ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Κεφάλαιο 11: Πιθανοτική ανάκτηση πληροφορίας.

Πιθανοτική Ανάκτηση Πληροφορίας

- Βασική ιδέα: Διάταξη εγγράφων με βάση *την πιθανότητα να είναι συναφή με το ερώτημα*
- Τυχαία μεταβλητή R
- Διάταξη με βάση το

$$P(R=1|q, d)$$

Αρχή Πιθανοτικής Κατάταξης

«Αν η απόκριση ενός συστήματος ανάκτησης πληροφορίας σε κάθε ερώτημα είναι η *κατάταξη των εγγράφων κατά φθίνουσα πιθανότητα συνάφειας* προς τον χρήστη που υπέβαλε το ερώτημα, όπου οι πιθανότητες εκτιμώνται με τον ακριβέστερο δυνατό τρόπο βάσει των όποιων δεδομένων είναι διαθέσιμα στο σύστημα για τον σκοπό αυτό, η συλλογική *αποτελεσματικότητα είναι η βέλτιστη εφικτή* με βάση αυτά τα δεδομένα» [*van Rijsbergen* 1979]

Probability Ranking Principle (RPR):

decreasing probability of relevance => best overall effectiveness

Δυαδικό Μοντέλο Ανεξαρτησίας (Binary Independence Model (BIP))

Θα αναπαραστήσουμε κάθε έγγραφο *d* (και αντίστοιχα, το ερώτημα) ως ένα δυαδικό (Boolean) διάνυσμα *(M: #όρων)*

$$\vec{x} = (x_1, \ldots, x_M)$$

$x_i = 1$ ανν το έγγραφο περιέχει τον όρο i

Υπόθεση ανεξαρτησίας όρων (independence assumption): Οι όροι εμφανίζονται στα έγγραφα ανεξάρτητα ο ένας από τον άλλο

Επίσης, η συνάφεια ενός εγγράφου είναι ανεξάρτητη από τη συνάφεια των άλλων εγγράφων *(ανεξαρτησία εγγράφων)*

Δυαδικό Μοντέλο Ανεξαρτησίας

Επειδή μας ενδιαφέρει *η διάταξη* θα χρησιμοποιήσουμε τη σχετική πιθανότητα - odds

$$O(R \mid q, \vec{x}) = \frac{p(R=1 \mid q, \vec{x})}{p(R=0 \mid q, \vec{x})}$$

Από τον κανόνα του Bayes

$$O(R \mid q, \vec{x}) = \frac{p(R=1 \mid q, \vec{x})}{p(R=0 \mid q, \vec{x})} = \frac{\dfrac{p(R=1 \mid q)\, p(\vec{x} \mid R=1, q)}{p(\vec{x} \mid q)}}{\dfrac{p(R=0 \mid q)\, p(\vec{x} \mid R=0, q)}{p(\vec{x} \mid q)}}$$

Δυαδικό Μοντέλο Ανεξαρτησίας

$$O(R \mid q, \vec{x}) = \frac{p(R=1 \mid q, \vec{x})}{p(R=0 \mid q, \vec{x})} = \frac{p(R=1 \mid q)}{p(R=0 \mid q)} \frac{p(\vec{x} \mid R=1, q)}{p(\vec{x} \mid R=0, q)}$$

Σταθερά για κάθε ερώτημα

Πρέπει να εκτιμήσουμε αυτήν την ποσότητα

Εκ των προτέρων πιθανότητα (prior probability) να ανακτήσουμε συναφές (μη συναφές) ερώτημα

Αν έχουμε ανακτήσει ένα συναφές (μη συναφές) έγγραφο, αυτό να έχει αναπαράσταση $\vec{x}$

Δυαδικό Μοντέλο Ανεξαρτησίας

Θα χρησιμοποιήσουμε την *υπόθεση ανεξαρτησίας των όρων:* η παρουσία ή απουσία ενός όρου σε ένα έγγραφο είναι ανεξάρτητη από την παρουσία ή απουσία οποιουδήποτε άλλου όρου – Naïve Bayes conditional independent assumption)

$$\frac{p(\vec{x} \mid R=1,q)}{p(\vec{x} \mid R=0,q)} = \prod_{i=1}^{M} \frac{p(x_i \mid R=1,q)}{p(x_i \mid R=0,q)}$$

