19ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών, 3-5 Νοεμβρίου 2010, Αθήνα

Εννοιολογική Διεύρυνση Ερωτημάτων με τη Χρήση Θησαυρού: μια εμπειρική μελέτη

Άννα Μάστορα(1)
Μαρία Μονόπωλη(2)
Σαράντος Καπιδάκης(1)

(1)Εργαστήριο Ψηφιακών Βιβλιοθηκών & Ηλεκτρονικής Δημοσίευσης, Τμήμα Αρχειονομίας - Βιβλιοθηκονομίας, Ιόνιο Πανεπιστήμιο, Ιωάννου Θεοτόκη 72, 49100, Κέρκυρα
(2)Βιβλιοθήκη, Τράπεζα της Ελλάδος, Ελ. Βενιζέλου 21, 102 50, Αθήνα

Δομή της παρουσίασης

- Εισαγωγή - Βασικές έννοιες
- Στόχος της έρευνας
- Επισκόπηση του πεδίου
- Μεθοδολογία
- Αποτελέσματα
- Συμπεράσματα
- Μελλοντική έρευνα

Εισαγωγή - Βασικές έννοιες

- Γνωσιακή Ανάκτηση Πληροφορίας
  (Cognitive Information Retrieval)
- Εννοιολογική Διεύρυνση Ερωτημάτων
  (Conceptual Query Expansion)
- Συστήματα Οργάνωσης της Γνώσης (ΣΟΓ)
  (Knowledge Organisation Systems)
- Πολυ-αναπαράσταση της Γνώσης
  (Knowledge Poly-representation)
- Χρήστες, μη-ειδικοί
  (Users, non-experts)

Στόχος της έρευνας

Διερεύνηση των προτύπων συμπεριφοράς των χρηστών κατά την επιλογή και δόμηση των όρων αναζήτησης

- καταγράφοντας την αντίληψη για την αναπαράσταση της γνώσης που έχουν οι χρήστες (μη-ειδικοί),

σε αντιπαραβολή με

- την αναπαράσταση της γνώσης σε υφιστάμενα Συστήματα Οργάνωσης της Γνώσης (ΣΟΓ),

με σκοπό

να εντοπίσουμε τις περιπτώσεις στις οποίες μπορούν να χρησιμεύσουν τα ΣΟΓ

για εννοιολογική διεύρυνση των ερωτημάτων

Επισκόπηση του πεδίου

Διεύρυνση των Ερωτημάτων

- εντάσσεται στην προσπάθεια βελτίωσης της διαδικασίας Ανάκτησης Πληροφοριών (Information Retrieval) εμπλουτίζοντας το αρχικό ερώτημα με όρους που θα αυξήσουν την απόδοση της ανάκτησης
- καλείται να αντιμετωπίσει:
  - το σημασιολογικό χάσμα (semantic gap) μεταξύ των όρων αναζήτησης και της αποτύπωσης του περιεχομένου
  - τις δυνατότητες των συστημάτων αναζήτησης
  - το χρήστη και τις συμπεριφορικές αλλαγές του
    - ♦καθώς και το συνδυασμό όλων αυτών

Μεθοδολογία (1)

Εργαστηριακό Πείραμα

- Σενάριο αναζητήσεων για εντοπισμό σχετικών τεκμηρίων (ελεγχόμενη πληροφοριακή ανάγκη)
  - Θέματα:
    - Αποδημητικά πουλιά
    - Οπωροφόρα δέντρα
    - Προστασία του περιβάλλοντος
    - Φαινόμενο του θερμοκηπίου
    - Εναλλακτικές μορφές ενέργειας
- Συμμετέχοντες
  - 48 φοιτητές του τμήματος Αρχειονομίας – Βιβλιοθηκονομίας του Ιονίου Πανεπιστημίου (προ- & μετα-πτυχιακοί)

Μεθοδολογία (2)

