Ανάκτηση πληροφορίας

Ενότητα 3: Μοντελοποίηση: Boolean μοντέλο

Απόστολος Παπαδόπουλος
Τμήμα Πληροφορικής

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Το Boolean μοντέλο

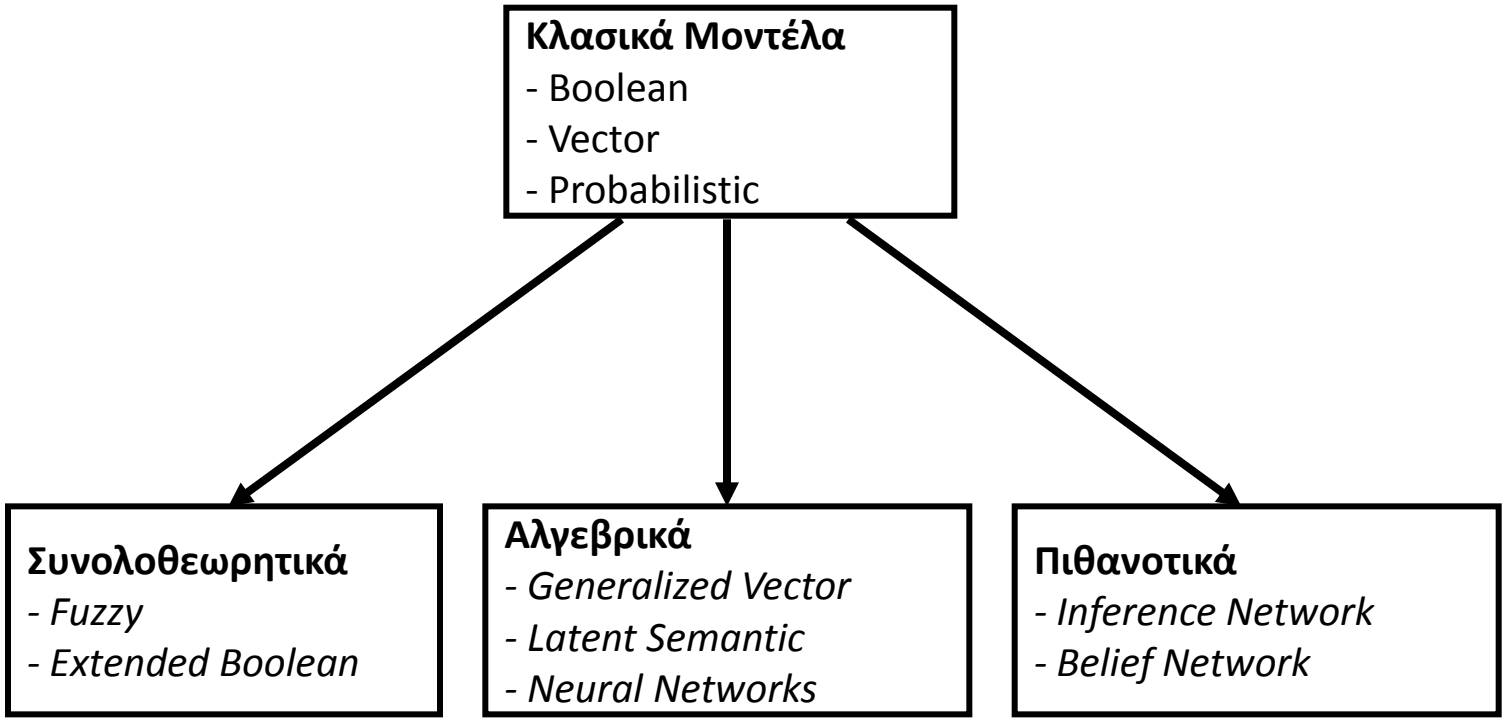
Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

1. Το Boolean μοντέλο

Το Boolean μοντέλο

Μοντέλα IR


Κλασικά Μοντέλα
- Boolean
- Vector
- Probabilistic
Συνολοθεωρητικά
- *Fuzzy*
- *Extended Boolean*
Αλγεβρικά
- *Generalized Vector*
- *Latent Semantic*
- *Neural Networks*
Πιθανοτικά
- *Inference Network*
- *Belief Network*

Χαρακτηριστικά Μοντέλων IR

Ένα μοντέλο IR χαρακτηρίζεται από:

- **D**, σύνολο λογικών όψεων κειμένων.
- **Q**, σύνολο λογικών όψεων ερωτημάτων.
- **F,** πλαίσιο μοντελοποίησης κειμένων, ερωτημάτων και συσχετισμών τους.
- **R(q,d),** συνάρτηση βαθμολόγησης.

