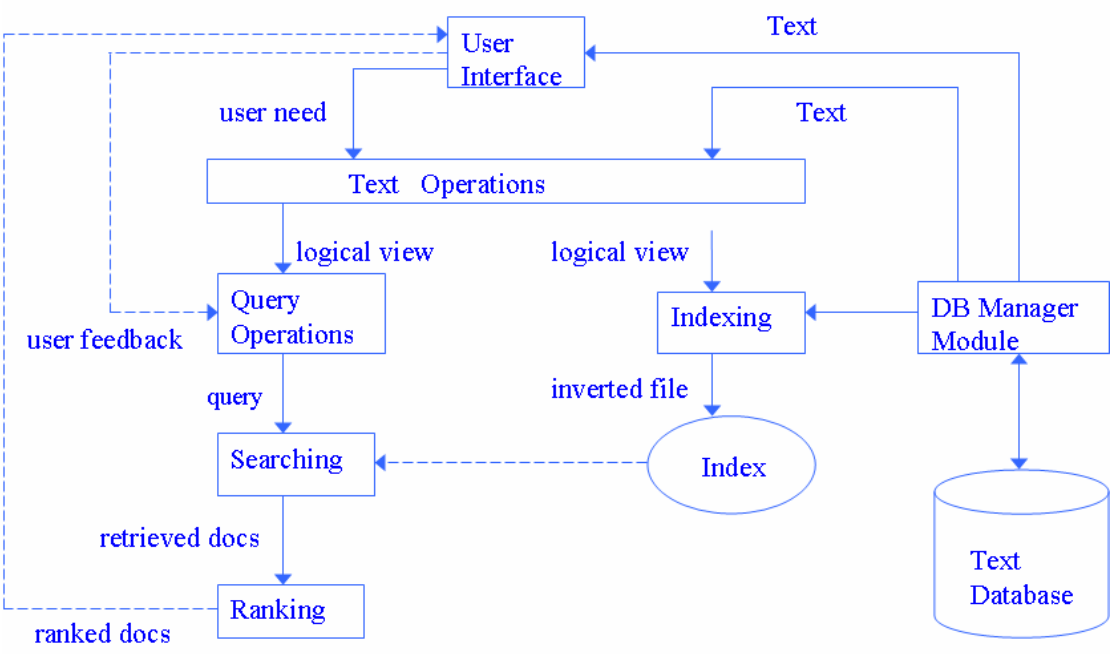
Vector Model vs. Boolean Model
CS-463 Information Retrieval
Spring 2006

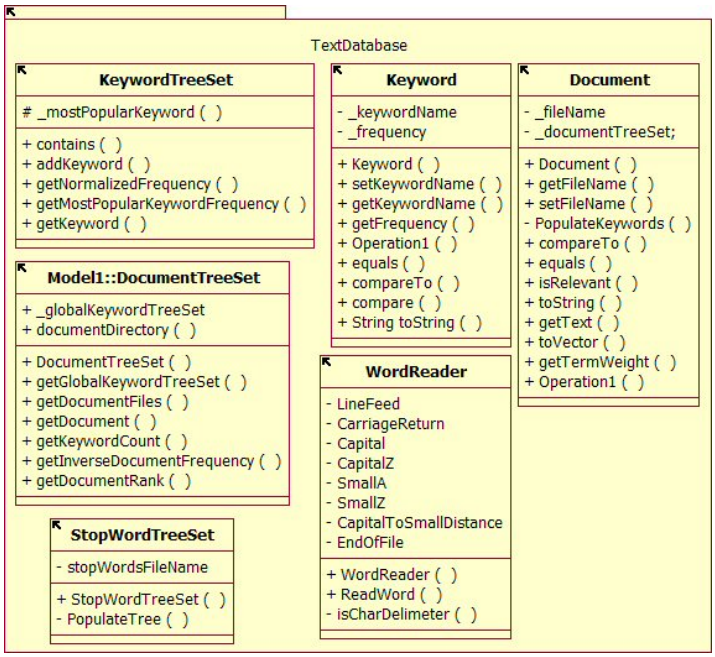
Costas Vandikas (vandikas@csd.uoc.gr)

Τεχνική παρουσίαση του εργαλείου που υλοποιεί μια Real-Time σύγκριση των δύο μοντέλων



Ένα τυπικό σύστημα ανάκτησης πληροφορίας αποτελείται από τα παραπάνω τμήματα. Για λόγους απλότητας η εφαρμογή μας εστιάζεται σε 3 μόνο από τα παραπάνω συστατικά. Τα τμήματα αυτά είναι, το Σύστημα Κειμένων (Text Database), ο μηχανισμός αναζήτησης (Searching) και η Διεπαφή του χρήστη (User Interface). Οι υπόλοιπες διαδικασίες εξαιτίας του μεγέθους της εφαρμογής συμπεριλαμβάνονται μέσα στα τρία αυτά πακέτα. Τα πακέτα αυτά υλοποιούνται ως εξής:

1. Σύστημα Κειμένων



Το σύστημα κειμένων υλοποιείται από τις παραπάνω 6 κλάσεις.

WordReader: Εκτελεί την ανάγνωση ενός αρχείου κειμένου, χαρακτήρα-χαρακτήρα απορρίπτοντας όλους τους χαρακτήρες εκτός από αυτούς που βρίσκονται μεταξύ των γραμμάτων a και z. Για διαχωριστικά μεταξύ των λέξεων χρησιμοποιούνται οι χαρακτήρες κενό, κόμμα, τελεία, παύλα και αλλαγή γραμμής. Επιπλέον τα κεφαλαία γράμματα μετατρέπονται σε μικρά. Οι απλοποιήσεις αυτές αλλοιώνουν το κείμενο όμως κατορθώνουνε να αντλήσουν όσο το δυνατό περισσότερα ουσιαστικά για να φτιάξουμε με αυτά τα keywords κάθε κειμένου.

StopWordTreeSet: Δενδρική δομή που περιέχει τα stop words που θα αγνοήσει η εφαρμογή όταν θα χτίζει τα index terms ενός document. Τα stop words βρίσκονται στον κατάλογο support και ο χρήστης μπορεί να τα τροποποιήσει. Συμφώνα με τις προδιαγραφές της εργασίας στα stop words υπάρχουν τα άρθρα the, a και an.

