2020

Learning to rank

ΜΕ ΕΜΦΑΣΗ ΣΤΗΝ ΠΡΟΣΕΓΓΙΣΗ ΚΑΤΑ ΣΗΜΕΙΑ

ΕΥΑ ΝΤΟΥΡΟΥ (Α.Μ. 235854)

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ – ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Υ/Η ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Περιεχόμενα

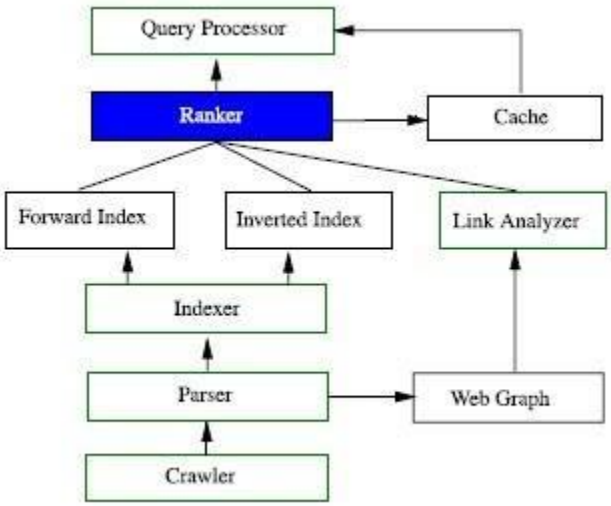
ΕΙΣΑΓΩΓΗ

Μπορεί κανείς να φανταστεί το Διαδίκτυο σαν μια βιβλιοθήκη, η οποία κάθε στιγμή επεκτείνεται, μαζεύει περισσότερα βιβλία, περιοδικά, έγγραφα αλλα και δέχεται ταυτόχρονα ολοένα και περισσότερες επισκέψεις. Μια τέτοια βιβλιοθήκη, προφανώς, θα χρειαζόταν όχι μόνο ένα άψογο σύστημα ταξινόμησης και οργάνωσης των περιεχομένων της, αλλά και ικανούς βιβλιοθηκάριους, οι οποίοι θα μπορούσαν να αναλύουν τις ανάγκες και τις απαιτήσεις των επισκεπτών (χρηστών) της βιβλιοθήκης, να βρίσκουν τα βιβλία/κείμενα που εξυπηρετούν τις ανάγκες αυτές και να τους τα προτείνουν. Παράλληλα, για να εξυπηρετηθούν όλοι, θα πρέπει αυτά να γίνονται αποδοτικά και όσο το δυνατόν γρηγορότερα.

Αναλογικά, η αναζήτηση στο Διαδίκτυο γίνεται με ένα σύστημα δεικτοδότησης και βαθμολόγησης των ιστοσελίδων, έτσι ώστε η ερώτηση (query) του κάθε χρήστη να εξυπηρετείται, με αποτελέσματα σχετικά με την ερώτηση και χρήσιμα. Αντιστοιχίζουμε, δηλαδή το σύστημα οργάνωσης της παραπάνω φανταστικής βιβλιοθήκης με το σύστημα βαθμολόγησης/δεικτοδότησης κειμένων και τους (ακούραστους) βιβλιοθηκάριους με τις **μηχανές αναζήτησης** που αναλαμβάνουν να ανακτήσουν την πληροφορία που ζητάει ο χρήστης.

Στην παρούσα εργασία εξετάζεται η σχέση του **Machine Learning** με τη βαθμολόγηση των αποτελεσμάτων αναζήτησης και το πώς αυτό χρησιμοποιείται στον τομέα της Εκμάθησης Βαθμολόγησης (**Learning to Rank**) . Για καλύτερη κατανόηση, παρουσιάζεται συνοπτικά και ο τρόπος λειτουργίας μηχανών αναζήτησης μέσω χαρακτηριστικών μοντέλων βαθμολόγησης σελίδων και κειμένων. Τέλος, εμβαθύνουμε στην προσέγγιση κατά σημεία(**Pointwise Approach**), παρουσιάζοντας συνοπτικά χαρακτηριστικούς αλγορίθμους και τις αντίστοιχες αποδόσεις τους.

Μηχανές Αναζήτησης



Σχήμα 1

Πώς λειτουργεί, όμως, μια μηχανή αναζήτησης; Το παραπάνω σχήμα αποτελεί μια περιγραφή των δομικών στοιχείων που συναποτελούν μια μηχανή αναζήτησης. Για να φτάσει μια σελίδα στο διαδίκτυο στα χέρια (ή πιο σωστά στα μάτια) του χρήστη, περνάει, κατα μία έννοια από όλα τα παραπάνω στάδια.

- Ο **crawler**, αρχικά, συλλέγει ιστοσελίδες και λοιπά έγγραφα από το Δίκτυο, σύμφωνα με κάποιες **στρατηγικές προτεραιότητας.**
- Ο **parser** τα αναλύει και παράγει όρους δεικτοδότησης (**index terms**)[1] και γράφημα υπερσύνδεσης (**hyperlink graph**)[2] για καθένα από αυτά.

---

[1] **Index term=** είναι ένας όρος ο οποίος συνοψίζει το κύριο νόημα του κειμένου. Χρησιμοποιείται ως λέξη-κλειδί για τη δεικτοδότηση και ανάκτηση του εκάστοτε κειμένου(π.χ. Από μια μηχανή αναζήτησης).

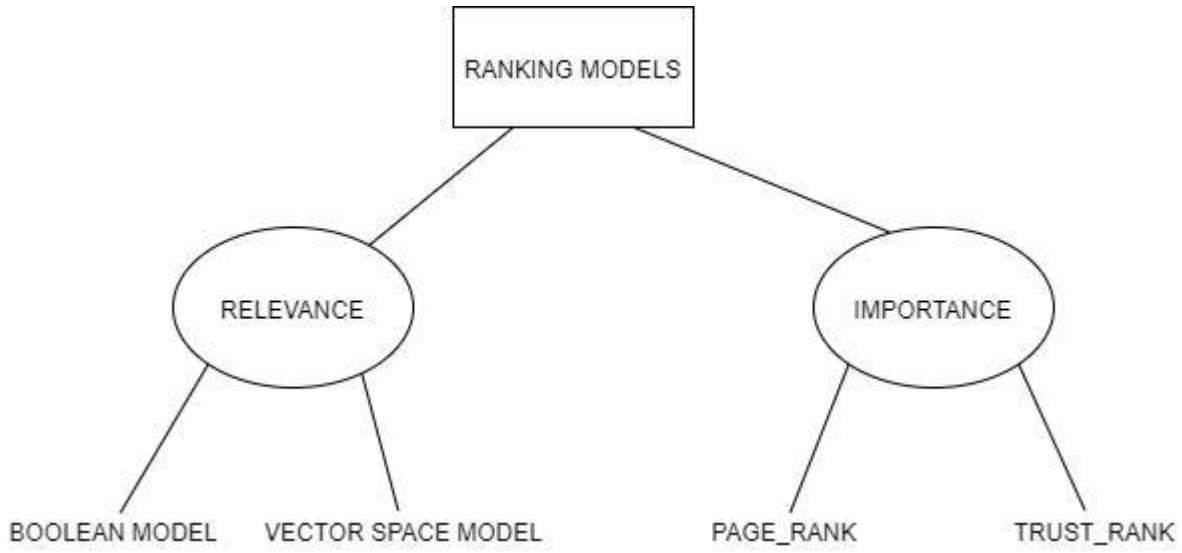
[2] **Hyperlink graph=** ένα hyperlink (ή και απλά link) αναφέρεται σε δεδομένα στα οποία μπορεί να βρεθεί ο χρήστης με ένα κλικ. Μπορεί να "δείχνει" σε ολόκληρο κείμενο ή ένα κομμάτι κειμένου. Ένα γράφημα υπερσυνδέσεων χρησιμοποιείται από τις μηχανές αναζήτησης για να ξεχωρίσουν σελίδες "υψηλής ποιότητας" (χρήσιμες) και παράλληλα να αφαιρούν από το γράφημα σελίδες spam. Χρησιμοποιούνται, λοιπόν για το ranking των links

- Ο **indexer** παίρνει την έξοδο του parser και παράγει τους δείκτες ή γενικότερα τις δομές δεδομένων που θα του επιτρέψουν γρήγορη αναζήτηση των κειμένων.
- Ο **link analyzer** παίρνει ως είσοδο γράφημα Δικτύου (Web graph) και καθορίζει τη σημασία κάθε σελίδας. Η σημασία κάθε σελίδας χρησιμοποιείται για τη βαθμολόγηση της.
- Ο **query processor** αποτελεί την διεπαφή μεταξύ χρήστη και μηχανής αναζήτησης. Πραγματοποιεί και την επεξεργασία των ερωτημάτων (π.χ. αφαίρεση stop words) και παράγει όρους δεικτοδότησης τους οποίους να καταλαβαίνει μια μηχανή αναζήτησης.
- Ο **ranker** αναλαμβάνει να ταιριάξει queries με κείμενα, παίρνοντας απευθείας και τα queries και τα κείμενα ως είσοδο και υπολογίζοντας ένα matching score με βάση ορισμένους τύπους.

Σημειώνεται, πως ο ranker αποτελεί ίσως το κυριότερο κομμάτι μιας μηχανής αναζήτησης, και γι' αυτό και υπάρχει τεράστιο ενδιαφέρον στην έρευνα και την ανάπτυξη τεχνολογιών **βαθμολόγησης σελίδων**. Αποτελεί όμως και πρόβλημα για διάφορες εφαρμογές της ανάκτησης πληροφοριών, όπως π.χ. το πρόβλημα ανάκτησης κειμένων, το οποίο προσεγγίζεται με χρήση ευρετικών μοντέλων βαθμολόγησης και πιο πρόσφατα και μέσω χρήσης μηχανικής μάθησης για την κατασκευή αποτελεσματικών μοντέλων βαθμολόγησης.

Μοντέλα Βαθμολόγησης Σελίδων

Από την πληθώρα μοντέλων βαθμολόγησης σελίδων στην επιστήμη της ανάκτησης πληροφορίας, διακρίνουμε δύο κύριες κατηγορίες: τα **μοντέλα βαθμολόγησης βάσει σχετικότητας** και τα **μοντέλα βαθμολόγησης βάσει σημαντικότητας**.