Λέξεις Κλειδιά (Keywords)

- Χρησιμοποιούνται σαν αντιπρόσωποι όλου του κειμένου και βοηθούν στη σύντομη περιγραφή του κειμένου (περίληψη).
- Απαιτείται προσοχή στην επιλογή τους, έτσι ώστε τα κείμενα να διαχωρίζονται κατάλληλα.
- Το πλήθος των όρων είναι συνήθως μεγάλο και προηγείται απαλοιφή τετριμμένων λέξεων (π.χ., άρθρα, σύνδεσμοι κλπ).

Παράδειγμα

Κείμενο 1	Κείμενο 2	Κείμενο 3
… η γεωργική επανάσταση	… η βιομηχανική επανάσταση	… η επανάσταση υψηλής τεχνολογίας

Η επιλογή της λέξης *επανάσταση* σαν λέξη κλειδί για τα τρία κείμενα δημιουργεί πρόβλημα. Γιατί;

Παρατήρηση

- Όλες οι λέξεις κλειδιά (αλλιώς όροι) δεν έχουν την ίδια βαρύτητα για τις προτιμήσεις των χρηστών. Κάποιες λέξεις μπορεί να είναι σημαντικές ενώ κάποιες άλλες λιγότερο σημαντικές.
- Έστω ti ένας όρος και dj ένα έγγραφο. Το βάρος του όρου ti στο έγγραφο dj συμβολίζεται ως w(ti,dj) >= 0 (ή απλούστερα wij) και δηλώνει το πόσο σημαντικός είναι ο όρος ti σε σχέση με το έγγραφο dj.

Ορισμός

- Έστω m αριθμός των όρων και $T=\{t_1,\ldots,t_m\}$ το σύνολο των μοναδικών όρων. Εάν ο όρος t_i δεν εμφανίζεται στο έγγραφο d_j τότε $w(t_i,d_j)=0$. Διαφορετικά, $w(k_i,d_j) > 0$.
- Άρα σε κάθε κείμενο $\boldsymbol{d}_j$ αντιστοιχεί ένα m-διάστατο διάνυσμα βαρών $(w_{1,j}, w_{2,j}, \ldots, w_{m,j})$

Κλασικά Μοντέλα IR-1

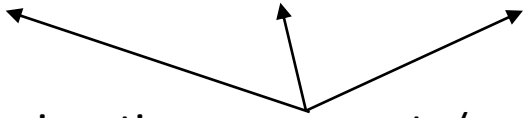
- Κάθε κείμενο αντιπροσωπεύεται από ένα σύνολο χαρακτηριστικών λέξεων (keywords).
- Ένα keyword είναι χρήσιμο για να θυμόμαστε το βασικό θέμα του κειμένου.
- Συνήθως τα keywords είναι ουσιαστικά, τα οποία από μόνα τους έχουν νόημα.
- Ωστόσο, οι μηχανές αναζήτησης θεωρούν ότι όλες οι λέξεις του κειμένου είναι keywords (full text representation).