Άρα:

$$O(R \mid q,\vec{x}) = O(R \mid q) \prod_{i=1}^{M} \frac{p(x_i \mid R=1,q)}{p(x_i \mid R=0,q)}$$

Δυαδικό Μοντέλο Ανεξαρτησίας

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{i=1}^{n} \frac{p(x_i \mid R=1, q)}{p(x_i \mid R=0, q)}$$

Αφού τα x_i είναι είτε 1 είτε 0, χωρίζουμε το γινόμενο

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{x_i=1} \frac{p(x_i=1 \mid R=1, q)}{p(x_i=1 \mid R=0, q)} \prod_{x_i=0} \frac{p(x_i=0 \mid R=1, q)}{p(x_i=0 \mid R=0, q)}$$

Δυαδικό Μοντέλο Ανεξαρτησίας

Για τον όρο *i*, ορίζουμε ως p_i την πιθανότητα να εμφανίζεται σε **συναφές** έγγραφο και ως r_i την πιθανότητα να εμφανίζεται σε **μη συναφές** έγγραφο

$$p_i = p(x_i = 1 | R = 1, q)$$

$$r_i = p(x_i = 1 | R = 0, q)$$

	Έγγραφο	Συναφές (R=1)	Μη συναφές (R=0)
Όρος i εμφανίζεται	$x_i = 1$	p_i	r_i
Όρος i δεν εμφανίζεται	$x_i = 0$	$(1 - p_i)$	$(1 - r_i)$

Δυαδικό Μοντέλο Ανεξαρτησίας

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{x_i = 1} \frac{p(x_i = 1 \mid R = 1, q)}{p(x_i = 1 \mid R = 0, q)} \prod_{x_i = 0} \frac{p(x_i = 0 \mid R = 1, q)}{p(x_i = 0 \mid R = 0, q)}$$

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{x_i = 1} \frac{p_i}{r_i} \prod_{x_i = 0} \frac{(1 - p_i)}{(1 - r_i)}$$

Δυαδικό Μοντέλο Ανεξαρτησίας

Αν υποθέσουμε ότι οι όροι που δεν εμφανίζονται στο ερώτημα είναι το ίδιο πιθανό να εμφανίζονται σε συναφή και σε μη συναφή έγγραφα

$$p_i = r_i$$

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{\substack{x_i=1 \\ q_i=1}} \frac{p_i}{r_i} \prod_{\substack{x_i=0 \\ q_i=1}} \frac{(1-p_i)}{(1-r_i)}$$

Δυαδικό Μοντέλο Ανεξαρτησίας

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{\substack{x_i = q_i = 1}} \frac{p_i}{r_i} \prod_{\substack{x_i = 0 \\ q_i = 1}} \frac{1 - p_i}{1 - r_i}$$

Οι όροι του ερωτήματος που εμφανίζονται

Οι όροι του ερωτήματος που δεν εμφανίζονται

Μοιράζουμε το μεσαίο όρο (=1) στα αριστερά και δεξιά (είναι οι όροι του ερωτήματος που εμφανίζονται στο έγγραφο)

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{\substack{x_i = 1 \\ q_i = 1}} \frac{p_i}{r_i} \prod_{\substack{x_i = 1 \\ q_i = 1}} \left(\frac{1 - r_i}{1 - p_i} \cdot \frac{1 - p_i}{1 - r_i} \right) \prod_{\substack{x_i = 0 \\ q_i = 1}} \frac{1 - p_i}{1 - r_i}$$

$$O(R \mid q, \vec{x}) = O(R \mid q) \prod_{\substack{x_i = q_i = 1}} \frac{p_i (1 - r_i)}{r_i (1 - p_i)} \prod_{\substack{q_i = 1}} \frac{1 - p_i}{1 - r_i}$$

Οι όροι του ερωτήματος που εμφανίζονται στο έγγραφο

Όλοι οι όροι

Δυαδικό Μοντέλο Ανεξαρτησίας

$$O(R \mid q, \vec{x}) = O(R \mid q) \cdot \prod_{x_i = q_i = 1} \frac{p_i (1 - r_i)}{r_i (1 - p_i)} \cdot \prod_{q_i = 1} \frac{1 - p_i}{1 - r_i}$$