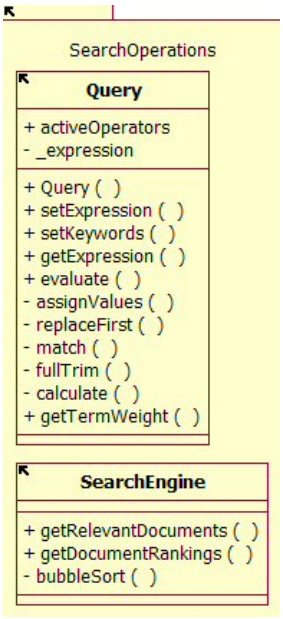
Keyword: Η κλάση αυτή καταχωρεί κάθε λέξη-κλειδί (ή αλλιώς index term) που βρίσκουμε σε ένα κείμενο. Επίσης καταχωρεί και την συχνότητα της λέξης. Το χαρακτηριστικό αυτό το εκμεταλλευόμαστε στο Διανυσματικό μοντέλο.

KeywordTreeSet: Δενδρική δομή που καταχωρεί keywords. Η δομή αυτή είναι υλοποιημένη έτσι ώστε όταν διαβάζουμε ένα keyword ελέγχεται εάν έχει ήδη καταχωρηθεί έτσι ώστε να βρεθεί η συχνότητα του. Επιπλέον ελέγχουμε την συχνότητα του για να βρούμε το keyword με την μέγιστη συχνότητα. Γνωρίζοντας την συχνότητα κάθε keyword και το keyword με την μέγιστη συχνότητα μπορούμε να υπολογίσουμε την κανονικοποιημένη συχνότητα του keyword.

Document: Βοηθητική κλάση που επεκτείνει την κλάση KeywordTreeSet έτσι ώστε να υλοποιήσει ένα κείμενο. Ένα κείμενο έχει ένα όνομα αρχείου και μια συλλογή από keywords. Ορίζοντας το όνομα αρχείου το Document ανοίγει το αντίστοιχο αρχείο από το σκληρό δίσκο και διαβάζει τα περιεχόμενα του βρίσκοντας έτσι τα keywords του.

DocumentTreeSet: Η δομή αυτή είναι η καθολική συλλογή κειμένων (Documents) και αντλεί τα κείμενα τις διαβάζοντας τα περιεχόμενα του κατάλογο documents. Για κάθε αρχείο που υπάρχει στον κατάλογο αυτό δημιουργεί ένα document με όνομα το όνομα του αρχείου. Το Document μετά φροντίζει να αποκτήσει τα keywords του. Επίσης διαθέτει μια συλλογή με όλα τα keywords που υπάρχουν σε όλα τα Documents (και σε πόσα Documents υπάρχουν). Η πληροφορία αυτή μας είναι απαραίτητα για τον υπολογισμό της αντίστροφης συχνότητας ενός keyword η οποία υλοποιείται σε αυτή την κλάση.

2. Διαδικασία Αναζήτησης



Ο φορέας των αιτημάτων της αναζήτησης είναι η κλάση **Query**: Η κλάση αυτή είναι ένα KeywordTreeSet το οποίο φροντίζει να βρει τις λέξεις κλειδιά που έδωσε ο χρήστης. Το αίτημα του χρήστη δέχεται την ίδια προεπεξεργασία με αυτή που δέχονται και τα κείμενα.

Η κλάση **SearchEngine** διαθέτει δύο μεθόδους. Η μέθοδος **getRelevantDocuments** υλοποιεί το Λογικό μοντέλο ενώ η μέθοδος **getDocumentRankings** το Διανυσματικό.

Το **μοντέλο Boolean** υλοποιείται από τη μέθοδο isRelevant του Document. Η διαδικασία ξεκινάει από την getRelevantDocuments και φροντίζει να δώσει το Query σε κάθε κείμενο και να δει εάν το κείμενο είναι σχετικό ή όχι με το Query. Από την θεωρία βρίσκουμε το βαθμό της ομοιότητας ενός κειμένου από τον τύπο :

$$sim(d_j, q) = \begin{cases} 1 & \alpha\nu\ \exists \vec{q}_{cc} \,|\, (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall k_i, g_i(\vec{d}_j) = g_i(\vec{q}_{cc})) \\ 0 & \end{cases}$$

Ο τύπος αυτός μας λέει πως αν υπάρχει ένα τμήμα της κανονικής διαζευκτικής μορφής του Query και το τμήμα αυτό ανήκει στην κανονική διαζευκτική μορφή του Query και για κάθε Keyword του συστήματος το βάρος του Keyword στο κείμενο είναι ίσο με το βάρος του Keyword στο Query. Με άλλα λόγια ο βαθμός ομοιότητας είναι 1 όταν τα keywords που δεν υπάρχουν στο κείμενο δεν υπάρχουν στο Query και τα Keywords που υπάρχουν στο Query υπάρχουν και στο κείμενο. Εάν ο βαθμός ομοιότητας είναι 1 το κείμενο είναι σχετικό ενώ εάν είναι 0 το κείμενο δεν είναι σχετικό. Στην εργασία μας για λόγους απλότητας αλλά και ευελιξίας ακολουθήσαμε μια διαφορετική τακτική. Η τακτική αυτή υλοποιείται από τη συνάρτηση evaluate της **Query**. Συγκεκριμένα δημιουργούμε ένα string που έχει τα keywords του Query καθώς και τους τελεστές που όρισε ο χρήστης. Έπειτα περνάμε αυτό το String από κάθε κείμενο (μέσα από τη συνάρτηση της Document - isRelevant) και αντικαθιστούμε κάθε keyword του Query με το βάρος του στο κείμενο. Το βάρος αυτό είναι μια δυαδική τιμή που είναι 1 αν υπάρχει το keyword στο κείμενο και 0 εάν δεν υπάρχει. Έπειτα στο string αυτό υπολογίζουμε τους δυαδικούς τελεστές με προτεραιότητα στο AND και μετά στο OR.