Εικόνα 1 Διάγραμμα Ειδών Μοντέλων Βαθμολόγησης

Βαθμολόγηση Βάσει Σχετικότητας (Relevance Ranking Models)

Τα μοντέλα αυτής της κατηγορίας παράγουν μια λίστα κειμένων βαθμολογημένων σύμφωνα με το κατα πόσο σχετίζονται με το ερώτημα του χρήστη. Για κάθε ένα κείμενο στην είσοδο υπολογίζεται ένα σκορ, βάσει της σχετικότητάς του κειμένου αυτού με την ερώτηση του χρήστη. Η τελική κατάταξη προκύπτει από την διάταξη των κειμένων με γνώμονα το σκορ τους.

Παρακάτω γίνεται αναφορά σε δύο χαρακτηριστικά παραδείγματα αυτού του είδους μοντέλου:

- **Boolean Model:** χαρακτηριστικό παράδειγμα πρωταρχικού μοντέλου αυτής της κατηγορίας. Στα πλαίσια του Boolean Μοντέλου, η σχετικότητα είναι δυαδική: το κείμενο είτε θα περιέχει τον όρο της αναζήτησης και επομένως θα ανακτηθεί, είτε δεν θα τον περιέχει και δε θα ανακτηθεί. Προφανώς, με τον τρόπο αυτό γίνεται πρόβλεψη του εαν το κείμενο είναι σχετικό ή όχι, δεν γίνεται, όμως πρόβλεψη του βαθμού σχετικότητας.Αυτό συμβαίνει γιατί όλα τα κείμενα θεωρούνται ισάξια όσον αφορά τη σχετικότητά τους με τον όρο της αναζήτησης.
- **Vector Space Model:** στα πλαίσια αυτού του μοντέλου, τα κείμενα και οι ερωτήσεις αναπαρίστανται ως διανύσματα μεγέθους n (όπου n ο αριθμός των index terms στην αναζήτησή μας). Ένα κείμενο, δηλαδή έχει τη μορφή $D_i = (d_{i1}, d_{i2}, \ldots., d_{in})$ όπου κάθε d_{ij} το βάρος του όρου j στο κείμενο i . Οι ομοιότητες και η σχετικότητα υπολογίζεται με χρήση του εσωτερικού γινόμενου δύο τέτοιων διανυσμάτων. Χαρακτηριστικά και πολύ δημοφιλή παραδείγματα του μοντέλου αυτού αποτελούν τα TF-IDF και BM25.s

Βαθμολόγηση Βάσει Σημαντικότητας(Importance Ranking Models)

Τα μοντέλα αυτής της κατηγορίας βαθμολογούν τα κείμενα με βάση το πόσο σημαντικά είναι αυτά. Αυτομάτως, βέβαια, δημιουργείται το ερώτημα του τί ακριβώς είναι η σημαντικότητα και πώς την ορίζουμε. Για να απαντήσουμε στο ερώτημα αυτό και να διερευνήσουμε την κατηγορία μοντέλων , μελετούμε ένα πολύ χαρακτηριστικό παράδειγμά της: το μοντέλο PageRank. Αναφορικά, το PageRank είναι και ο αλγόριθμος που χρησιμοποιείται από την πλέον διαδεδομένη μηχανή αναζήτησης Google, επομένως η χρησιμότητα και η σημαντικότητά του είναι δεδομένες.

PageRank : Χρησιμοποιεί τη δομή υπερσυνδέσεων του Διαδικτύου και την πιθανότητα ένας χρήστης-surfer που «κλικάρει» σε τυχαία links να φτάσει στη σελίδα που βαθμολογείται. Γενικά, χρησιμοποιείται ο τύπος :

$$PR(d_u) = \sum_{d_v \in B_u} \frac{PR(d_v)}{U(d_v)}$$

- Αυτό που είναι σημαντικό να παρατηρήσουμε στην παραπάνω σχέση είναι πώς ο βαθμός κάθε σελίδας d_u εξαρτάται από το βαθμό d_v κάθε σελίδας του B_u όπου B_u το σύνολο σελίδων που περιλαμβάνουν links προς τη σελίδα d_u, πρός το σύνολο σελίδων U, οι οποίες είναι αυτές που περιλαμβάνουν links από τη σελίδα d_v προς άλλες του διαδικτύου. Σημειώνεται πως, σε αυτή την παράγραφο, η χρήση του όρου «βαθμός σελίδας» αναφέρεται συγκεκριμένα στον PageRank βαθμό της σελίδας. Είναι σημαντικό, επίσης να λάβουμε υπ'όψην πως ένας χρήστης δεν ακολουθεί πάντα hyperlinks πάνω στο γράφημα, δηλαδή δεν «περπατά» αποκλειστικά το γράφημα, αλλά πολλές φορές φτάνει σε κάποιο κόμβο (σελίδα) του τυχαία. Γι' αυτό και ο παραπάνω τύπος μπορεί να επεκταθεί για να περιλαμβάνει αυτή την περίπτωση. Σε κάθε περίπτωση, ο αλγόριθμος PageRank επιδέχεται πολλών ειδών τροποποιήσεων και βελτιώσεων, ανάλογα πάντα με το είδος της εφαρμογής για την οποία χρησιμοποιείται.

TrustRank : Εισήχθη το 2004 από τους Zoltan Gyongyi ,Hector Garcia-Molina και Jan Pederse σε paper με τίτλο **«Combating Web Spam with TrustRank**» και όπως μαρτυρά και ο τίτλος της έκθεσής τους, έχει ως σκοπό την καταπολέμηση του SPAM. Ο αλγόριθμος αυτός ταξινομεί τις σελίδες λαμβάνοντας υπ' όψην την αξιοπιστία μιας ιστοσελίδας, όταν αποφασίζει για τη σημαντικότητά της. Είναι αποτελεσματικός, αλλά χρειάζεται ανθρώπινη «βοήθεια», αφού αρχικά πρέπει ένα σύνολο σελιδών να οριστούν ως σελίδες-πηγές. Στη συνέχεια η αξιοπιστία των σελίδων αυτών διαδίδεται στις σελίδες του γραφήματος συνδέσεων.

Learning To Rank

Ορισμός – Τί είναι το «Learning to Rank»

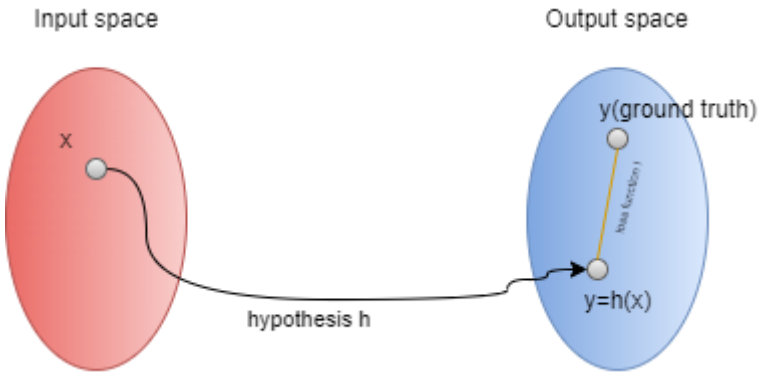
Στα μοντέλα βαθμολόγησης που αναφέρθηκαν εισάγονται συνήθως παράμετροι, οι οποίες μπορεί να καθορίζουν το βάρος ενός στοιχείου της βαθμολόγησης ή να εξομαλύνουν το αποτέλεσμα. Για να επιτευχθεί καλή απόδοση κατά τη βαθμολόγηση, αυτές οι παράμετροι θα πρέπει, προφανώς να λάβουν τις κατάλληλες τιμές. Αυτό επιτυγχάνεται μέσω ενός συνόλου κειμένων, το οποίο χρησιμοποιείται για να "εκπαιδεύσει" το μοντέλο, και κατ'επεκταση να βρει τις κατάλληλες παραμέτρους. Ένα ακόμη θέμα είναι να βρεθεί ένας τρόπος να συνδυαστούν με αποτελεσματικό τρόπο δύο ή περισσότερα από την πληθώρα μοντέλων που έχουν προταθεί.

Παρουσιάζονται όμως προβλήματα κατα τη διαδικασία προσδιορισμού παραμέτρων, καθώς παρατηρείται πως ένα μοντέλο το οποίο είναι τέλεια ρυθμισμένο στο αρχικό σύνολο κειμένων δεν αποδίδει καλά σε νέα ερωτήματα.Έτσι μπορεί να παρουσιαστούν φαινόμενα over-fitting και under-fitting. Παράλληλα, ούτε ο συνδιασμός μοντέλων γίνεται εύκολα.Έτσι, οι παραπάνω διαδικασίες εμπλέκουν τη μηχανική μάθηση, η οποία μπορεί αποτελεσματικά να ρυθμίσει παραμέτρους αυτόματα, ή να συνδυάσει αποτελεσματικά στοιχεία, αποφεύγοντας ταυτόχρονα το over-fitting.

Συνολικά, η παραπάνω διαδικασία, δηλαδή η εμπλοκή της μηχανικής μάθησης στην κατασκευή των ranking models για συστήματα ανάκτησης πληροφορίας (όπως μηχανές αναζήτησης) ονομάζεται **learning to rank**.

Δομή του Machine Learning

Αναγνωρίζουμε τρία βασικά συστατικά που πλαισιώνουν τη δομή του learning to rank: Το *input space,* το *output space,* και το *hypothesis space*. Στο σχήμα 2 φαίνεται η παραπάνω δομή.



Σχήμα 2 Βασική δομή του Machine Learning

Ένα αντικείμενο (συνήθως feature vector) του input space, αντιστοιχίζεται σε αντικείμενο y του output space μέσω του hypothesis, το οποίο περιλαμβάνει κλάσεις συναρτήσεων mapping .

Για την εκπαίδευση του συστήματος χρησιμοποιείται ένα εξειδικευμένο σύνολο εκπαίδευσης, για το οποίο γνωρίζουμε τα input space και output space,δηλαδή το "ground truth"* του και επομένως μπορούμε να μετρήσουμε το βαθμό ακρίβειας της πρόβλεψης του hypothesis, μέσω ενός loss function.