Σταθερά για κάθε ερώτημα

Η μοναδική ποσότητα που πρέπει να εκτιμηθεί

Θα χρησιμοποιήσουμε το log αυτής της ποσότητας
Retrieval Status Value (RSV) (τιμή κατάστασης ανάκτησης)

$$RSV = \log \prod_{x_i = q_i = 1} \frac{p_i (1 - r_i)}{r_i (1 - p_i)} = \sum_{x_i = q_i = 1} \log \frac{p_i (1 - r_i)}{r_i (1 - p_i)}$$

Τιμή Κατάστασης Ανάκτησης (RSV)

$$RSV = \log \prod_{x_i = q_i = 1} \frac{p_i(1-r_i)}{r_i(1-p_i)} = \sum_{x_i = q_i = 1} \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

$$RSV = \sum_{x_i = q_i = 1} c_i \qquad c_i = \log \frac{p_i(1-r_i)}{r_i(1-p_i)} = \boxed{\log \frac{p_i}{1-p_i} + \log \frac{1-r_i}{r_i}}$$

Τιμή Κατάστασης Ανάκτησης (RSV)

$$RSV = \sum_{x_i = q_i = 1} c_i \qquad c_i = \log \frac{p_i}{1 - p_i} + \log \frac{1 - r_i}{r_i}$$

	Έγγραφο	Συναφές (R=1)	Μη συναφές (R=0)
Όρος i εμφανίζεται	$x_i = 1$	p_i	r_i
Όρος i δεν εμφανίζεται	$x_i = 0$	$(1 - p_i)$	$(1 - r_i)$

Τα c_i (log odd ratios) έχουν το ρόλο των βαρών σε αυτό το μοντέλο – αθροιστικά για κάθε όρο της ερώτησης

Για έναν όρο:

- 0 αν equal odds - η πιθανότητα να εμφανίζεται και να μην εμφανίζεται σε συναφές (μη συναφές) έγγραφο είναι ½
- + αν $p_i > 1\text{-}p_i$ (εμφανίζεται σε περισσότερα συναφή) και $r_i < 1 - r_i$ (εμφανίζεται σε λιγότερα μη συναφή)

Πως θα τα υπολογίσουμε;

Θεωρητική Εκτίμηση των RSV συντελεστών

Για κάθε όρο *i*, κατασκευάζουμε έναν πίνακα μετρητών
Έστω *N* έγγραφα στη συλλογή, *S* συναφή, και ο όρος εμφανίζεται σε συνολικά σε *n* έγγραφα από τα οποία τα συναφή είναι *s*

Έγγραφα	Συναφή (R=1)	Μη συναφή (R=0)	Συνολικά
$x_i = 1$	s	n - s	n
$x_i = 0$	S - s	N – n – (S – s)	N - n
	S	N - S	N

Εκτιμήσεις:

$$p_i \approx \frac{s}{S} \qquad r_i \approx \frac{(n-s)}{(N-S)}$$

$$c_i \approx K(N,n,S,s) = \log \frac{s/(S-s)}{(n-s)/(N-n-S+s)}$$

Αγνοεί τους όρους που δεν εμφανίζονται

Εκτίμηση των RSV συντελεστών (στην πράξη)

Αν τώρα υποθέσουμε ότι τα μη συναφή έγγραφα είναι περίπου όσα όλη η συλλογή, τότε το r_i (η πιθανότητα εμφάνισης του όρου σε μη συναφές έγγραφο) είναι n/N

$$\log \frac{1-r_i}{r_i} = \log \frac{N-n-S+s}{n-s} \approx \log \frac{N-n}{n} \approx \log \frac{N}{n} = IDF!$$

Εκτίμηση των RSV συντελεστών (στην πράξη)

- Το p_i (πιθανότητά εμφάνισης σε συναφή έγγραφα) είναι πιο δύσκολο να εκτιμηθεί
- Πιθανοί τρόποι εκτίμησης του p_i:
  - Από συναφή έγγραφα, αν κάποια από αυτά είναι γνωστά
  - Με μία σταθερά – τότε απλώς χρησιμοποιούμε το idf (with p_i=0.5)

$$RSV = \sum_{x_i = q_i = 1} \log \frac{N}{n_i}$$

  - Ανάλογο των εμφανίσεων του όρου στη συλλογή

έχει προταθεί: 1/3 + 2/3 df_i/N

Πιθανοτική Ανάδραση Συνάφειας (Relevance Feedback)