Παράδειγμα :

Query: hello |world
Κείμενο 1: 0 | 0 (αποτέλεσμα 0)
Κείμενο 2: 1 | 1 (αποτέλεσμα 1) (το κείμενο 2 έχει βαθμό σχετικότητας 1)
Κείμενο 3: 0 | 1 (αποτέλεσμα 1) (το κείμενο 3 έχει βαθμό σχετικότητας 1)

Το αποτέλεσμα που βρίσκουμε με την εφαρμογή των τελεστών είναι ο τελικός βαθμός ομοιότητας του κειμένου. Η μεθοδολογία αυτή μας επιτρέπει να έχουμε συνδυασμούς τελεστών μέσα στο Query μας και μπορεί να επεκταθεί έτσι ώστε να χειρίζεται και παρενθέσεις. (η δυνατότητα χειρισμού παρενθέσεων δεν υπάρχει στην εφαρμογή μας αυτή την στιγμή). Είναι ισοδύναμη με τον παραπάνω τύπο γιατί πετυχαίνει το ίδιο αποτέλεσμα δίνοντας σημασία μόνο στα Keywords που υπάρχουν (τα keywords που δεν υπάρχουν μας δίνουν 0 (το 0 αποτελεί ουδέτερο στοιχείο στην κανονική διαζευκτική μορφή και επομένως μπορεί να παραληφθεί).

Το **μοντέλο Vector** υλοποιείται από την μέθοδο getDocumentRank του **DocumentTreeSet**. Η μέθοδος αυτή υλοποιεί τον τύπο εύρεσης του βαθμού ομοιότητας του κειμένου με το Query

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \times \left|\vec{q}_j\right|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} {w_{i,j}}^2} \times \sqrt{\sum_{i=1}^{t} {w_{i,q}}^2}}$$

Ο τύπος αυτός μας λέει πως για να βρούμε τον βαθμό ομοιότητας ενός κειμένου με ένα Query θα πρέπει για κάθε Keyword του συστήματος να βρούμε το άθροισμα του γινομένου του βάρος του keyword αυτού στο κείμενο j με το βάρος του keyword στο query και αυτό να το διαιρέσουμε με το μέτρο του διανύσματος του document επί το μέτρο του διανύσματος του Query. Το μέτρο του διανύσματος του κειμένου υπολογίζεται από το άθροισμα των τετραγώνων των βαρών για κάθε keyword του συστήματος στο κείμενο. Το μέτρο του διανύσματος του Query υπολογίζεται από το άθροισμα των τετραγώνων των βαρών για κάθε keyword του συστήματος στο Query. Για να υπολογίσουμε το βάρος ενός keyword σε ένα κείμενο χρησιμοποιούμε τον τύπο :

$$w_{i,j} = f_{i,j} \times idf_i \text{ όπου } f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \text{ και } idf_i = \log\frac{N}{n_i}$$

Η $f_{i,j}$ ονομάζεται κανονικοποιημένη συχνότητα και υπολογίζεται από το πηλίκο της συχνότητας του keyword στο κείμενο με την μέγιστη συχνότητα (την συχνότητα της λέξης που εμφανίζεται περισσότερες φορές στο κείμενο). Ο λόγος που χρησιμοποιούμε αυτό τον τύπο είναι για να εξαλείψουμε την ανομοιογένεια που δημιουργείται μεταξύ ενός μεγάλου και ενός μικρού κειμένου. Ένα μεγάλο κείμενο μπορεί να έχει μια keyword που εμφανίζεται 10 φορές όμως ένα μικρό δεν έχει. Αν χρησιμοποιήσουμε την ωμή συχνότητα βλέπουμε πως η λέξη αυτή είναι σημαντική στο πρώτο κείμενο όμως αυτό είναι λάθος γιατί η λέξη μπορεί να είναι εξίσου σημαντική και στο μικρό κείμενο.

λέξη	**συχνότητα $freq_{i,j}$**	**συχνότητα της ποιο συχνής keyword $max_l freq_{l,j}$**	**κανονικοποιημένη συχνότητα $f_{i,j}$**
Spell	1	5	0.2
the	5	5	1
game	3	5	0.6

Από τον πίνακα βλέπουμε πως οι λέξεις που εμφανίζονται σπάνια έχουν μικρή κανονικοποιημένη συχνότητα ενώ αυτές που εμφανίζονται συχνά έχουν μεγάλη. Ο υπολογισμός της κανονικοποιημένης συχνότητας υλοποιείται από τη συνάρτηση getNormalizedFrequency() του KeywordTreeSet. Θυμίζουμε πως το Document είναι KeywordTreeSet όπως και το Query επομένως έχουμε τη δυνατότητα (μέσω υπέρβασης της συνάρτησης κατά την κληρονομικότητα) να υπολογίσουμε τη κανονικοποιημένη συχνότητα τόσο για ένα keyword του Document όσο και για ένα keyword του Query.

Το idf ονομάζεται αντίστροφη συχνότητα και ο στόχος του είναι να δείξει πόσο σημαντικό είναι το keyword σε ολόκληρη τη συλλογή κειμένων. Για να το βρούμε αυτό υπολογίζουμε το πηλίκο του πλήθους των κειμένων με το αριθμό των κείμενων στο οποίο εμφανίζεται η λέξη. Η πληροφορία αυτή βρίσκεται στο globalKeywordTreeSet του DocumentTreeSet.

λέξη	**πλήθος κειμένων N**	**Πλήθος κειμένων στα οποία εμφανίζεται η λέξη ni**	**N/Ni**	**Log(N/Ni)**
Spell	20	5	4	0,6
the	20	20	1	0
game	20	15	1,33	0,12

Ο λόγος που χρησιμοποιούμε τον λογάριθμο είναι για να μηδενίσουμε το βάρος του keyword που εμφανίζεται σε όλα τα κείμενα. Το βάρος μιας keyword σε ένα κείμενο υπολογίζεται από την μέθοδο getTermWeight του Document.

