technology—Hellas, Heraklion, Crete, Greece, 1995.

[Doe96] M. Doerr. Authority Services in Global Information Spaces: A Requirements Analysis and Feasibility Study. Technical Report 163, Institute of Computer Science, FOountation for Research and Technology—Hellas, Heraklion, Crete, Greece, 1996.

[FT96] William Ford and William Topp. *Data Structures with C++*. Prentice Hall, 1996.

[GI94] M. R. Girardi and I. Ibrahim. A Similarity Measure for Retrieving Software Artifacts. University of Geneva, Centre Universitaire d' Informatique, 1994.

[GLN92] W. Gotthard, P. Lockemann, and A. Neufeld. System-Guided View Integration for Object Oriented Databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(1):1–22, 1992.

[ISO85] ISO. Documentation — Guidelines for the Development and Establishment of Multilingual Thesauri, February 1985. International Standard ISO 5964.

[ISO86] ISO. Documentation — Guidelines for the Development and Establishment of Monolingual Thesauri, November 1986. International Standard ISO 2788.

[Knu73] D. Knuth. *Fundamental Algorithms*, volume 1 of *The Art of Computer Programming*. Addison Wesley, 1973.

[Kri94] J. Kristensen. Expanding User Queries Using a Search-Aid Thesaurus. *Information Processing and Management*, 29(6):733–744, 1994.

[KS93] T. Z. Kalaboukis and M. M. Sintichakis. Suffix Stripping with Modern Greek. Department of Informatics, Athens University of Economics, 1993.

[LKL94] J. H. Lee, M. H. Kim, and Y. J. Lee. Ranking Documents in Thesaurus-Based Boolean Retrieval Systems. *Information Processing and Management*, 30(1):79–91, 1994.

[Maz94] Z. Mazur. Models of a Distributed Information Retrieval System Based on Thesauri with Weights. *Information Processing and Management*, 30(1):61–77, 1994.

[MBJK90] J. Mylopoylos, A. Borgida, M. Jarke, and M. Koubarakis. Telos: Representing Knowledge About Information Systems. *ACM Transactions on Information Systems*, 8(4):325–362, 1990.

[Mil91] J. Milstead. Specifications for Thesaurus Software. *Information Processing and Management*, 27(2):165–175, 1991.

[MN94] K. Mehlhorn and S. Näher. LEDA: A Platform for Combinatorial and Geometric Computing. Max-Planckt-Institut für Informatik, 66123 Saarbrücken, Germany, 1994.

[MNE88] M. V. Mannino, S. B. Navathe, and W. Effelsberg. A Rule-Based Approach for Merging Generalization Hierarchies. *Information Systems*, 13(3):257–272, 1988.

[MR88] H. Mili and R. Rada. Merging Thesauri: Principles and Evaluation. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 10(2):204–220, 1988.

[NU95] S. Näher and C. Uhrig. *The LEDA User Manual*, 1995.

[Pai91] C. Paice. A Thesaural Model of Information Retrieval. *Information Processing and Management*, 27(5):433–447, 1991.

[Por80] M. F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.

[RM87] R. Rada and B. K. Martin. Augmenting Thesauri for Information Systems. *ACM Transactions on Office Information Systems*, 5(4):378–392, 1987.

[RM89] R. Rada and H. Mili. A Knowledge-Intensive Learning System for Document Retrieval. In K. Morik, editor, *Knowledge Representation and Organization in Machine Learning*, volume 347 of *LNAI*, pages 65–87. Springer Verlag, Berlin, 1989.

[RW85] K. Ross and C. Wright. *Discrete Mathematics*. Prentice-Hall, 1985.

[Sal89] G. Salton. *Automatic Text Processing*. Addisson Wesley, 1989.

[SC96] G. Spanoudakis and P. Constantopoulos. Elaborating Analogies from Conceptual Models. *International Journal of Intelligent Systems*, 11(11):917–974, 1996.

[SC97] M. Sintichakis and P. Constantopoulos. A Method for Monolingual Thesauri Merging. Submitted to the 20th Iinternational ACM SIGIR Conference on Research and Development in Information Retrieval, 1997.

[Soe95] D. Soergel. The Arts and Architecture Thesaurus (AAT): A Critical Appraisal. *Visual Resources*, 10(4):369–400, 1995.

[SP94] S. Spaccapietra and C. Parent. View Iintegration: A Step Forward in Solving Structural Conflicts. *IEEE TRansactions on Knowledge and Data Engineering*, 6(2):258–274, 1994.

[Spa94] G. Spanoudakis. *Analogical Similarity of Objects: A Conceptual Modeling Approach*. PhD thesis, Department of Computer Science University of Crete, 1994.

[Sve89] E. Svenonius. Design of Controlled Vocabuaries. In *Encyclopedia of Library and Information Science*, pages 82–109. Marcel Dekker, 1989.

[TL82] D. C. Tsichritzis and F. H. Lochovsky. *Data Models*. Prentice-Hall, 1982.

[Tve77] A. Tversky. Features of Similarity. *Psychological Review*, 84(4):327–362, 1977.