- Περιβάλλον αναζήτησης: Β/Δ Ευωνύμου Οικολογικής Βιβλιοθήκης
  - ~14.400 βιβλιογραφικές εγγραφές (προσαρμοσμένες & τροποποιημένες: π.χ. όχι λογοτεχνία, όχι αγγλικά)
  - τοπική πρόσβαση σε έναν z39.50 server
  - δυνατότητα αναζήτησης μόνο με Θέμα, όχι Boolean τελεστές, μόνο δεξιά αποκοπή
- Θησαυρός EUROVOC (v. 4.2, ελληνική εκδοχή): για την εννοιολογική συσχέτιση των όρων
  - διεπιστημονικός, χωρισμένος σε θεματικές ενότητες, η συγκεκριμένη έκδοση περιλάμβανε 6.645 έννοιες
    - ♦χρησιμοποιήθηκε η αντίστοιχη ενότητα για κάθε θέμα ώστε να περιοριστούν οι επιδράσεις της πολυσημίας των λέξεων

Ορισμοί και διευκρινίσεις (1)

Χαρακτηρισμοί Όρων:

- Ειδικότερος: πιο ειδικός όρος ή περισσότερες λέξεις
- Γενικότερος: πιο γενικός όρος ή λιγότερες λέξεις
- Συνώνυμος: όρος με ίδια ή περίπου ίδια σημασία
- Παράλληλος: τίποτε από τα παραπάνω. Στο ίδιο επίπεδο της ιεραρχίας (κατά κανόνα)
- Παρεχόμενος: δινόταν από το σενάριο
- Λανθασμένος: δεν υπάρχει στην ελληνική γλώσσα
- Ακατάλληλος: είτε καμία εννοιολογική συσχέτιση με το θέμα αναζήτησης είτε όρος στα αγγλικά

Ορισμοί και διευκρινίσεις (2)

- Στην παρούσα έρευνα ασχοληθήκαμε με συσχέτιση εννοιών. Έτσι, δε λάβαμε υπόψη:
  - ορθογραφικά λάθη και λάθη στον τονισμό
  - αριθμούς: ενικό – πληθυντικό
- Σύνθεση (formulation) και Ανασύνθεση (reformulation) ερωτημάτων: το πρώτο ερώτημα και τα επόμενα
- Χρήση του παρεχόμενου όρου: ιδιαίτερης σημασίας
  - Στο στάδιο της σύνθεσης: 41,6% των αναζητήσεων
  - Στο στάδιο της ανασύνθεσης: 5,3% των αναζητήσεων
- Απλό περιβάλλον αναζήτησης ώστε να μην απαιτείται η εφαρμογή ιδιαίτερων τεχνικών αναζήτησης από τους συμμετέχοντες

Στάδιο σύνθεσης ερωτήματος

Κατηγοριοποίηση* Όρων κατά τη Σύνθεση Ερωτημάτων (%)

Γενικότερος όρος	18,1
Ειδικότερος όρος	25,2
Παράλληλος όρος	8,4
Συνώνυμος όρος	5,9
Λανθασμένος όρος	0
Ακατάλληλος όρος	0,8
Παρεχόμενος όρος	41,6

***Ο χαρακτηρισμός έγινε με βάση τον παρεχόμενο όρο**

Στάδιο ανασύνθεσης ερωτήματος

Κατηγοριοποίηση* Όρων κατά την Ανασύνθεση Ερωτημάτων (%)

Γενικότερος όρος	20,0
Ειδικότερος όρος	20,3
Παράλληλος όρος	47,6
Συνώνυμος όρος	5,3
Λανθασμένος όρος	0,8
Ακατάλληλος όρος	6,0

***Ο χαρακτηρισμός έγινε με βάση τον προηγούμενο όρο**

Μοναδικοί όροι & Αριθμός λέξεων ανά όρο

- Συνολικός αριθμός όρων: 843
- Συνολικός αριθμός λέξεων: 1372
- Μοναδικές λέξεις: 205 (=15%)
  - Δε δικαιολογείται. Η θεματολογία δεν ήταν αλληλο-επικαλυπτόμενη σε τέτοιο βαθμό

Αριθμός λέξεων/ όρο	*%*
Όροι με 1 λέξη	57,7
Όροι με 2 λέξεις	33,6
Όροι με 3 λέξεις	8,8