Κλασικά Μοντέλα IR-2

- t_i ένας όρος (index term, keyword)
- d_j ένα έγγραφο
- m συνολικός αριθμός όρων
- $T = \{t_1, t_2, \ldots, t_m\}$ σύνολο keywords
- $w_{ij} >= 0$ βάρος μεταξύ t_i, d_j
- $w_{ij} = 0$ το t_i δε βρίσκεται στο έγγραφο d_j
- $vec(d_j) = (w_{1j}, w_{2j}, \ldots, w_{tj})$ διάνυσμα που σχετίζεται με το έγγραφο d_j
- $g_i(vec(d_j)) = w_{ij}$ συνάρτηση που επιστρέφει το βάρος που σχετίζεται με τα t_i και d_j

Boolean Μοντέλο-1

- Απλό, βασίζεται στη **Θεωρία Συνόλων.**
- Διατύπωση ερωτημάτων ως λογικές εκφράσεις
  - ακριβής σημαντική (exact semantics)
  - απλός φορμαλισμός.
- Ένας όρος είναι είτε παρόν είτε απών από το έγγραφο, επομένως $w_{ij} \in \{0,1\}$
- Για παράδειγμα
  - $q = (t_1 \vee t_2) \wedge t_3$
  - $qdnf = (1,1,1) \vee (0,1,1) \vee (1,0,1)$ (disjunctive normal form)



conjunctive components (qcc)

Boolean Μοντέλο-2

- Πίνακας αληθείας του ερωτήματος $(t1 \vee t2) \wedge t3$

t1	t2	t3	Διάνυσμα	Έκφραση	*απάντηση*
0	0	0	(0 , 0, 0)	$\neg t1 \wedge \neg t2 \wedge \neg t3$	0
0	0	1	(0, 0, 1)	$\neg t1 \wedge \neg t2 \wedge t3$	0
0	1	0	(0, 1, 0)	$\neg t1 \wedge t2 \wedge \neg t3$	0
0	1	1	(0, 1, 1)	$\neg t1 \wedge t2 \wedge t3$	1
1	0	0	(1, 0, 0)	$t1 \wedge \neg t2 \wedge \neg t3$	0
1	0	1	(1, 0, 1)	$t1 \wedge \neg\ t2 \wedge t3$	1
1	1	0	(1, 1, 0)	$t1 \wedge t2 \wedge \neg t3$	0
1	1	1	(1, 1, 1)	$t1 \wedge t2 \wedge t3$	1