Το παραπάνω μοντέλο χρησιμοποιείται ολοένα και περισσότερο σε μεθόδους learning to rank, ιδιαίτερα τα τελευταία χρόνια. Συγκεκριμένα, τα μοντέλα learning to rank τα οποία μελετάμε περιγράφονται από τα παρακάτω χαρακτηριστικά:

1. **Βασίζονται στα χαρακτηριστικά(feature based):** Αυτό στην πράξη σημαίνει ότι όλα τα στοιχεία του input space είναι διανύσματα χαρακτηριστικών (feature vectors) τα οποία αντικατοπτρίζουν τη συνάφεια των κειμένων στο ερώτημα. Τα διανύσματα αυτά είναι της μορφής $x = \Phi(d, q)$ όπου d το σχετικό κείμενο, q το ερώτημα και Φ συνάρτηση που εξάγει τα χαρακτηριστικά(feature extractor). Τα χαρακτηριστικά αυτά μπορεί να είναι η συχνότητα εμφάνισης ενός όρου στο κείμενο, τα αποτελέσματα που έχει δώσει ένα άλλο μοντέλο (π.χ. BM25 ή PageRank) ή σχέσεις μεταξύ αυτού του κειμένου και κάποιου άλλου.
2. **Χρησιμοποιούν discriminative training:** Αυτό σημαίνει πως η εκπαίδευση του συστήματος γίνεται όπως περιγράφηκε παραπάνω, μέσω των τεσσάρων διακριτών συστατικών στοιχείων (input space, output space, hypothesis και loss function).

*__Ground truth:__ με τον όρο αυτό αναφερόμαστε στην πληροφορία η οποία λαμβάνεται μέσω άμεσης παρατήρησης, και όχι μέσω τεκμηρίων ή συμπερασμάτων. Στην επιστήμη

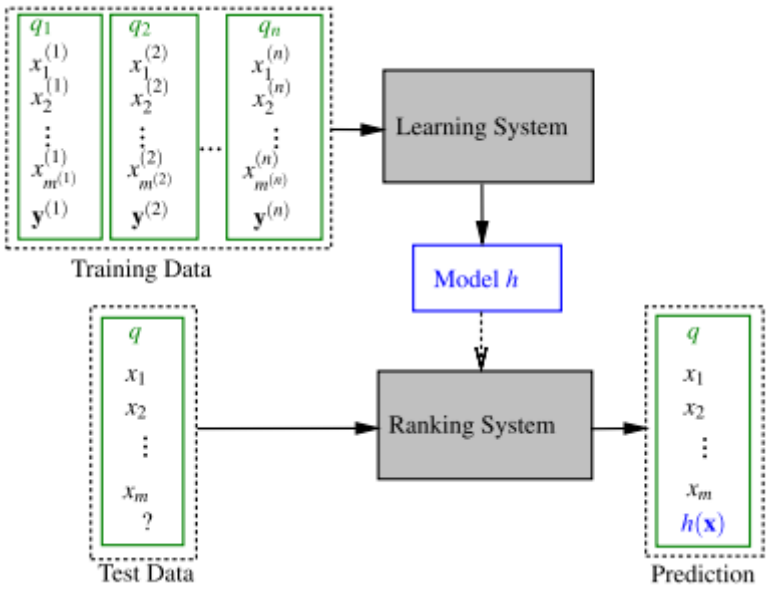
της μηχανικής μάθησης, το ground truth εκφράζει το ιδανικό αναμενόμενο αποτέλεσμα και χρησιμοποιείται για την απόδειξη ή μη της υπόθεσης του μοντέλου. Μπορεί να συγκριθεί με τα benchmarks.

Δομή Learning to Rank

Κατ'αρχας το learning to rank αποτελεί supervised training, επομένως απαιτεί ένα σύνολο κειμένων το οποίο θα το εκπαιδεύσει. Το σύνολο εκπαίδευσης (training set) του συστήματος μοιάζει με δοκιμαστικό σύνολο και αποτελείται από n ερωτήματα q_i με i=1:n και τα σχετιζόμενα κείμενα τα οποία αντιπροσωπεύονται από διανύσματα $x^{(i)} = \{x_j{}^{(i)}\}^{m^{(i)}}{}_{j=1}$ όπου $m^{(i)}$ ο αριθμός κειμένων που σχετίζονται με το ερώτημα i, καθώς και από το αντίστοιχο αποτέλεσα συσχέτισης $y^{(i)}$. Τα ερωτήματα q_i τροφοδοτούνται σε ένα αλγόριθμο μάθησης, ο οποίος μαθαίνει τον τρόπο συνδυασμού των στοιχείων ώστε να προκύπτει το ground truth y το οποίο δίνεται από το σύνολο μάθησης.

Ως αποτέλεσμα της παραπάνω διαδικασίας μάθησης έχουμε ένα μοντέλο βαθμολόγησης h το οποίο μπορεί να χρησιμοποιηθεί στο σύστημα βαθμολόγησης. Έτσι προχωράμε στη φάση δοκιμών, τροφοδοτόντας το ranking system με δοκιμαστικά κείμενα μορφής όμοιας με τα σύνολα μάθησης και λαμβάνοντας τα κείμενα ταξινομημένα.

Η δομή που περιγράφηκε παραπάνω σκιαγραφείται στο σχήμα 3:



Σχήμα 3 Σχήμα εκμάθησης βαθμολόγησης (Learning to rank) Πηγή: (Liu, 2011)

Οι learning to rank αλγόριθμοι κατηγοριοποιούνται στους

- pointwise
- pairwise
- listwise

Η παραπάνω κατηγοριοποίηση γίνεται με κριτήριο τα διαφορετικά input spaces.Παρ' όλα αυτά, σε κάθε προσέγγιση οι αλγόριθμοι που χρησιμοποιούνται μπορεί να διαφέρουν με έναν ή παραπάνω τρόπο στα τέσσερα βασικά στοιχεία της μηχανικής μάθησης.

POINTWISE APPROACH

Input space: περιλαμβάνει ένα διάνυσμα χαρακτηριστικών για κάθε κείμενο

Output space: περιλαμβάνει έναν βαθμό συνάφειας για κάθε κείμενο.

Για να εξάγουμε από τα αποτελέσματα (judgements) το ground truth, κινούμαστε, ανάλογα την περίπτωση ως εξής:

- Εάν έχουμε ως αποτέλεσμα έναν βαθμό συνάφειας l_j τότε το ground truth για κείμενο x_j ορίζεται ως $y_j = l_j$, δηλαδή το αποτέλεσμα-judgement αποτελεί απ'ευθείας το ground truth.
- Εαν έχουμε ως αποτέλεσμα ένα ζεύγος προτίμησης $l_{u,v}$τότε το ground truth λαμβάνεται αν μετρήσει κανείς τη συχντότητα με την οποία ένα κείμενο "νικά" άλλα κείμενα
- Εαν έχουμε ως αποτέλεσμα μια απόλυτη δίάταξη π_l γίνεται χρήση συνάρτησης χαρτογράφησης (mapping function)

Hypothesis space: Περιέχει συναρτήσεις οι οποίες έχουν ως είσοδο τα διανύσματα του input space και παράγουν το βαθμό σχετικότητας του κειμένου, οι οποίες ονομάζονται scoring functions.

Loss function: Εξετάζει την ακρίβεια της πρόβλεψης για κάθε ένα κείμενο ξεχωριστά.

Κατα την προσέγγιση αυτή, θεωρούμε πως βαθμός συνάφειας είναι ένας αριθμός, σύμφωνα με τον οποίο ταξινομούνται τα κείμενα. Το πρόβλημα προσεγγίζεται μέσω οπισθοδρόμισης, κατα την οποία για κάθε ζεύγος κειμένου-ερωτήματος προβλέπεται ο αντίστοιχος βαθμός συνάφειας του κειμένου. Για το σκοπό αυτό αξιοποιούνται συχνά αλγόριθμοι οπισθοδρόμισης και ταξινόμησης.

Μερικές από τις χρήσεις της προσέγγισης αυτής εντοπίζονται σε αλγορίθμους μορφολογικής ανάλυσης της ιαπωνικής γλώσσας καθώς και, σε συνδιασμό με το τρίτο κατα σειρά, listwise, στην ανάκτηση πληροφορίας από RF (Random-Forset) δέντρα

αποφάσεων, ιδιαίτερα για πολύ μεγάλα δεδομένα, στα οποία χρησιμοποιείται listwise προσέγγιση για την αρχική φάση κατασκευής των δέντρων και pointwise για τη μεταγενέστερη.

PAIRWISE APPROACH

Input space: περιλαμβάνει ζεύγη κειμένων, τα οποία αναπαρίστανται από διανύσματα χαρακτηριστικών

Output space: περιλαμβάνει προτίμιση ανα ζεύγος, δηάδή μια τιμή στο διάστημα {+1,-1}.

Για να εξάγουμε από τα αποτελέσματα (judgements) το ground truth, κινούμαστε, ανάλογα την περίπτωση ως εξής:

- Εαν το αποτέλεσμα είναι ένας βαθμός συνάφειας l_j, τότε ορίζουμε την προτίμηση στο ζεύγος κειμένων (x_u, x_v)ως: $y_{u,v} = 2 \cdot I_{\{l_u > l_v\}} - 1$.
- Εαν το αποτέλεσμα είναι απευθείας η προτίμηση στο ζεύγος, τότε εξίσου απευθείας ορίζουμε: $y_{u,v} = l_{u,v}$.
- Εαν το αποτέλεσμα δίνεται ως απόλυτη διάταξη π_l, τότε ορίζουμε: $y_{u,v} = 2 \cdot I_{\{\pi_l(u) < \pi_l(v)\}} - 1$.

Hypothesis space: περιλαμβάνει συναρτήσεις h δύο μεταβλητών, οι οποίες παίρνουν ως ορίσματα τα δύο κείμενα και εξάγουν μια σχετική ταξινόμηση ανάμεσά τους, δηλαδή μια προτίμηση του ενός έναντι του άλλου.

Loss function: μετρά την ασυνέπεια των αποτελεσμάτων της συνάρτησης $h(x_u, x_v)$ και του ground truth $y_{u,v}$.

Η προσέγγιση αυτή είναι δυαδική, καθώς από ένα ζεύγος κειμένων προτιμάται αυτό με το μεγαλύτερο βαθμό συνάφειας. Αποσκοπεί στο να μειώσει στο ελάχιστο την αναστροφή κειμένων. Παρα τα πλεονεκτήματα που προσφέρει, παρατηρήθηκε πως η προσέγγιση listwise ξεπερνά την pairwise. Από πειραματικά δεδομένα (Zhe Chao, 2007), φαίνεται πώς και το loss function της pairwise λειτουργεί “ανα ζεύγη¨, κάτι που τελικά αποτελεί κακή προσέγγιση του ground truth.