Για τον υπολογισμό του βάρους μιας keyword στο Query χρησιμοποιούμε τον τύπο που προτάθηκε από τους Buckley και Salton

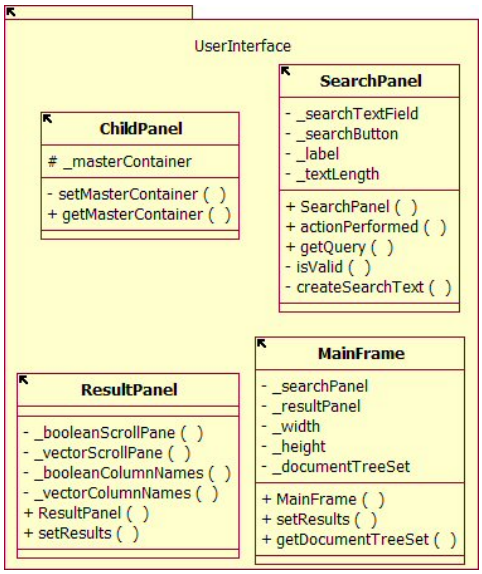
$$w_{i,q} = \left(0.5 + \frac{0.5\, freq_{i,q}}{\max_l freq_{l,q}}\right) \times \log\frac{N}{n_i}.$$

Ο τύπος αυτός μας λέει πως για να υπολογίσουμε το βάρος ενός keyword στο Query χρειάζεται να βρούμε την κανονικοποιημένη συχνότητα του keyword στο Query (με τον ίδιο τρόπο που το βρίσκουμε και στο κείμενο – η μέθοδος υλοποιείται στο KeywordTreeSet που είναι η κλάση από την οποία κληρονομεί το Query) και να το πολλαπλασιάσουμε με την αντίστροφη συχνότητα του Keyword η οποία βρίσκεται από τη συλλογή των κειμένων. Επειδή η τιμή αυτή είναι μικρή προσθέτουμε το 0.5. Η μέθοδος αυτή υλοποιείται από την getTermWeight της Query.

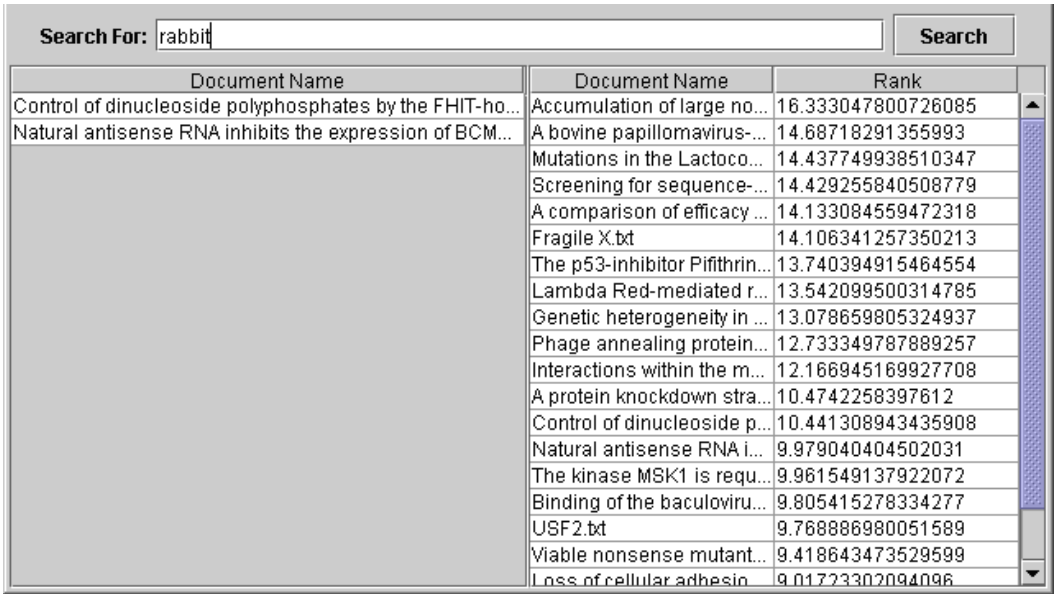
3. Διεπαφή Χρήστη

Η διεπαφή χρήστη αποτελείται από δύο μέρη. Το **SearchPanel** και το **ResultPanel**. Το **SearchPanel** είναι το επάνω μέρος της εφαρμογής και φροντίζει να τροφοδοτήσει το **Query** με το αίτημα του χρήστη ελέγχοντας πρώτα εάν το αίτημα του είναι ορθό. Ο τελεστής **AND** συμβολίζεται με το + ενώ ο τελεστής **OR** με το |. Ένα αίτημα αναζήτησης είναι ορθό όταν υπάρχουν keywords η μια μετά την άλλη χωρίς κανένα τελεστή (το σύστημα θεωρεί πως χρησιμοποιείται ο τελεστής **OR**) ή σε περίπτωση που οριστεί τελεστής αυτός θα πρέπει να υπάρχει σε κάθε λέξη από το δεύτερο κείμενο και μετά δηλαδή :

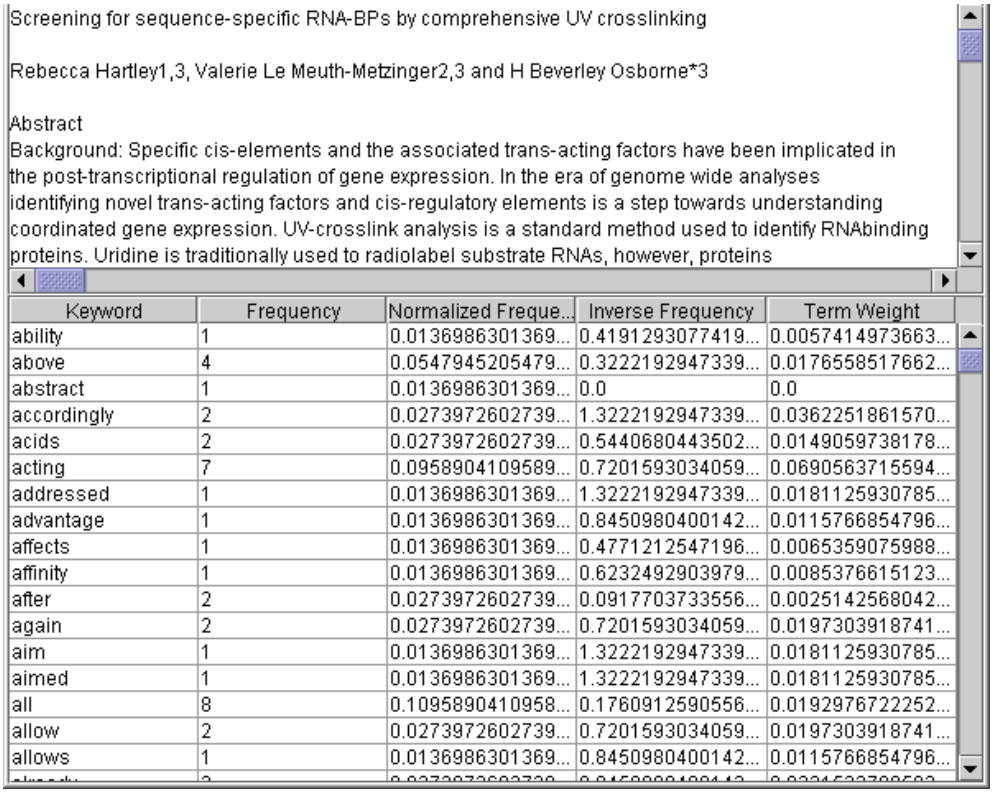
control |function (<=> control function)
control +function
+control +function (ΛΑΘΟΣ)
(το ΣΩΣΤΟ είναι control +function)
control +function something (ΛΑΘΟΣ)
(το ΣΩΣΤΟ είναι control +function (+ ή |)function