Εννοιολογική συσχέτιση όρων με το θησαυρό

Πόσες από τις (μοναδικές) λέξεις των όρων αναζήτησης συσχετίζονται εννοιολογικά με λέξεις του Θησαυρού:

Από τις 205 συσχετίστηκαν οι 124 (60,5%)

- Διευκρίνιση: οι συσχετίσεις έγιναν εκ των υστέρων από την ερευνητική ομάδα. Οι χρήστες παρείχαν τους όρους κάνοντας
  - πολλαπλές αναζητήσεις για κάθε θέμα
  - χωρίς υποβοηθούμενη διεύρυνση των ερωτημάτων τους

- Προκύπτει ικανό ποσοστό εννοιολογικών συσχετίσεων
- Ισάριθμες ευκαιρίες σημείων εκκίνησης για εννοιολογική διεύρυνση ερωτημάτων

Επισήμανση

- Η έρευνά μας βασίστηκε σε ελληνικά δεδομένα, με υποβαλλόμενα ερωτήματα στα ελληνικά και αξιοποίηση της ελληνικής εκδοχής του εργαλείου συσχέτισης, δηλαδή του Θησαυρού.

  Αποφύγαμε, έτσι, τα συνήθη προβλήματα που αντιμετωπίζουν ανάλογες έρευνες όταν καλούνται να αποδώσουν έννοιες σε διαφορετικές γλώσσες, δηλαδή, αμφισβητήσιμες αποδόσεις όρων και, κατ' επέκταση και, συσχετίσεις εννοιών.

Συμπεράσματα (1)

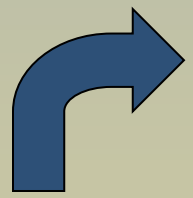
- Ο παρεχόμενος όρος επηρεάζει σημαντικά την επιλογή του πρώτου όρου αναζήτησης (σύνθεση)
- Οι χρήστες κυρίως επιλέγουν ειδικότερο ή γενικότερο όρο ⟶ ευνοείται η αξιοποίηση της ιεραρχίας του Θησαυρού για διεύρυνση των ερωτημάτων
- Η χρήση παράλληλου όρου θα πρέπει να ευνοείται αρχικά από το θέμα.
- Μικρό ποσοστό χρήσης συνώνυμων όρων ⟶ *η διεύρυνση ερωτημάτων με τη χρήση ενός ΣΟΓ για πρόταση συνωνύμων θα ήταν χρήσιμη
- Μικρό ποσοστό μοναδικών λέξεων ⟶ οι χρήστες μπορούν να γίνουν προβλέψιμοι (αν ευνοεί και το γνωστικό πεδίο ή και όχι...)

Συμπεράσματα (2)

- Στο 60,5% των περιπτώσεων μπορούμε να εφαρμόσουμε άμεση συσχέτιση των όρων αναζήτησης με όρους του Θησαυρού. Θεωρούμε ότι είναι μια καλή εκκίνηση για διεύρυνση των ερωτημάτων χωρίς επιπλέον βήματα. Ωστόσο...
- Χρήση, κυρίως, μίας λέξης για τη δόμηση των όρων αναζήτησης

  Στην περίπτωση που δεν μπορούμε να επιτύχουμε άμεση εννοιολογική συσχέτιση μεταξύ των όρων αναζήτησης και ενός ΣΟΓ (όπως στην περίπτωση πολυθεματικού περιεχομένου), θα μπορούσαμε να πετύχουμε καλύτερη απόδοση της εννοιολογικής συσχέτισης όρων αν οι χρήστες χρησιμοποιούσαν περισσότερες λέξεις για τη δόμηση των όρων αναζήτησης .