LISTWISE APPROACH

Input space: Περιλαμβάνει κείμενα x τα οποία σχετίζονται σε κάποιο βαθμό με το ερώτημα.

Output space: Περιλαμβάνει μία μετάθεση των κειμένων του input space σε μια βαθμολογημένη λίστα Y όπου ο δείκτης y(i) δείχνει τη θέση i του κείμένου στη τελική λίστα.

Για να εξάγουμε από τα αποτελέσματα (judgements) το ground truth, κινούμαστε, ανάλογα την περίπτωση ως εξής:

- Εάν το αποτέλεσμα δίνεται ως βαθμός l_j τότε οι μεταθέσεις οι οποίες συμφωνούν με το αποτέλεσμα θεωρούνται ground truth.Ως παράδειγμα, μια μετάθεση π_y είναι σύμφωνη με το βαθμό συνάφειας l_j εάν: $\forall\ u, v :\ l_u > l_v$ ισχύει πάντα $\pi_y(u) < \pi_y(v)$
- Παρομοίως, εάν το αποτέλεσμα δίνεται ως ζεύγη προτιμήσεων, τότε οι μεταθέσεις αυτές οι οποίες συμφωνούν με τα ζεύγη προτιμήσεων θεωρούνται ground truth. Στην περίπτωση αυτή τα u,v θα πρέπει να ικανοποιούν τη συνθήκη $l_{u,v} = +1$.
- Εάν το αποτέλεσμα δίνεται ως απόλυτη διάταξη π_l τότε ορίζουμε απευθείας:

$$\pi_y = \pi_l$$

Hypothesis space: περιλαμβάνει συναρτήσεις πολλών μεταβλητών οι οποίες πράττουν πάνω σε σύνολο κειμένων, προβλέποντας μια μετάθεσή τους. Για το σκοπό αυτό, συνήθως χρησιμοποιείται αρχικά μια συνάρτηση βαθμολόγησης, η οποία αναθέτει έναν αριθμό (βαθμό) σε κάθε κείμενο κι έπειτα τα κείμενα ταξινομούνται κατά φθίνουσα σειρά. Η μετάθεση που προκύπτει αποτελεί την πρόβλεψη της υπόθεσης h.

Σε αντίθεση με τις προηγούμενες προσεγγίσεις, στις οποίες οι έξοδος συχνά δεν είναι κατάλληλη για να τροφοδοτήσει τη διαδικασία εκπαίδευσης του μοντέλου, στην προσέγγιση κατά λίστες (listwise) παρατηρούμε πως και η είσοδος (input space) και η έξοδος (output space) είναι σύνολα κειμένων. Έτσι το output space το οποίο διευκολύνει τη διαδικασία μάθησης είναι ίδιο με το output space της διαδικασίας στην προσέγγιση κατά λίστες, και επομένως αυτό μπορεί να χρησιμοποιηθεί.

Loss function: Διακρίνουμε δύο είδη: τις συναρτήσεις οι οποίες συνδέονται ρητά με τα μέτρα αξιολόγησης και αυτές που δεν συνδέονται. Επιπρόσθετα και λόγω του ότι κάποια δομικά στοιχεία των συναρτήσεων αυτών στην προσέγγιση κατά λίστες φαίνεται να ανήκουν στις κατά ζεύγη ή κατά σημεία, ξεχωρίζουμε μια συνάρτηση απώλειας (loss function) βάσει του ότι:

- Ορίζεται βάσει των κειμένων εκπαίδευσης που σχετίζονται με ένα ερώτημα,
- Δεν αναλύεται απόλυτα σε απλό άθροισμα κειμένων ή ζευγαριών κειμένων και
- Δίνει έμφαση στη δημιουργία βαθμολογημένης λίστας, της οποίας η θέσεις των κειμένων της είναι εμφανείς.

Η παρούσα εργασία εμβαθύνει στην πρώτη προσέγγιση στο learning to rank: pointwise approach.

Προσέγγιση κατά σημεία (Pointwise Approach)

Η συγκεκριμένη προσέγγιση καθορίζεται από το ότι αντιμετωπίζει το κάθε κείμενο ξεχωριστά και αποσκοπεί στην παραγωγή ταξινομημένης λίστας των κειμένων, βάσει του βαθμού τους.

Υλοποιείται με αλγορίθμους οπισθοδρόμησης και ταξινόμησης. Πιο συγκεκριμένα, χρησιμοποιούνται:

1. αλγόριθμοι οπισθοδρόμησης, όπου η έξοδος αποτελείται από βαθμούς συνάφειας εκφρασμένους σε πραγματικούς αριθμούς,
2. αλγόριθμοι ταξινόμησης, όπου η έξοδος αποτελείται από μη διατεταγμένες κατηγορίες
3. αλγόριθμοι γραμμικής παλινδρόμησης, όπου η έξοδος αποτελείται από διατεταγμένες κατηγορίες.

ΑΛΓΟΡΙΘΜΟΙ ΟΠΙΣΘΟΔΡΟΜΙΣΗΣ

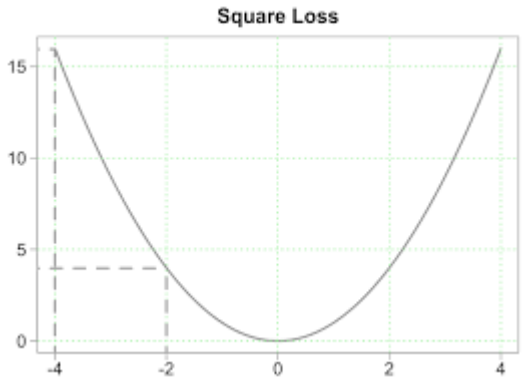
Οι αλγόριθμοι αυτής της κατηγορίας χαρακτηρίζονται από αναίρεση των αποτελεσμάτων σε κάποιο υπολογιστικό βήμα, οπισθοδρόμηση σε προηγούμενο και λήψη διαφορετικών αποφάσεων. Λόγω του ότι στην οπισθοδρόμηση θεωρούμε πως η μεταβλητή που προσπαθούμε να προβλέψουμε είναι συνεχής, στην περίπτωση της βαθμολόγησης με αυτό το είδος αλγορίθμου θεωρούμε το κείμενο στην είσοδο ως μια συνεχή μεταβλητή. Το σύστημα μαθαίνει τη συνάρτηση βαθμολόγησης ελαχιστοποιώντας το σφάλμα για το σύνολο εκπαίδευσης.

Ένας χαρακτηριστικός αλγόριθμος της κατηγορίας αυτής που χρησιμοποιείται στο learning to rank είναι η *βαθμολόγηση υποσυνόλων με οπισθοδρόμηση*(subset ranking with regression). Ορίζουμε:

- $\boldsymbol{x} = \{x_j\}_{j=1}^{m}$ το σύνολο κειμένων που σχετίζεται με το ερώτημα q
- $\boldsymbol{y} = \{y_j\}_{j=1}^{m}$ τα ground truth των κειμένων του συνόλου **x**
- $\boldsymbol{f}$ η συνάρτηση βαθμολόγησης που χρησιμοποιείται για την κατάταξη των κειμένων του συνόλου **x**
- $\boldsymbol{L}(f; x_j, y_j) = (y_j - f(x_j))^2$ η συνάρτηση απώλειας, η οποία όπως φαίνεται ορίζεται ως τετραγωνικό σφάλμα.

Είναι γνωστό πως το τετραγωνικό σφάλμα γραφικά απεικονίζεται όπως στο σχήμα 4. Επομένως βλέπουμε πως το σφάλμα μηδενίζεται για

$$L = (y_j - f(x_j))^2 = 0 \Rightarrow y_j - f(x_j) = 0 \Rightarrow y_j = f(x_j) \qquad (1)$$



Σχήμα 4

Εάν η (1) δεν ισχύει, τότε το σφάλμα αυξάνεται τετραγωνικά.

Επομένως, εάν ένα κείμενο θεωρείται συναφές με το ερώτημα και επομένως το ground truth του είναι $y_j = 1$, τότε η συνάρτηση βαθμολόγησης f θα πρέπει να παράγει πάντα 1, ώστε να μην υπάρχει σφάλμα. Μια όμως πιο ισχυρή πρόβλεψη συνάφειας για το κείμενο θα ήταν $f(x_j) = 2$. Μια τέτοια πρόβλεψη όμως παράγει σφάλμα. Για την επίλυση του προβλήματος αυτού γίνεται χρήση βάρους στη συνάρτηση, ώστε να μειωθεί το σφάλμα στα κείμενα που βρίσκονται ψηλότερα στην κατάταξη όσον αφορά τη συνάφειά τους στο ερώτημα.

Για την μέτρηση της ποιότητας της βαθμολόγησης και της αποτελεσματικότητας του αλγορίθμου χρησιμοποιείται NDCG (Normalized Discounted Cumulative Gain). Το NDCG αποτελεί κανονικοποίηση του DCG , το οποίο με τη σειρά του μειώνει λογαριθμικά τη βαθμολογημένη συνάφεια του κειμένου όσο χαμηλότερα βρίσκεται αυτό στη λίστα που προκύπτει από τη βαθμολόγηση βάσει του ερωτήματος q. Σύμφωνα με (Liu, 2011) και (Cossock & Zhang, 2006), το τετραγωνικό σφάλμα αποτελεί άνω φράγμα του σφάλματος βαθμολόγησης βάσει NDCG. Επιπρόσθετα, εφ' όσον ενδιαφερόμαστε για τη σειρά με την οποία εμφανίζονται τα κείμενα στην τελική βαθμολογημένη λίστα, ακόμα και στην περίπτωση που προκύπτει μεγάλο σφάλμα κατά την οπισθοδρόμηση , είναι δυνατόν να έχουμε ιδανικό αποτέλεσμα, εφ' όσον η σχετική σειρά των προβλέψεων είναι η ίδια με αυτή του ground truth.