Το ResultPanel είναι το κάτω μέρος της οθόνης και αποτελείται από δύο πίνακες όπου ο αριστερός περιέχει τα αποτελέσματα της αναζήτησης με το λογικό μοντέλο και ο δεξής τα αποτελέσματα αναζήτησης με το διανυσματικό μοντέλο καθώς και τις βαθμολογίες τους.



Επιπλέον κάνοντας **διπλό-κλικ** πάνω σε κάποια γραμμή του πίνακα αποτελεσμάτων εμφανίζεται ένα παράθυρο που δείχνει τα keywords που βρεθήκανε για κάθε κείμενο, την συχνότητα τους, την κανονικοποιημένη συχνότητα την αντίστροφη συχνότητα, και το βάρος του keyword στο συγκεκριμένο κείμενο.


Screening for sequence-specific RNA-BPs by comprehensive UV crosslinking

Rebecca Hartley1,3, Valerie Le Meuth-Metzinger2,3 and H Beverley Osborne*3

Abstract
Background: Specific cis-elements and the associated trans-acting factors have been implicated in the post-transcriptional regulation of gene expression. In the era of genome wide analyses identifying novel trans-acting factors and cis-regulatory elements is a step towards understanding coordinated gene expression. UV-crosslink analysis is a standard method used to identify RNAbinding proteins. Uridine is traditionally used to radiolabel substrate RNAs, however, proteins

Keyword	Frequency	Normalized Freque...	Inverse Frequency	Term Weight
ability	1	0.0136986301369...	0.4191293077419...	0.0057414973663...
above	4	0.0547945205479...	0.3222192947339...	0.0176558517662...
abstract	1	0.0136986301369...	0.0	0.0
accordingly	2	0.0273972602739...	1.3222192947339...	0.0362251861570...
acids	2	0.0273972602739...	0.5440680443502...	0.0149059738178...
acting	7	0.0958904109589...	0.7201593034059...	0.0690563715594...
addressed	1	0.0136986301369...	1.3222192947339...	0.0181125930785...
advantage	1	0.0136986301369...	0.8450980400142...	0.0115766854796...
affects	1	0.0136986301369...	0.4771212547196...	0.0065359075988...
affinity	1	0.0136986301369...	0.6232492903979...	0.0085376615123...
after	2	0.0273972602739...	0.0917703733556...	0.0025142568042...
again	2	0.0273972602739...	0.7201593034059...	0.0197303918741...
aim	1	0.0136986301369...	1.3222192947339...	0.0181125930785...
aimed	1	0.0136986301369...	1.3222192947339...	0.0181125930785...
all	8	0.1095890410958...	0.1760912590556...	0.0192976722252...
allow	2	0.0273972602739...	0.7201593034059...	0.0197303918741...
allows	1	0.0136986301369...	0.8450980400142...	0.0115766854796...

Στον κατάλογο Application υπάρχουν τρία shortcut για να εκτελεστεί η εφαρμογή.

Το **Plain** εκτελεί την εφαρμογή με τις παραμέτρους που ορίζονται από το βιβλίο (Modern Information Retrieval).

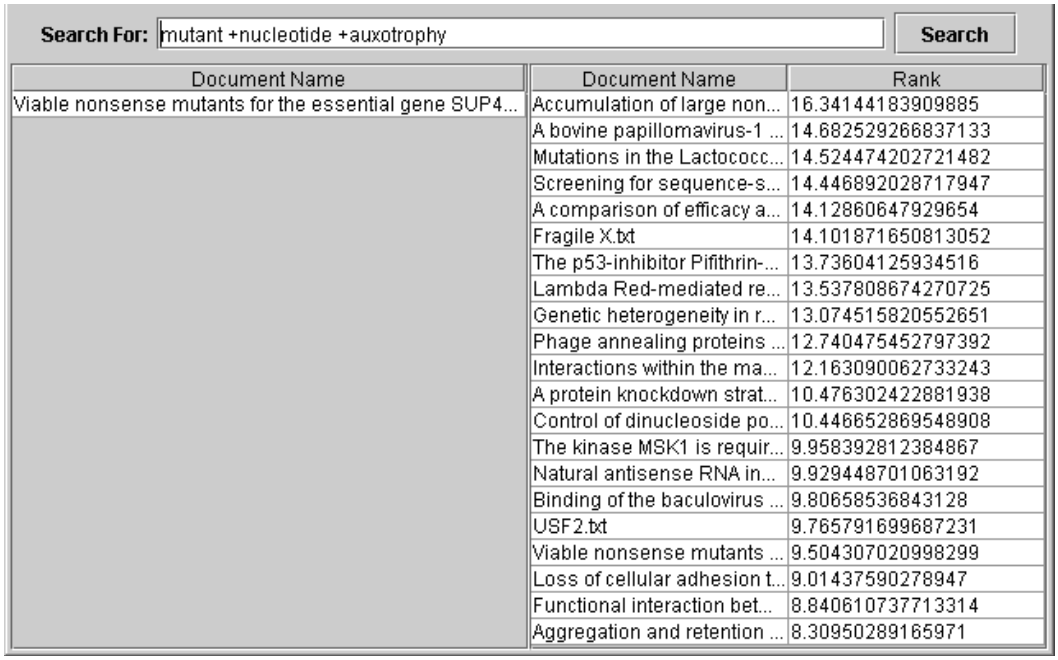
Το **Extra** εκτελεί την εφαρμογή χρησιμοποιώντας για τον υπολογισμό του **Query** τον τύπο

$$w_{i,q} = \frac{freq_{i,q}}{\max_l freq_{l,q}} \times \log \frac{N}{n_i}$$ (χωρίς το 0.5)

Το **Alternate** εκτελεί την εφαρμογή αντιπαραθέτοντας κάθε keyword του query μέσα στη συνάρτηση ομοιότητας και όχι με κάθε keyword του συστήματος.