1. Υποθέτουμε κάποιες *αρχικές τιμές για τα* p_i και r_i - τις οποίες χρησιμοποιούμε για να ανακτήσουμε ένα *αρχικό σύνολο* συναφών εγγράφων (εγγράφων με R=1)
2. Αλληλοεπιδρούμε με το χρήστη για να βελτιώσουμε αυτές τις τιμές: οι χρήστες χαρακτηρίζουν ένα σύνολο έγγραφων *V* ως συναφή (R = 1) και μη συναφή (R = 0)
3. Επανα-υπολογίζουμε τα p_i και r_i
   - Ή τα συνδυάζουμε με τις αρχικές μας εκτιμήσεις χρήση Bayesian prior):

$$p_i^{(2)} = \frac{|V_i| + \kappa p_i^{(1)}}{|V| + \kappa}$$

κ: prior weight

4. Επαναλαμβάνουμε, έως σύγκλιση

PRP και BIM

(+) Λογικές προσεγγίσεις για τις πιθανότητες

(-) Περιοριστικές υποθέσεις:

- Ανεξαρτησία όρων
- Ανεξαρτησία συνάφειας εγγράφων
- Οι όροι που δεν εμφανίζονται στο ερώτημα δεν επηρεάζουν το αποτέλεσμα
- Δυαδική αναπαράσταση (αγνοούμε tf, μήκος εγγράφου, κλπ)

Στάθμιση Okapi BM25

- BM25 “Best Match 25”
- Αναπτύχθηκε στα πλαίσια του συστήματος Okapi
- Πιθανοτικό μη δυαδικό μοντέλο που λαμβάνει υπόψη συχνότητες όρων και μήκη εγγράφων

Αρχικές Εκδοχές του BM25

Εκδοχή 1:

$$c_i^{BM25v1}(tf_i) = c_i^{BIM}\frac{tf_i}{k_1 + tf_i}$$

Εκδοχή 2 (απλοποίηση με IDF):

$$c_i^{BM25v2}(tf_i) = \log\frac{N}{df_i} \times \frac{(k_1+1)tf_i}{k_1 + tf_i}$$

k_1 tuning παράμετρος, = 0 no tf

Στάθμιση BM25 με Κανονικοποίηση με το Μήκους Εγγράφου

Μεγαλύτερα έγγραφα πιο πιθανόν να έχουν μεγάλες τιμές tf_i

Μεγάλα έγγραφα

- πλεονασμός (verbosity) το tf_i που παρατηρούμε είναι πού μεγάλο
- γενικότερου σκοπού – το tf_i που παρατηρούμε μπορεί να είναι ακριβές

Πρέπει να σταθμίσουμε αυτά τα δύο

Στάθμιση BM25 με Κανονικοποίηση με το Μήκους Εγγράφου

- Μήκος εγγράφου

$$dl = \sum_{i \in V} tf_i \qquad B = \left((1-b) + b \frac{dl}{avdl} \right)$$

- *avdl*: μέσο μήκος εγγράφου στη συλλογή
- b - Συντελεστής κανονικοποίησης μήκους

Κανονικοποίηση

$$tf_i' = \frac{tf_i}{B}$$

Okapi BM25

$$RSV^{BM\,25} = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1) tf_i}{k_1((1-b) + b\frac{dl}{avdl}) + tf_i}$$

- k_1 ελέγχει τη διαβάθμιση της συχνότητα όρου (term frequency scaling)
  - k_1 = 0 δυαδικό μοντέλο; k_1 = μεγάλο – μετράει η καθαρή συχνότητα
- b ελέγχει την κανονικοποίηση του μήκους εγγράφου
  - b = 0 μη κανονικοποίηση; b = 1 είναι η σχετική συχνότητα (πλήρης κανονικοποίηση)
- Συνήθως, *k_1 στο 1.2–2* και το *b γύρω στο 0.75*
- Υπάρχουν και εκδοχές του BM25 που συμπεριλαμβάνουν και βάρη στους όρους του ερωτήματος καθώς και ανάδραση συνάφειας.

ΤΕΛΟΣ (τμήματος) Κεφαλαίου 11

Ερωτήσεις?

Χρησιμοποιήθηκε κάποιο υλικό των:

- ✓ *Pandu Nayak and Prabhakar Raghavan, CS276:Information Retrieval and Web Search (Stanford)*
- ✓ *Hinrich Schütze and Christina Lioma, Stuttgart IIR class*