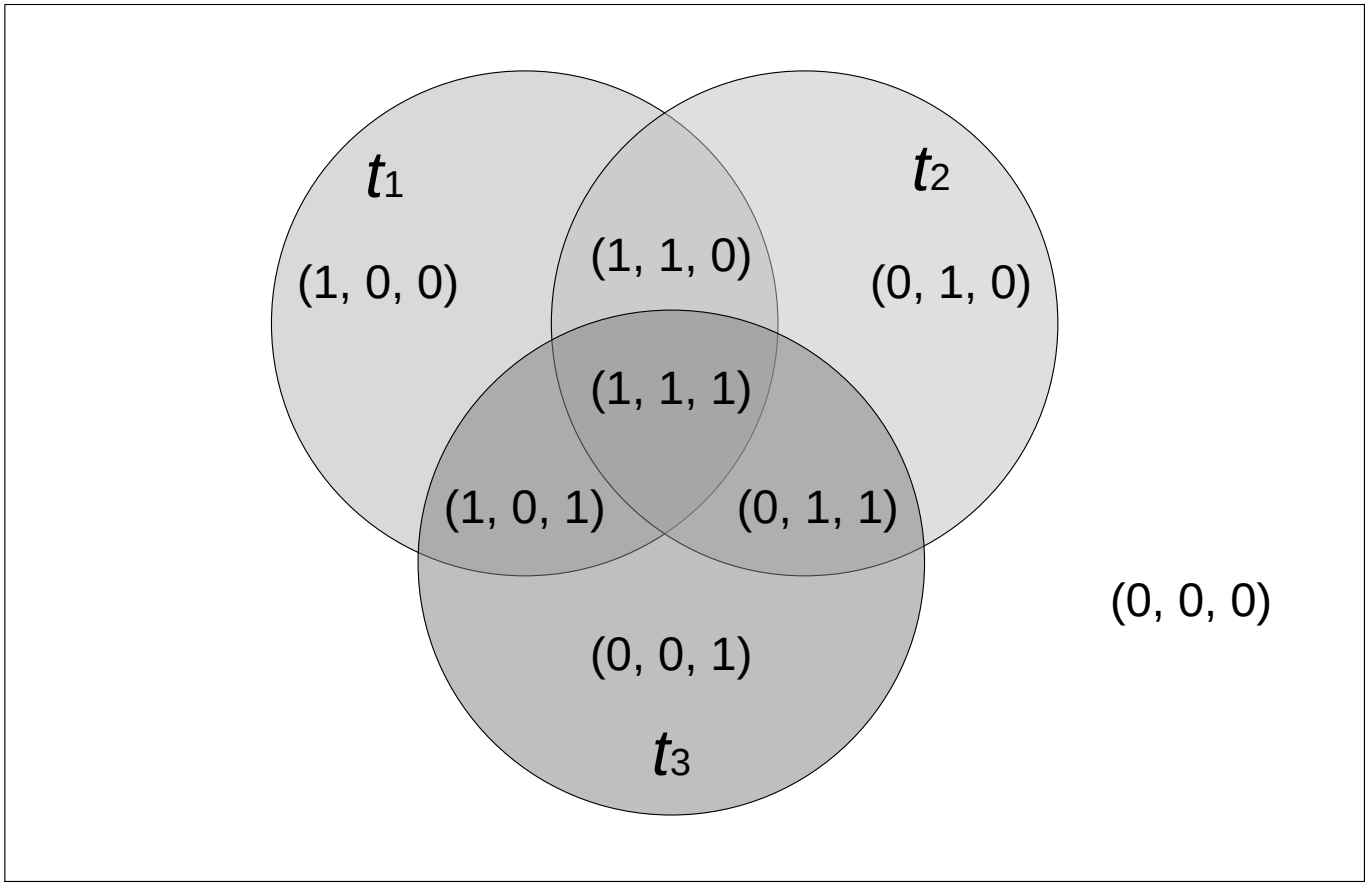
Boolean Μοντέλο-3

- Ομοιότητα στο Boolean μοντέλο
- $\text{Sim}(q,d_j) = 1$, αν $\exists\ \text{vec}(qcc) \in \text{vec}(qdnf) \mid \forall t_i, \text{gi}(\text{vec}(dj)) = \text{gi}(\text{vec}(qcc)$

  0, διαφορετικά

Boolean Μοντέλο-4

- $q = (t_1 \vee t_2) \wedge t_3$



Μειονεκτήματα Boolean Μοντέλου

- Δεν υπάρχει υποστήριξη για μερική ταύτιση (partial matching).
- Δεν υπάρχει βαθμολόγηση των αποτελεσμάτων.
- Η ερώτηση πρέπει να διατυπωθεί με λογική έκφραση, το οποίο δεν είναι πάντα εύκολο για όλες τις κατηγορίες χρηστών.
- Τα ερωτήματα που διατυπώνονται είναι τις περισσότερες φορές πολύ απλοϊκά.
- Το Boolean μοντέλο άλλοτε επιστρέφει πάρα πολλά έγγραφα (απλές λογικές εκφράσεις) και άλλοτε πάρα πολύ λίγα (πολύπλοκες λογικές εκφράσεις).

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά - Μη Εμπορική Χρήση - Όχι Παράγωγα Έργα 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».

Ο δικαιούχος μπορεί να παρέχει στον αδειοδόχο ξεχωριστή άδεια να χρησιμοποιεί το έργο για εμπορική χρήση, εφόσον αυτό του ζητηθεί.

Ως **Μη Εμπορική** ορίζεται η χρήση:

- που δεν περιλαμβάνει άμεσο ή έμμεσο οικονομικό όφελος από την χρήση του έργου, για το διανομέα του έργου και αδειοδόχο
- που δεν περιλαμβάνει οικονομική συναλλαγή ως προϋπόθεση για τη χρήση ή πρόσβαση στο έργο
- που δεν προσπορίζει στο διανομέα του έργου και αδειοδόχο έμμεσο οικονομικό όφελος (π.χ. διαφημίσεις) από την προβολή του έργου σε διαδικτυακό τόπο

[1] http://creativecommons.org/licenses/by-nc-nd/4.0/

Τέλος ενότητας

Επεξεργασία: <Μαυρίδης Απόστολος>
Θεσσαλονίκη, <Εαρινό εξάμηνο 2013-2014>

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.