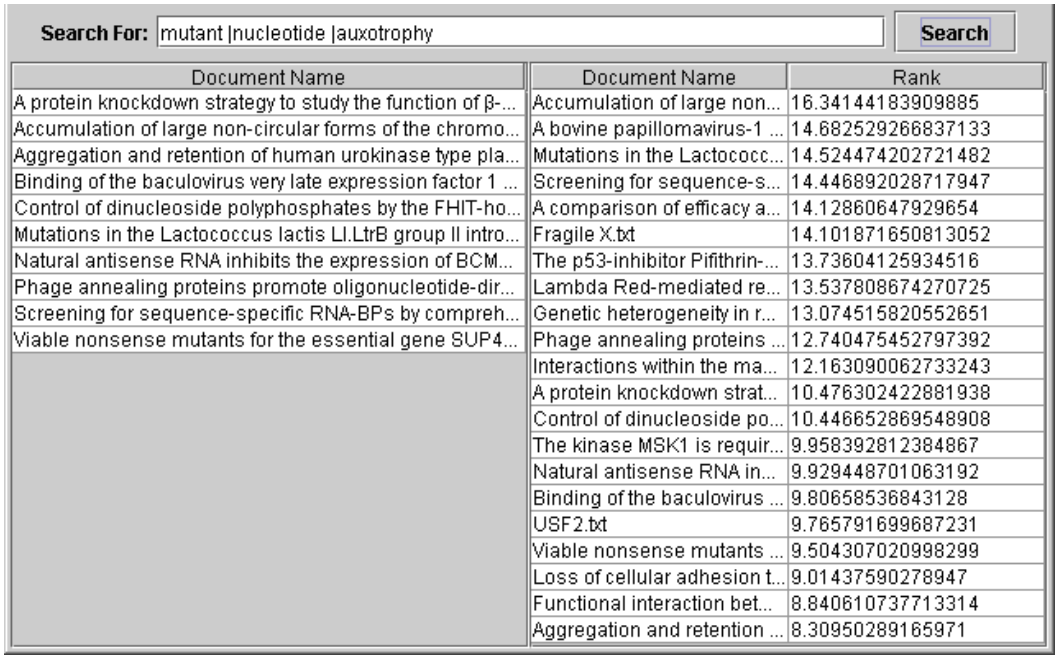
Ποιοτική σύγκριση Boolean – Vector Model

Έστω ότι εκτελούμε την αναζήτηση με τις λέξεις “mutant +nucleotide +auxotrophy” (mutant AND nucleotide AND auxotrophy). Το αποτέλεσμα της αναζήτησης φαίνεται στο screen shot που ακολουθεί.



Document Name	Document Name	Rank
Viable nonsense mutants for the essential gene SUP4...	Accumulation of large non...	16.341441839098 85
	A bovine papillomavirus-1 ...	14.682529266837133
	Mutations in the Lactococc...	14.524474202721482
	Screening for sequence-s...	14.446892028717947
	A comparison of efficacy a...	14.12860647929654
	Fragile X.txt	14.101871650813052
	The p53-inhibitor Pifithrin-...	13.73604125934516
	Lambda Red-mediated re...	13.537808674270725
	Genetic heterogeneity in r...	13.074515820552651
	Phage annealing proteins ...	12.740475452797392
	Interactions within the ma...	12.163090062733243
	A protein knockdown strat...	10.476302422881938
	Control of dinucleoside po...	10.446652869548908
	The kinase MSK1 is requir...	9.958392812384867
	Natural antisense RNA in...	9.929448701063192
	Binding of the baculovirus ...	9.80658536843128
	USF2.txt	9.765791699687231
	Viable nonsense mutants ...	9.504307020998299
	Loss of cellular adhesion t...	9.01437590278947
	Functional interaction bet...	8.840610737713314
	Aggregation and retention ...	8.30950289165971

Από την εικόνα βλέπουμε την αδυναμία του λογικού μοντέλου να κάνει μερική ταύτιση. Το κείμενο που επέστρεψε περιέχει και τους τρεις όρους που βάλαμε στην αναζήτηση. Αυτό είναι ένα γνωστό μειονέκτημα του λογικού μοντέλου και είναι η περίπτωση που επιστρέφει πολύ λίγα κείμενα. Αλλάζοντας τον τελεστή σε OR (mutant OR nucleotide OR auxotrophy) βλέπουμε πως το λογικό μοντέλο κάνει μερική ταύτιση με συνέπεια να επιστρέφει πάρα πολλά κείμενα.



Document Name	Document Name	Rank
A protein knockdown strategy to study the function of β-...	Accumulation of large non...	16.341441839098 85
Accumulation of large non-circular forms of the chromo...	A bovine papillomavirus-1 ...	14.682529266837133
Aggregation and retention of human urokinase type pla...	Mutations in the Lactococc...	14.524474202721482
Binding of the baculovirus very late expression factor 1 ...	Screening for sequence-s...	14.446892028717947
Control of dinucleoside polyphosphates by the FHIT-ho...	A comparison of efficacy a...	14.12860647929654
Mutations in the Lactococcus lactis LI.LtrB group II intro...	Fragile X.txt	14.101871650813052
Natural antisense RNA inhibits the expression of BCM...	The p53-inhibitor Pifithrin-...	13.73604125934516
Phage annealing proteins promote oligonucleotide-dir...	Lambda Red-mediated re...	13.537808674270725
Screening for sequence-specific RNA-BPs by compreh...	Genetic heterogeneity in r...	13.074515820552651
Viable nonsense mutants for the essential gene SUP4...	Phage annealing proteins ...	12.740475452797392
	Interactions within the ma...	12.163090062733243
	A protein knockdown strat...	10.476302422881938
	Control of dinucleoside po...	10.446652869548908
	The kinase MSK1 is requir...	9.958392812384867
	Natural antisense RNA in...	9.929448701063192
	Binding of the baculovirus ...	9.80658536843128
	USF2.txt	9.765791699687231
	Viable nonsense mutants ...	9.504307020998299
	Loss of cellular adhesion t...	9.01437590278947
	Functional interaction bet...	8.840610737713314
	Aggregation and retention ...	8.30950289165971