ΑΛΓΟΡΙΘΜΟΙ ΤΑΞΙΝΟΜΗΣΗΣ

ΔΥΑΔΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

- **Support Vector Machine** (Liu, 2011)**,** (Nallapati, 2004)
  Όταν χρησιμοποιείται η μέθοδος αυτή, ο αλγόριθμος ταξινομεί τα δεδομένα σε μια από τις δύο κατηγορίες: σχετικά με το ερώτημα q, ή μη σχετικά. Τα δεδομένα, ακόμα και άπειρων διαστάσεων, διαχωρίζονται από υπερεπίπεδο το οποίο, από τον ορισμό του, είναι πάντα μιας λιγότερης διάστασης από το περιβάλλοντα χώρο, καθιστώντας έτσι το διαχωρισμό των κατηγοριών εφικτό.
  Με δεδομένα κείμενα $\boldsymbol{x} = \{x_j\}_{j=1}^{m}$ και οι δυαδικοί βαθμοί συσχέτισής τους με το ερώτημα q, $\mathbf{y}= \{y_j\}_{j=1}^{m}$, θεωρούμε τα σχετικά κείμενα ως θετικά παραδείγματα $(y_j = +1)$ και τα μη σχετικά ως αρνητικά, $(y_j = -1)$ και η βαθμολόγηση ανάγεται σε

πρόβλημα δυαδικής ταξινόμησης.
Εάν, επίσης υποθέσουμε γραμμική συνάρτηση βαθμολόγησης της μορφής $f(w,x) = w^T x$, όπου w το διάνυσμα βάρους , το οποίο το σύστημα το μαθαίνει μέσω του εκπαιδευτικού συνόλου τότε η μέθοδος SVM διατυπώνεται ως:

$$\boldsymbol{min}\frac{\mathbf{1}}{\mathbf{2}}\|\boldsymbol{w}\|^2 + \boldsymbol{\lambda}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{m^{(i)}}}\boldsymbol{\xi_j^{(i)}}$$

με:

$$w^T x_j^{(i)} \leq -1 + \xi_j^{(i)}, \gamma\iota\alpha\ y_j^{(i)} = 0$$
$$w^T x_j^{(i)} \geq 1 - \xi_j^{(i)}, \gamma\iota\alpha\ y_j^{(i)} = 1$$
$$\xi_j^{(i)} \geq 0, \gamma\iota\alpha\ j = 1{:}\,m^{(i)}, i = 1{:}\,n$$

Το πρόβλημα που αντιμετωπίζει το μοντέλο SVM είναι εάν το κάθε κείμενο ταξινομείται στη σωστή κατηγορία ή όχι. Συχνά, λόγω του κατωφλιού που επιλέγεται για να διαχωρίσει τις κατηγορίες, τα δεδομένα μπορεί να ταξινομηθούν σε κατηγορία Α, ακόμα κι αν απέχουν περισσότερο από το κοντινότερο δεδομένο της κατηγορίας Α, παρά από το αντίστοιχο της Β. Το σφάλμα ή απώλεια για κάθε κείμενο είναι $\xi_j^{(i)}$ σε περίπτωση που η έξοδος του μοντέλου για ένα σχετικό κείμενο $(y_j = +1)$ είναι μικρότερη του +1.

Το παραπάνω πρόβλημα είναι επιλύσιμο στη δυαδική μορφή του και μέσω του kernel trick, το σύστημα μπορεί να μάθει και μη γραμμική συνάρτηση, επεκτείνοντας τη γραμμική λύση στο kernel space.
Πειραματικά δεδομένα δείχνουν πώς το μοντέλο SVM είναι συγκρίσιμο ή και καλύτερο από γλωσσικά μοντέλα, όπως φαίνεται στην εικόνα 2, όπου συγκρίθηκαν οι αποδόσεις των μοντέλων σε δοκιμαστικά δεδομένα ερωτήματος που αφορά εύρεση της σωστής αρχικής σελίδας για το ερώτημα ενός χρήστη (homepage finding). Το MRR(mean reciprocal rank) εκφράζει τη μέση αμοιβαία κατάταξη ενός κειμένου για όλα τα ερωτήματα, το Success (επιτυχία) εκφράζει το ποσοστό των ερωτημάτων για τα οποία η απάντηση βρίσκεται στα κορυφαία 10 κείμενα και το Failure (αποτυχία) εκφράζει το ποσοστό των ερωτημάτων των οποίων η απάντηση βρίσκεται στα κορυφαία 100 κείμενα.

Model	MRR	Success %	Failure %
full-featured SVM	0.52	77.93	11.03
LM baseline	0.35	57.93	15.86
SVM baseline	0.28	52.41	17.9

Εικόνα 2 Σύγκριση SVM και LM, Πηγή: (Nallapati, 2004)

- **Logistic Regression** (Liu, 2011)**,** (Gey, 1994)
  Παρά το γεγονός ότι καλείται «οπισθοδρόμηση» (regression), η τεχνική αυτή αποτελεί μια από τις πιο δημοφιλείς τεχνικές ταξινόμησης. Προβλέπει εάν το κείμενο είναι σχετικό με το ερώτημα ή όχι, υπολογίζοντας πιθανότητα του να ισχύει μια από τις δύο περιπτώσεις βάσει των όρων που υπάρχουν στο κείμενο. Δεδομένων ενός ερωτήματος, κειμένου και όρου για τα οποία γνωρίζουμε τη δυαδική συσχέτιση, υπολογίζουμε το λογάριθμο της πιθανότητας ένας όρος t_k, ο οποίος εμφανίζεται και στο κείμενο και στο ερώτημα της «τριάδας» του, να είναι σχετικός (true/false):

$$\log O(R \mid q_i, d_j, t_k) = c_0 + c_1 v_1 + \cdots + c_l v_n$$ (Gey, 1994)

  Επομένως, το άθροισμα όλων των επιμέρους λογαρίθμων πιθανοτήτων για κάθε όρο θα δίνει την πιθανότητα συσχέτισης για το ερώτημα q_i

$$\log O(R \mid q_i, d_j) = \sum_{k=1}^{q} [\log O(R \mid q_i, d_j, t_k) - \log O(R)]$$ (Gey, 1994)

  Άρα λογάριθμος της πιθανότητας το κείμενο x_j να σχετίζεται με το ερώτημα ορίζονται ως:

$$\log\left(\frac{P(R|x_j)}{1 - P(R|x_j)}\right) = c + \sum_{t=1}^{T} w_t x_{j,t}$$

  Όπου c σταθερά (Liu, 2011)

  Το οποίο ισοδυναμεί με τον παρακάτω ορισμό για την πιθανότητα συσχέτισης ενός κειμένου με ένα ερώτημα:

$$P(R|x_j) = \frac{1}{1 + e^{-c - \sum_{t=1}^{T} w_t x_{j,t}}}$$

  Το επόμενο βήμα είναι να εκπαιδευτεί το παραπάνω μοντέλο με ένα σύνολο «δεδομένων εξάσκησης» (training data), ώστε να υπολογιστεί η παράμετρος w_t μεγιστοποιώντας την πιθανότητα.
  Οι όροι που χρησιμοποιούνται σύμφωνα με (Gey, 1994) είναι: **απόλυτη συχνότητα ερωτήματος (query absolute frequency - QAF)**, **σχετική συχνότητα ερωτήματος (query relative frequency - QRF)** η οποία ισούται με την απόλυτη συχνότητα ερωτήματος προς το συνολικό αριθμό εμφανίσεων όλων των όρων στο ερώτημα, **απόλυτη συχνότητα κειμένου (document absolute frequency - DAF)**, **σχετική συχνότητα κειμένου (document relative frequency - DRF)** η οποία, ομοίως με QRF ισούται με DAF προς το συνολικό αριθμό εμφανίσεων όλων των όρων στο κείμενο, **σχετική συχνότητα σε όλα τα κείμενα(relative frequency in all documents-RFAD)** η οποία είναι απλά ο συνολικός αριθμός εμφανίσεων του όρου στα κείμενα προς το συνολικό

αριθμό εμφανίσεων όλων των όρων σε όλα τα κείμενα και **συχνότητα ανεστραμμένου κειμένου (inverse document frequency- IDF).** Για να μετριαστεί η επιρροή της πληροφορίας συχνότητας («Δεν είναι ρεαλιστικό να υποθέσουμε πως 50 εμφανίσεις ενός όρου στο κείμενο είναι πέντε φορές πιο σημαντικές από 10 εμφανίσεις»- (Gey, 1994), σελ 3) και για να εξομαλυνθεί η κατανομή, οι παραπάνω έξι όροι λογαριθμίζονται. Έτσι προκύπτει η παρακάτω σχέση η οποία εκφράζει το μοντέλο:

$$Z_{t_j} = \log O(R \mid t_j) = c_0 + c_1 log(QAF) + c_2 log(QRF)$$

$$+ c_3 log(DAF) + c_4 log(DRF) + c_5 log(IDF) + c_6 log(RFAD)$$ (Gey, 1994)

Το παραπάνω μοντέλο δοκιμάστηκε σε τυπικά δεδομένα συλλογών Cranfield και συγκρίθηκε με το μοντέλο vector space. Τα αποτελεσματα του πίνακα 1 δείχνουν το κομμάτι των σχετικών κειμένων που ανακτήθηκαν σε δεδομένο σημείο των δοκιμών (recall) καθώς και το κομμάτι των ανακτημένων κειμένων τα οποία είναι σχετικά (precision).

Logistic Inference versus tfidf/cosine Vector Space Performance Cranfield Collection: Averages over 225 queries		
Recall	Logistic Precision	Vector Space Precision
0.00	0.8330	0.7787
0.10	0.8116	0.7440
0.20	0.7129	0.6434
0.30	0.6021	0.5301
0.40	0.5161	0.4380
0.50	0.4503	0.3814
0.60	0.3698	0.2994
0.70	0.2859	0.2267
0.80	0.2280	0.1882
0.90	0.1640	0.1379
1.00	0.1464	0.1251
11-pt Avg:	0.4655	0.4084
% Change:		-12.3

Πίνακας 1 Αποτελεσματα δοκιμών. Πηγή: (Gey, 1994)

Παρατηρούμε πως τα δεδομένα αποδεικνύουν πως το μοντέλο Logistic Regression αποδίδει καλύτερα από το μοντέλο Vector Space.

- **Boosting Tree** (Li, Burges, & Wu, 2007)
  Ο αλγόριθμος που προτείνεται στο (Li, Burges, & Wu, 2007) χρησιμοποιεί αλγόριθμο gradient boosting με την παρακάτω σχέση για τη μέτρηση σφάλματος:

$$L_{\varphi}(\hat{y}_j, y_j) = \sum_{j=1}^{m} \sum_{k=1}^{K} (-logP(\hat{y}_j = k) I_{\{y_j = k\}}) \qquad \text{(Liu, 2011)}$$

Όπου $\hat{y}_j$ η πρόβλεψη του μοντέλου για το κείμενο x_j

Το σφάλμα $L(\hat{y}_j, y_j)$ παίρνει τιμές στο διάστημα [0,1] και χρησιμοποιείται ώστε να διατηρείται καλός ο λόγος bias-variance. Εάν η τιμή του είναι 1, τα δεδομένα ταιριάζουν υπερβολικά τέλεια στο training set (low bias) κάτι που όμως οδηγεί στο να μην ανταποκρίνεται το μοντέλο καλά σε δοκιμαστικά δεδομένα (high variance).