Ακολουθεί ένα παράδειγμα



Συμπεράσματα (2): παράδειγμα…

Όρος αναζήτησης: ***περιβάλλον***

Αποτέλεσμα λεξιλογικής συσχέτισης στον EUROVOC (19 έννοιες). *Ποια χρειαζόμαστε;*

οικογένεια	UF	οικογενειακό **περιβάλλον**
κοινωνική τάξη	UF	κοινωνικό **περιβάλλον**
οικολογικός τουρισμός	UF	τουρισμός που σέβεται το **περιβάλλον**
αγροτική κατοικία	UF	αγροτικό **περιβάλλον**
αστική κατοικία	UF	αστικό **περιβάλλον**
σχολικό **περιβάλλον**	UF	πανεπιστημιακό **περιβάλλον**
εργασιακό **περιβάλλον**	UF	επαγγελματικό **περιβάλλον**
επίπτωση στο **περιβάλλον**	UF	επίδραση στο **περιβάλλον**
περιβαλλοντική εκπαίδευση	UF	ευαισθητοποίηση στο **περιβάλλον**
περιβαλλοντικό πρότυπο	UF	πρότυπο σχετικό με το **περιβάλλον**
[…]		

Όρος αναζήτησης: ***προστασία του περιβάλλοντος***

Αποτέλεσμα λεξιλογικής συσχέτισης στον EUROVOC (2 έννοιες). *Ποια χρειαζόμαστε;*

οικονομικό μέσο για το περιβάλλον	UF	οικονομικό μέσο για την **προστασία του περιβάλλοντος**
προστασία του περιβάλλοντος		

Συμπεράσματα (3)

Αν και η χρήση περισσότερων όρων αναζήτησης κατά το στάδιο της αναζήτησης απαιτεί ιδιαίτερη προσοχή από την πλευρά του χρήστη, καθώς μπορεί να περιορίσει σημαντικά τον αριθμό ανακτημένων αποτελεσμάτων ή να μην επιστρέψει κανένα, κατά τη διαδικασία υποβοηθούμενης διεύρυνσης ερωτημάτων μπορεί να χρησιμεύει ιδιαίτερα στην περίπτωση:

- που μία λέξη έχει πολλές έννοιες (poly-representation), οπότε αν συσχετιζόταν αρχικά και με άλλες λέξεις του όρου αναζήτησης, τότε θα ήταν δυνατή η εξαγωγή ασφαλέστερων συμπερασμάτων ως προς το γνωστικό πεδίο στο οποίο ανήκει το ερώτημα για να προχωρήσουμε ασφαλέστερα στην εννοιολογική διεύρυνση του ερωτήματος
- λέξεων που δε συσχετίζονται αρχικά ούτε με απλή λεξιλογική συσχέτιση με το Θησαυρό παρέχοντας εναλλακτικό και ικανό σημείο εκκίνησης για εννοιολογική διεύρυνση του ερωτήματος

Μελλοντική έρευνα

- Ποσοτική επιβεβαίωση των αρχικών μας αποτελεσμάτων με περισσότερα ΣΟΓ (π.χ. οντολογίες, ταξινομίες, σημασιολογικά δίκτυα), χρήστες και ερωτήματα αυτοματοποιώντας τη διαδικασία συσχέτισης με ανάλογο εργαλείο για ευρύτερη εφαρμογή

- Διερεύνηση της δυνατότητας εφαρμογής μοντέλων για εννοιολογική συσταδοποίηση ερωτημάτων (conceptual query clustering)

Ευχαριστούμε

- τον κ. Σάκη Κουρουζίδη, διευθυντή της Ευωνύμου Οικολογικής Βιβλιοθήκης, καθώς και την κα Κατερίνα Τοράκη, για την άδεια αξιοποίησης των δεδομένων του καταλόγου της βιβλιοθήκης

- τους συμμετέχοντες σε όλα τα στάδια της έρευνάς μας

Ευχαριστώ για την προσοχή σας!

Άννα Μάστορα

Email επικοινωνίας: mastora[at]ionio[dot]gr

Εργαστήριο Ψηφιακών Βιβλιοθηκών & Ηλεκτρονικής Δημοσίευσης, Τμήμα Αρχειονομίας - Βιβλιοθηκονομίας, Ιόνιο Πανεπιστήμιο, Ιωάννου Θεοτόκη 72, 49100, Κέρκυρα