Από τα παραπάνω βλέπουμε πως η αποτελεσματικότητα του λογικού μοντέλου εξαρτάται πλήρως από την σύνταξη του Query που γράφει ο χρήστης. Δυστυχώς ο μέσος χρήστης δυσκολεύεται να συντάξει εύκολα ένα λογικό Query με αποτέλεσμα να ταλαντεύεται μεταξύ των πάρα πολλών και των πολύ

λίγων αποτελεσμάτων που επιστρέφει το λογικό μοντέλο. Σε ακραίες περιπτώσεις ίσως να χρειαστεί να κάνει δύο αναζητήσεις με παρόμοια κριτήρια και μετά να ενοποιήσει τα σύνολα των αποτελεσμάτων στο μυαλό του για να καταλήξει στο τελικό συμπέρασμα. Επομένως η απλότητα της υλοποίησης ενός συστήματος αναζήτησης (που είναι το σημαντικότερο πλεονέκτημα του μοντέλου) με το λογικό μοντέλο αντισταθμίζεται από την πολυπλοκότητα της διεπαφής χρήστη για την συγγραφή των Query. Το κυριότερο του μειονέκτημα είναι πως οι βαθμοί ομοιότητας είναι 1 ή 0 με αποτέλεσμα στην περίπτωση που μας επιστραφούνε περισσότερα από ένα κείμενα να μην ξέρουμε ποιο κείμενο είναι ποιο σημαντικό (δίνει μεγαλύτερη σημασία στα κριτήρια αναζήτησης μας).

Το διανυσματικό μοντέλο αγνοεί τους τελεστές και όπως φαίνεται από τα αποτελέσματα επιστρέφει ακριβώς τα ίδια κείμενα με τις ίδιες βαθμολογίες και για τα δύο Query.

Έστω ότι κάνουμε μια αναζήτηση με μια μόνο λέξη, την λέξη auxotrophy.

Search For: auxotrophy **Search**

Document Name	Document Name	Rank
Viable nonsense mutants for the essential gene SUP4...	Accumulation of large non...	16.330171571890475
	A bovine papillomavirus-1 ...	14.684596519426677
	Mutations in the Lactococc...	14.4352074692051
	Screening for sequence-s...	14.426714867003328
	A comparison of efficacy a...	14.130595741350795
	Fragile X.txt	14.103857148690006
	The p53-inhibitor Pifithrin-...	13.737975249487283
	Lambda Red-mediated re...	13.539714753906601
	Genetic heterogeneity in r...	13.076356670054523
	Phage annealing proteins ...	12.731107461271424
	Interactions within the ma...	12.164802586438942
	A protein knockdown strat...	10.472381341982304
	Control of dinucleoside po...	10.43045371121763
	The kinase MSK1 is requir...	9.959794921854904
	Natural antisense RNA in...	9.929227701930392
	Binding of the baculovirus ...	9.80368855724016
	USF2.txt	9.7671666915441
	Viable nonsense mutants ...	9.487248923164145
	Loss of cellular adhesion t...	9.015645097734666
	Functional interaction bet...	8.841855467085793
	Aggregation and retention ...	8.30241290084893

Το λογικό μοντέλο μας επιστρέφει μόνο ένα κείμενο ενώ το διανυσματικό μας επιστρέφει όλα τα κείμενα του συστήματος με τις βαθμολογίες τους και συγκεκριμένα δίνει την $18^{η}$ θέση στο κείμενο που μας επέστρεψε το λογικό μοντέλο. Ανοίγοντας τα κείμενα που βρίσκονται σε υψηλότερες θέσεις παρατηρούμε πως δεν διαθέτουνε την λέξη κλειδί που αναζητήσαμε και προφανώς η βαθμός τους είναι λάθος. Το πρόβλημα αυτό έγκειται στον τύπο

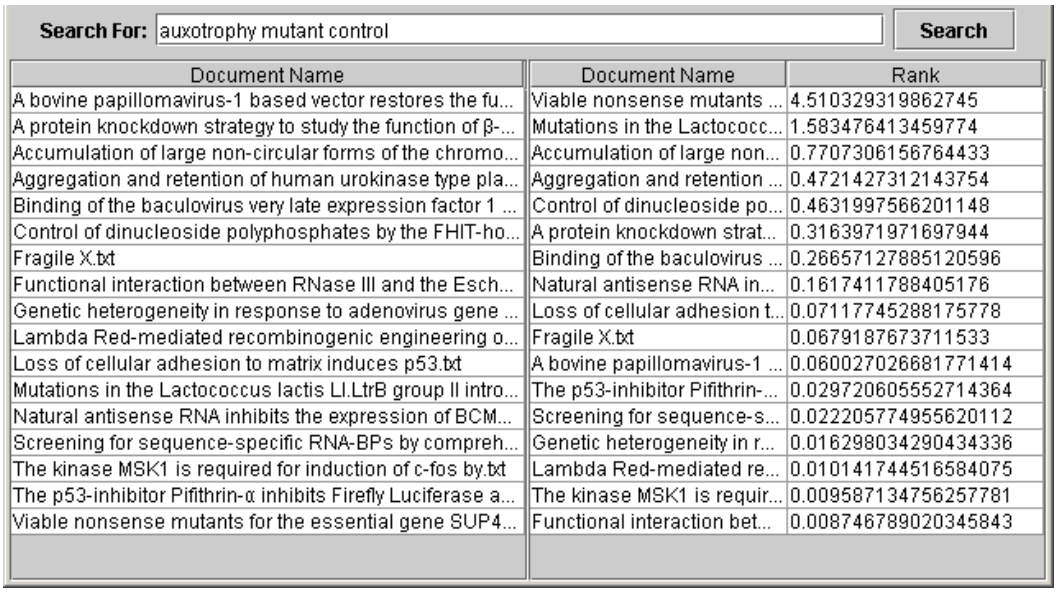
$$w_{i,q} = \left(0.5 + \frac{0.5\, freq_{i,q}}{\max_l \ freq_{l,q}} \right) \times \log \frac{N}{n_i}$$