Δεδομένων κειμένων x_j, μια συνάρτηση $F_k(x_j, w)$ καθορίζει το βαθμό στον οποίο το κείμενο x_j ανήκει στην τάξη k. Μετά την ολοκλήρωση της ταξινόμησης , κατά τη φάση των δοκιμών, τα αποτελέσματα της ταξινόμησης μετατρέπονται σε βαθμολόγηση.

Αρχικά, η έξοδος τους ταξινομητή μετατρέπεται σε πιθανότητα (το κείμενο x_j να ανήκει στην τάξη k) με χρήση της παρακάτω σχέσης:

$$P(\hat{y}_j = k) = \frac{e^{F_k(x_j, w)}}{\sum_{s=1}^{K} e^{F_k(x_j, w)}} \qquad \text{(Liu, 2011)}$$

Έπειτα, με χρήση μιας συνεχούς αύξουσας συνάρτησης του βαθμού συσχέτισης της τάξης k με το ερώτημα που εξετάζεται(έστω $g(\cdot)$), το τελικό αποτέλεσμα βαθμολόγησης ορίζεται ως:

$$f(x_j) = \sum_{k=1}^{K} g(k) \cdot P(\hat{y}_j = k) \qquad \text{(Liu, 2011)}$$

Στην εικόνα 3 παρουσιάζεται ο αλγόριθμος όπως εμφανίζεται στο (Li, Burges, & Wu, 2007)

0: $\tilde{y}_{i,k} = 1$, if $y_i = k$, and $\tilde{y}_{i,k} = 0$ otherwise.
1: $F_{i,k} = 0$, $k = 0$ to $K-1$, $i = 1$ to N
2: For $m = 1$ to M Do
3: For $k = 0$ to $K-1$ Do
4: $p_{i,k} = \exp(F_{i,k}) / \sum_{s=0}^{K-1} \exp(F_{i,s})$
5: $\{R_{j,k,m}\}_{j=1}^{J} = J$-terminal node regression tree for $\{\tilde{y}_{i,k} - p_{i,k},\ \mathbf{x}_i\}_{i=1}^{N}$
6: $\beta_{j,k,m} = \frac{K-1}{K} \frac{\sum_{\mathbf{x}_i \in R_{j,k,m}} \tilde{y}_{i,k} - p_{i,k}}{\sum_{\mathbf{x}_i \in R_{j,k,m}} (1-p_{i,k}) p_{i,k}}$
7: $F_{i,k} = F_{i,k} + \nu \sum_{j=1}^{J} \beta_{j,k,m} 1_{\mathbf{x}_i \in R_{j,k,m}}$
8: End
9: End

Εικόνα 3 Ο αλγόριθμος Boosting Tree. Πηγή: (Li, Burges, & Wu, 2007)

Πειραματικά δεδομένα που παρουσιάζονται στο (Li, Burges, & Wu, 2007) δείχνουν πως ο προτεινόμενος Boosting Tree αλγόριθμος McRank υπερτερεί άλλων αλγορίθμων βασισμένων σε οπισθοδρόμηση ή ζεύγη. Στην εικόνα 4 παρουσιάζονται τα αποτελέσματα αυτά. Ο McRank εξετάστηκε και στην περίπτωση τακτικής ταξινόμησης και σε απλή ταξινόμηση σε τέσσερα ήδη datasets (Artificial, Web-1, Web-2, Web-3), Στην εικόνα 4 φαίνονται τα ποσοστά μέσων βαθμών NDCG για τους τέσσερις αλγορίθμους.

Datasets	Ordinal Classification	Classification	Regression, p-value	LambdaRank, p-value
Artificial [5]	85.0 (9.5)	83.7 (9.9)	82.9 (10.2), 0	74.9, (12.6), 0
Web-1 [5]	72.4 (24.1)	72.2 (24.1)	71.7 (24.4), 0.021	71.2 (24.5), 0.0002
Web-2 [13]	—	75.8 (23.8)	74.7 (24.4), 0.023	74.3 (24.3), 0.003
Web-3	72.5 (26.5)	72.4 (27.3)	72.0 (27.6), 0.017	71.3 (28.8), 3.8×10^{-7}

Εικόνα 4 Πειραματικά αποτελέσματα απόδοσης Boosting Tree αλγορίθμων. Πηγή: (Li, Burges, & Wu, 2007)

- **Association Rule Mining** (Veloso, Almeida, Gonçalves, & Meira , 2008)
  Σε αυτή την κατηγορία, βλέπουμε αλγορίθμους οι οποίοι προτάθηκαν από ερευνητές του πεδίου της εξόρυξης δεδομένων. Συγκεκριμένα στο (Veloso, Almeida, Gonçalves, & Meira , 2008), προτείνεται μέθοδος η οποία χρησιμοποιεί κανόνες συσχέτισης, δηλαδή πρότυπα τα οποία περιγράφουν συσχετίσεις της μορφής $X \rightarrow Y$ και αναζητά στα κείμενα που χρησιμοποιούνται για εκπαίδευση του συστήματος κανόνες r ικανούς να κατηγοριοποιούν κάποιο κείμενο σε κατηγορία y.
  Οι κανόνες αυτοί είναι της μορφής $X \rightarrow r_1$ και η ποιότητά τους μετράται με δείκτες υποστήριξης (support) και αυτοπεποίθησης (confidence). Το support ενός κανόνα ορίζεται ως $\sigma(X \rightarrow r_1)$ και αποτελεί το κομμάτι των δεδομένων εκπαίδευσης που περιέχει χαρακτηριστικά X και βαθμό r_1.Το confidence ενός κανόνα ορίζεται ως $\theta(X \rightarrow r_1)$ και αποτελεί την υποσυνθήκη πιθανότητα ένα κείμενο, αν περιέχει χαρακτηριστικό X να έχει βαθμό r_1. Επιλέγονται οι κανόνες που έχουν $\sigma > \sigma_{min}$(support threshold) και ταυτόχρονα $\theta > \theta_{min}$ (confidence threshold). Επομένως, σκοπός του association rule είναι να ελαχιστοποιήσει τη συνάρτηση απώλειας, η οποία ορίζεται ως:
  $$-\sum_F H(P(y,F),\sigma_{min})H(P(y|F),\theta_{min}) \qquad \text{(Liu, 2011)}$$

Όπου F ο κανόνας r ο οποίος ορίζεται ως εύρος χαρακτηριστικών (ως παράδειγμα στο (Liu, 2011) *αναφέρεται το BM25=[0.56-0.70]), y η κατηγορία στην οποία εξετάζουμε εάν ανήκει ένα κείμενο, και* $H(a,b) = \begin{cases} 1, a \geq b \\ 0, αλλιώς \end{cases}$.

Στους αλγορίθμους αυτούς δεν αρκεί να βαθμολογήσουμε τα αποτελέσματα της παραπάνω διαδικασίας μέσω της scoring function f, αλλά κανονικοποιούμε το confidence P(y|F) ενός κανόνα r σύμφωνα με τη παρακάτω σχέση

$$\bar{\theta}(y) = \frac{\sum_r P(y|F)}{\sum_r 1}$$

κι έπειτα λαμβάνουμε το βαθμό s του κειμένου ως:

$$s = \sum_y y \frac{\bar{\theta}(y)}{\sum_y \bar{\theta}(y)}$$

Τέλος, τα κείμενα ταξινομούνται βάσει των βαθμών τους κατά φθίνουσα σειρά.

Πειραματικά, τρεις διαφορετικοί αλγόριθμοι ($\mathcal{R}, \mathcal{R}_d, \mathcal{R}_d^q$) εξετάστηκαν. Τα κατώφλια ορίστηκαν μέσω συνόλου επικύρωσης (validation set) ως $\sigma_{min} = 0{,}001$ και $\theta_{min} = 0{,}25$ με αποτέλεσμα να παραχτούν τα αποτελέσματα που φαίνονται στην εικόνα 5.

Trial	AR			Ranking SVM	RankBoost	FRank	ListNet	AdaRank		MHR
	$\mathcal{R}$	$\mathcal{R}_d$	$\mathcal{R}_d^q$					MAP	NDCG	
1	**0.355** (0.15)	**0.379** (0.06)	**0.379** (0.06)	0.334	0.340	0.345	0.346	0.341	0.348	0.329
2	**0.452** (0.37)	**0.463** (0.10)	**0.475** (0.03)	0.451	0.447	0.461	0.450	0.449	0.450	0.443
3	0.445 (0.90)	0.453 (0.82)	**0.469** (0.46)	0.460	0.446	0.449	0.467	0.458	0.457	0.456
4	0.512 (0.80)	**0.518** (0.44)	**0.525** (0.27)	0.511	0.506	0.515	0.517	0.507	0.509	0.502
5	0.457 (0.77)	**0.472** (0.37)	**0.472** (0.38)	0.480	0.464	0.463	0.468	0.454	0.447	0.471
Avg	0.444 (0.18)	**0.457** (0.18)	**0.464** (0.04)	0.447	0.440	0.446	0.449	0.442	0.442	0.440
Overall Improvements obtained by AR ($\mathcal{R}$)				-0.67%	0.91%	-0.45%	-1.11%	0.45%	0.45%	0.91%
Overall Improvements obtained by AR ($\mathcal{R}_d$)				2.24%	3.86%	2.47%	1.78%	3.39%	3.38%	3.86%
Overall Improvements obtained by AR ($\mathcal{R}_d^q$)				3.80%	5.45%	4.04%	3.34%	4.98%	4.98%	5.45%

Εικόνα 5 Αποτελέσματα πειραμάτων σε OHSUMED Πηγή: (Veloso, Almeida, Gonçalves, & Meira , 2008)

Παρατηρήθηκε αρκετά καλή απόδοση των προτεινόμενων αλγορίθμων. Οι (Veloso, Almeida, Gonçalves, & Meira , 2008) σημειώνουν, επίσης πως η μέθοδος αυτή είναι αρκετά αποτελεσματική, διότι εκμεταλλεύεται το ίδιο το ερώτημα και όρους που βρίσκονται σε αυτό, εξάγοντας μόνο χρήσιμη πληροφορία από τα κείμενα, με σκοπό την όσο το δυνατό πιο αποτελεσματική βαθμολόγηση.

ΑΛΓΟΡΙΘΜΟΙ ΤΑΚΤΙΚΗΣ ΟΠΙΣΘΟΔΡΟΜΗΣΗΣ

Έστω K ταξινομημένες κατηγορίες δεδομένων. Ζητείται συνάρτηση βαθμολόγησης f τέτοια ώστε, με χρήση K κατωφλιών $b_1 \leq b_2 \leq \cdots \leq b_{K-1} \leq b_K = \infty$ τα αποτελέσματα της συνάρτησης f να μπορούν να ταξινομηθούν σε διαφορετικές κατηγορίες. Για K=2, το πρόβλημα μετατρέπεται σε απλή δυαδική ταξινόμηση.

PERCEPTRON-BASED RANKING (PRanking) (Crammer & Springer, 2001)

Ο αλγόριθμος αυτός χωρίζει το χώρο των αποτελεσμάτων βάσει των προαναφερθέντων κατωφλιών, αναθέτοντας στο χώρο ανάμεσα από δύο κατώφλια $b_{r-1} < w \cdot x < b_r$ τον ίδιο βαθμό r. Στην προηγούμενη σχέση, w είναι ένα διάνυσμα παραμέτρων το οποίο χρησιμοποιείται για τις προβλέψεις, ενώ x ένα κείμενο ή, όπως αναφέρουν οι (Crammer & Springer, 2001), ένα «περιστατικό» (instance) εισόδου.

Η διαδικασία μάθησης γίνεται μέσω βρόγχου επανάληψης. Ο αλγόριθμος παρουσιάζεται στην εικόνα 6. Σε κάθε βήμα, προβλέπεται ο νέος βαθμός $\hat{y_j} = argmin_k\{w^T x_j - b_k < 0\}$ και έπειτα λαμβάνει το ground truth της πρόβλεψης, y_j και ανανεώνει τον κανόνα βαθμολόγησής του τροποποιώντας τα w,b. Εάν $\hat{y_j} \neq y_j$, ο αλγόριθμος έχει κάνει λάθος και επομένως υπάρχει κατώφλι b_k το οποίο πρέπει να «διαβεί» η τιμή $w^T x_j$.Έτσι, η τιμή$w^T x_j$ και το κατώφλι b_k πρέπει να μετακινηθούν το ένα προς το άλλο. Μια τέτοια περίπτωση δείχνει το σχήμα 5.

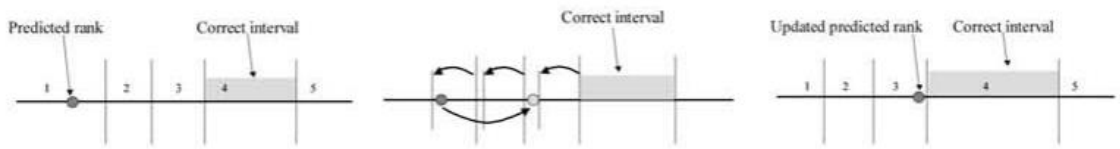
Η παραπάνω διαδικασία επαναλαμβάνεται μέχρι να συγκλίνει η διαδικασία εκπαίδευσης.

Initialize: Set $\mathbf{w}^1 = 0$, $b_1^1, \ldots, b_{k-1}^1 = 0, b_k^1 = \infty$.
Loop: For $t = 1, 2, \ldots, T$

- Get a new rank-value $\mathbf{x}^t \in \mathbb{R}^n$.
- Predict $\hat{y}^t = \min_{r\in\{1,\ldots,k\}}\{r : \mathbf{w}^t \cdot \mathbf{x}^t - b_r^t < 0\}$.
- Get a new label y^t.
- If $\hat{y}^t \neq y^t$ update $\mathbf{w}^t$ (otherwise set $\mathbf{w}^{t+1} = \mathbf{w}^t$, $\forall r : b_r^{t+1} = b_r^t$) :
  1. For $r = 1, \ldots, k-1$: If $y^t \leq r$ Then $y_r^t = -1$ Else $y_r^t = 1$.
  2. For $r = 1, \ldots, k-1$: If $(\mathbf{w}^t \cdot \mathbf{x}^t - b_r^t) y_r^t \leq 0$ Then $\tau_r^t = y_r^t$ Else $\tau_r^t = 0$.
  3. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t + (\sum_r \tau_r^t)\mathbf{x}^t$. For $r = 1, \ldots, k-1$ update: $b_r^{t+1} \leftarrow b_r^t - \tau_r^t$

Output : $H(\mathbf{x}) = \min_{r\in\{1,\ldots,k\}}\{r : \mathbf{w}^{T+1} \cdot \mathbf{x} - b_r^{T+1} < 0\}$.

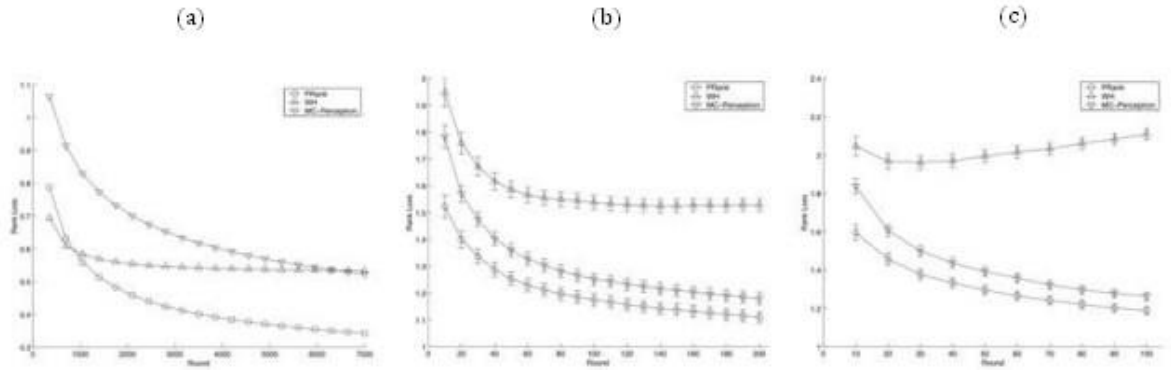
Εικόνα 6 Ο αλγόριθμος PRank. Πηγή: (Crammer & Springer, 2001)



Σχήμα 5 Παράδειγμα του κανόνα ανανέωσης για λάθος του αλγορίθμου. Πηγή: (Crammer & Springer, 2001)

Η απόδοση του PRank μετρήθηκε από (Crammer & Springer, 2001) σε δύο σύνολα δεδομένων. Το πρώτο ήταν συνθετικά δεδομένα, ενώ το δεύτερο ήταν κριτικές χρηστών για 100 και 200 ταινίες. Από τους 7,542 χρήστες, ένας επιλέχθηκε στην τύχη ώστε οι βαθμολογίες του να θεωρηθούν ground truth. Μετρήθηκε το μέσο σφάλμα ανα κύκλο αλγορίθμου, και τα αποτελέσματα συγκρίθηκαν με τους αλγορίθμους WH(Widrow-Hoff) και MCP(Multiclass generalization of the Perceptron algorithm). Τα

αποτελέσματα φαίνονται στην εικόνα 7 και δείχνουν πως, παρά την μεγαλύτερη πολυπλοκότητά του, ο PRank ξεπερνά σταθερά τους WH,MCP καθ' όλη τη διάρκεια εκτέλεσής του.



Σχήμα 6 Πειραματικά δεδομένα σε (a)συνθετικά δεδομένα, (b) κριτικές χρηστών σε 200 ταινίες, (c) κριτικές χρηστών σε 100 ταινίες. Πηγή: (Crammer & Springer, 2001)

LARGE MARGIN PRINCIPLES (Shashua & Levin, 2002)

Οι (Shashua & Levin, 2002) χρησιμοποιήσαν SVM(Support Vector Machines, βλ. ενότητα ΔΥΑΔΙΚΗ ΤΑΞΙΝΟΜΗΣΗ) για την εκμάθηση των παραμέτρων w, b_k του PRank. Διακρίνονται δύο στρατηγικές:

- *Fixed-margin:* Υπολογίζεται το $w^T x_j^{(i)}$ και στη συνέχεια, το κείμενο x_j κατατάσσεται στην ανάλογη κατηγορία, εφ' όσον η τιμή του $w^T x_j^{(i)}$ είναι μεγαλύτερη από κατώφλι b_{k-1} και μικρότερη από κατώφλι b_k. Οι αποστάσεις που επιτρέπονται είναι οι προκαθορισμένες ξ (soft margins) . Η διαδικασία αυτή περιγράφεται από τις παρακάτω σχέσεις (Liu, 2011):

$$min\frac{1}{2}\|w\|^2 + \lambda\sum_{i=1}^{n}\sum_{j=1}^{m^{(i)}}\sum_{k=1}^{K-1}(\xi_{j,k}^{(i)} + \xi_{j,k+1}^{(i)*})$$

Όπου ο όρος $min\frac{1}{2}\|w\|^2$ εισάγεται για έλεγχο της πολυπλοκότητας του μοντέλου.Έτσι:

$$w^T x_j^{(i)} - b_{k-1} \leq -1 + \xi_{j,k}^{(i)}, \quad \varepsilon\alpha\nu \;\; y_j^{(i)} = k$$

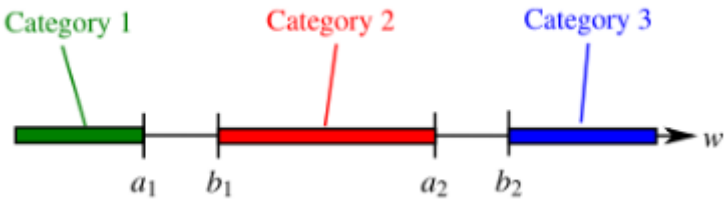
$$w^T x_j^{(i)} - b_{k-1} \geq -1 + \xi_{j,k+1}^{(i)*}, \quad \varepsilon\alpha\nu \;\; y_j^{(i)} = k+1$$

$$\xi_{j,k}^{(i)} \geq 0\,, \qquad \xi_{j,k+1}^{(i)*} \geq 0$$

- *Sum-of-margins:* Επιπλέον των παραπάνω, εισάγονται νέα κατώφλια a_k, τέτοια ώστε μια κατηγορία k να οριοθετείτε χαμηλά από το κατώφλι b_{k-1} και ψηλά από το κατώφλι a_k.Το παραπάνω μοντέλο, επομένως εκφράζεται ως (Liu, 2011):

$$min \sum_{k=1}^{K-1}(a_k - b_k) + \lambda \sum_{i=1}^{n}\sum_{j=1}^{m^{(i)}}\sum_{k=1}^{K-1}(\xi_{j,k}^{(i)} + \xi_{j,k+1}^{(i)*})$$

Όπου ο όρος $\sum_{k=1}^{K-1}(a_k - b_k)$ εκφράζει τις αποστάσεις μεταξύ των κατηγοριών, όπως φαίνεται στο σχήμα 7.



Σχήμα 7 Ο όρος $(b_k - a_k)$ εκφράζει ακριβώς την απόσταση μεταξύ των κατηγοριών k+1 και k.Πηγή: (Liu, 2011)

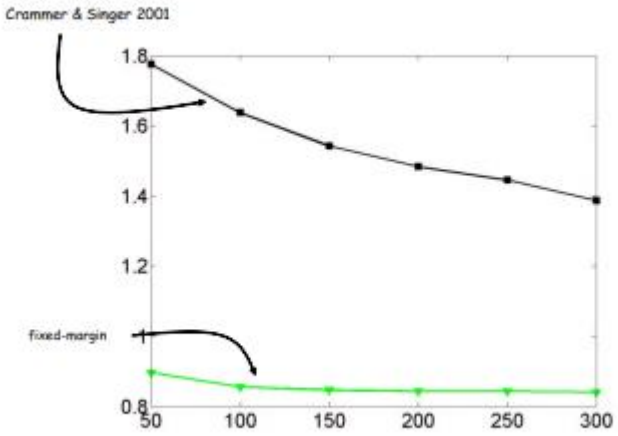
Έτσι:

$$w^T x_j^{(i)} \leq a_k + \xi_{j,k}^{(i)}, \quad \varepsilon\alpha\nu \;\; y_j^{(i)} = k$$

$$w^T x_j^{(i)} \geq b_k + \xi_{j,k+1}^{(i)*}, \quad \varepsilon\alpha\nu \;\; y_j^{(i)} = k+1$$

$$\|w\|^2 \leq 1, \quad \xi_{j,k}^{(i)} \geq 0\,, \qquad \xi_{j,k+1}^{(i)*} \geq 0$$

Για τον έλεγχο των επιδόσεων των παραπάνω μεθόδων, εξετάστηκαν από (Shashua & Levin, 2002) δεδομένα 1628 ταινιών, οι οποίες είχαν βαθμολογηθεί από 72916 χρήστες, με σκοπό να οργανωθούν σε κατηγορίες ("highly recommended","good"…."very bad") και να χρησιμοποιηθούν για να προβλέψουν τις βαθμολογίες ενός νέου χρήστη. Τα αποτελέσματα συγκρίθηκαν με αυτά του PRank. Τα αποτελέσματα εμφάνισαν μέσο σφάλμα 0,7,αριθμός σημαντικά βελτιωμένος σε σύγκριση με το μέσο σφάλμα του PRank, 1,25. Στην εικόνα 8, με πράσινη γραμμή φαίνονται τα προαναφερθέντα αποτελέσματα.



Εικόνα 7 Αποτελέσματα σύγκρισης PRank (μαύρη γραμμή) με fixed-margin (πράσινη γραμμή). Πηγή: (Shashua & Levin, 2002)

THRESHOLD-BASED LOSS FUNCTIONS (Rennie & Srebro, 2005)

Στην προηγούμενη ενότητα, υποθέσαμε γραμμική συνάρτηση σφάλματος. Στην παρούσα ενότητα, εξετάζονται δύο συναρτήσεις σφάλματος βασισμένες στο κατώφλι: σφάλμα άμεσου κατωφλιού (immediate-threshold loss) και σφάλμα όλων των κατωφλιών (all-threshold loss).

Έστω μια συνάρτηση βαθμολόγησης f(x) και μια συνάρτηση ποινής απόστασης(margin penalty) φ. Διαφορετικές τέτοιες συναρτήσεις φαίνονται στην εικόνα 8. Τότε, ορίζουμε το immediate-threshold loss ως:

$$L(f; x_j, y_j) = \varphi\left(f(x_j) - b_{y_j-1}\right) + \varphi(b_{y_j} - f(x_j)) \text{ (Liu, 2011)}$$

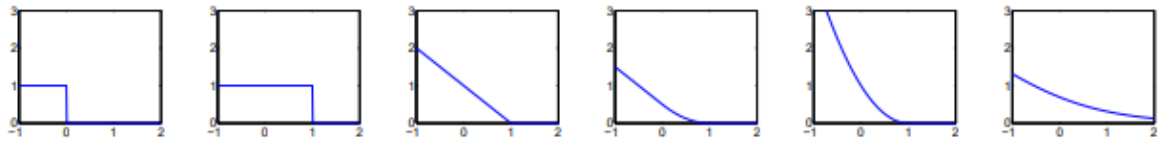
Όπου για κάθε ζευγάρι κειμένου-βαθμού (x_j, y_j) η σωστή κατηγορία θεωρούμε πως βρίσκεται μεταξύ των κατωφλιών b_{y_j-1}, b_{y_j}.

Το all-threshold loss ορίζεται ως:

$$L(f; x_j, y_j) = \sum_{k=1}^{K} \varphi((b_k - f(x_j))) \text{ (Liu, 2011)}$$

Όπου

$$s(k, y_j) = \begin{cases} -1, & k < y_j \\ +1, & k \geq y_j \end{cases}$$



Εικόνα 8 Διαφορετικές συναρτήσεις φ.Από αριστερά προς τα δεξιά: sign agreement, margin agreement, hinge, smooth hinge,modified least squares, logsitic. Πηγή: (Rennie & Srebro, 2005)

Πειραματικά, οι παραπάνω συναρτήσεις εξετάστηκαν με τα δεδομένα του MovieLens dataset από τους (Rennie & Srebro, 2005), παρόμοια με (Crammer & Springer, 2001) και (Shashua & Levin, 2002). Τα αποτελέσματά τους φαίνονται στην εικόνα 9 και δείχνουν βελτίωση στην περίπτωση του all-threshold σφάλματος.

	Multi-class Test MAE	Imm-Thresh Test MAE	All-Thresh Test MAE	
Mod. Least Squares	0.7486	0.7491	**0.6700**	(1.74e-18)
Smooth Hinge	0.7433	0.7628	**0.6702**	(6.63e-17)
Logistic	0.7490	0.7248	**0.6623**	(7.29e-22)
	Multi-class Test ZOE	Imm-Thresh Test ZOE	All-Thresh Test ZOE	
Mod. Least Squares	0.5606	0.5807	**0.5509**	(7.68e-02)
Smooth Hinge	0.5594	0.5839	**0.5512**	(1.37e-01)
Logistic	0.5592	0.5699	**0.5466**	(2.97e-02)

Εικόνα 9 Πειραματικά αποτελέσματα mean average loss-MAE και zero-on error-ZOE για τους παραπάνω αλγορίθμους σε σύγκριση με ταξινόμηση πολλαπλών τάξεων. Πηγή: (Rennie & Srebro, 2005)

ΣΥΜΠΕΡΑΣΜΑΤΑ

Το πρόβλημα της βαθμολόγησης κειμένων learning to rank έχει απασχολήσει πληθώρα ερευνητών, με το ενδιαφέρον να αυξάνεται ολοένα και περισσότερο τα τελευταία χρόνια. Ιδιαίτερα η αναζήτηση λύσης με τις μικρότερες δυνατές απώλειες και το ελάχιστο σφάλμα έχει πυροδοτήσει τεράστιο αριθμό ερευνών και προτάσεων για ολοένα και πιο αποτελεσματικές μεθόδους και αλγορίθμους.

Όσον αφορά το pointwise approach. δύο βασικά προβλήματα που εντοπίζονται από (Liu, 2011). Κατ' αρχάς, η προσέγγιση δεν λαμβάνει υπ' όψην πως κάποια κείμενα δεν είναι συναφή με το ερώτημα, με αποτέλεσμα ερωτήματα που έχουν μεγάλο βαθμό συναφών κειμένων να επηρεάζουν σε μεγαλύτερο βαθμό τη συνάρτηση απώλειας. Δεύτερον, δεδομένου του ότι η συνάρτηση απώλειας δεν λαμβάνει υπ' όψην τη θέση ενός κειμένου στην ταξινόμηση, μπορεί να επηρεαστεί από χαμηλά κείμενα, τα οποία δεν επηρεάζουν το χρήστη στην τελική ταξινόμηση.

Βιβλιογραφία

Cossock, D., & Zhang, T. (2006). Subset Ranking Using Regression. *Proceedings of the 19th Annual Conference on Learning Theory*, (σσ. 605-619).

Crammer, K., & Springer, Y. (2001). Pranking with Ranking. *Advances in Neural Information Processing Systems 14*, (σσ. 641-647).

Gey, F. C. (1994). Inferring Probability of Relevance. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (σσ. 222-231). Dublin Ireland: Springer-Verlag.

Li, P., Burges, C. J., & Wu, Q. (2007). McRank: Learning to Rank Using Multiple. *Advances in Neural Information Processing Systems 20*, (σσ. 845-852).

Liu, T.-Y. (2011). *Learning to Rank for Information Retrieval.* Springer.

Nallapati, R. (2004). Discriminative Models for Information Retrieval. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (σσ. 64-71). Sheffield United Kingdom: Association for Computing Machinery.

Rennie, J. D., & Srebro, N. (2005). Loss Functions for Preference Levels:Regression with Discrete Ordered Labels. *IJCAI 2005 Multidisciplinary Workshop on Advances in Preference Handling* (σσ. 180-185). Edinburgh: ACM.

Shashua, A., & Levin, A. (2002). Ranking with Large Margin Principle: Two Approaches. *Advances in Neural Information Processing Systems 15*, (σσ. 973-944).

Veloso, A. A., Almeida, H. M., Gonçalves, M. A., & Meira , W. (2008). Learning to rank at query-time using association rules. *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (σσ. 267-274). Singapore Singapore: Association for Computing Machinery.

Yining Wang, L. W. (2013). *A Theoretical Analysis of NDCG Ranking Measures.*

Zhe Chao, T. Q.-Y.-F. (2007, June). Learning to rank: from pairwise approach to listwise approach. *ICML '07: Proceedings of the 24th international conference on Machine learning*.