και συγκεκριμένα στον παράγοντα 0.5. Παρατηρούμε πως αν δώσουμε σε αυτό τον τύπο ένα keyword που δεν υπάρχει στο Query αλλά υπάρχει στα Keywords του συστήματος το αποτέλεσμα ισούται με 0.5 επί την αντίστροφη συχνότητα του Keyword. Αυτό προσθέτει ένα βάρος στο διάνυσμα του Query το οποίο δεν θα έπρεπε να υπάρχει εφόσον η λέξη κλειδί απουσιάζει. Περνώντας έτσι κάθε λέξη έχουμε ένα σχετικά μεγάλο σε τιμή Query το οποίο αλλοιώνει τα αποτελέσματα μας.

Μια λύση στο πρόβλημα αυτό είναι να παραλείψουμε το 0.5. Εκτελώντας το πρόγραμμα στην έκδοση **Extra** έχουμε ένα ικανοποιητικό αποτέλεσμα.

Document Name	Document Name	Rank
Viable nonsense mutants for the essential gene SUP4...	Viable nonsense mutants ...	4.119507505670972

Συνεχίζοντας τις δοκιμές μας με το **Extra** μπορούμε πλέον να παρατηρήσουμε καθαρά τα πλεονεκτήματα του διανυσματικού μοντέλου έναντι στο λογικό. Τοποθετώντας στο Query τις λέξεις “auxotrophy muntant control” παίρνουμε την παρακάτω Screen Shoot.



Search For: auxotrophy mutant control — Search

Document Name	Document Name	Rank
A bovine papillomavirus-1 based vector restores the fu...	Viable nonsense mutants ...	4.510329319862745
A protein knockdown strategy to study the function of β-...	Mutations in the Lactococc...	1.583476413459774
Accumulation of large non-circular forms of the chromo...	Accumulation of large non...	0.7707306156764433
Aggregation and retention of human urokinase type pla...	Aggregation and retention ...	0.4721427312143754
Binding of the baculovirus very late expression factor 1 ...	Control of dinucleoside po...	0.4631997566201148
Control of dinucleoside polyphosphates by the FHIT-ho...	A protein knockdown strat...	0.3163971971697944
Fragile X.txt	Binding of the baculovirus ...	0.26657127885120596
Functional interaction between RNase III and the Esch...	Natural antisense RNA in...	0.1617411788405176
Genetic heterogeneity in response to adenovirus gene ...	Loss of cellular adhesion t...	0.07117745288175778
Lambda Red-mediated recombinogenic engineering o...	Fragile X.txt	0.0679187673711533
Loss of cellular adhesion to matrix induces p53.txt	A bovine papillomavirus-1 ...	0.060027026681771414
Mutations in the Lactococcus lactis LI.LtrB group II intro...	The p53-inhibitor Pifithrin-...	0.029720605552714364
Natural antisense RNA inhibits the expression of BCM...	Screening for sequence-s...	0.022205774955620112
Screening for sequence-specific RNA-BPs by compreh...	Genetic heterogeneity in r...	0.016298034290434336
The kinase MSK1 is required for induction of c-fos by.txt	Lambda Red-mediated re...	0.010141744516584075
The p53-inhibitor Pifithrin-α inhibits Firefly Luciferase a...	The kinase MSK1 is requir...	0.009587134756257781
Viable nonsense mutants for the essential gene SUP4...	Functional interaction bet...	0.008746789020345843

Παρατηρούμε πως τα αποτελέσματα είναι ίδια στο πλήθος (διότι από το λογικό μοντέλο χρησιμοποιήθηκε ο τελεστής OR) όμως αλλάζει η σειρά. Παρατηρούμε πως το πρώτο κείμενο είναι αυτό που έχει και τις τρεις λέξεις. Τα επόμενα κείμενα αξιολογούνται ανάλογο με την σημαντικότητα των υπολοίπων λέξεων. Από τις παρατηρήσεις μας εξάγουμε τον παρακάτω πίνακα για τα βάρη στο κείμενο. Θυμίζουμε πως υψηλή κανονικοποιημένη συχνότητα σημαίνει πως η λέξη είναι σημαντική για το κείμενο (η λέξη με την μεγαλύτερη συχνότητα έχει κανονικοποιημένη συχνότητα 1). Όσο μεγαλύτερη είναι η αντίστροφη συχνότητα τόσο ποιο σημαντική είναι η λέξη για την συλλογή των κειμένων (όμως μια λέξη που εμφανίζεται σε όλα τα κείμενα έχει αντίστροφη συχνότητα 0).

Κανονικοποιημένη Συχνότητα	**Αντίστροφη Συχνότητα**	**Βάρος στο Κείμενο**
Υψηλή	Υψηλή	Πολύ Υψηλό
Χαμηλή	Υψηλή	Μέτριο
Υψηλή	Χαμηλή	Πολύ Χαμηλό

Μέσα από τη βαθμολόγηση αυτή βλέπουμε πως το κείμενο που καταλαμβάνει τη δεύτερη θέση δίνει κάποια αξία στα συγκεκριμένα keywords (μεγαλύτερη από αυτή που δίνουνε τα υποδεέστερα κείμενα) επομένως ο χρήστης αξίζει να το δει το κείμενο γιατί μπορεί να είναι πραγματικά αυτό που ψάχνει. Πέρα από την βαθμολόγηση που διευκολύνει τον χρήστη στο να βρει το κείμενο που θεωρεί ποιο σημαντικό το διανυσματικό μοντέλο επιτρέπει από την φύση του την μερική ταύτιση χωρίς να ασχοληθεί ο χρήστης με την σύνταξη του Query. Μια ακόμα δυνατότητα του διανυσματικού μοντέλου είναι η αλλαγή των αποτελεσμάτων όταν βάζουμε μια λέξη περισσότερες από μία φορές.

Search For: auxotrophy mutant control control control mutant [Search]

Document Name	Document Name	Rank
A bovine papillomavirus-1 based vector restores the fu...	Viable nonsense mutants ...	4.43664639643151
A protein knockdown strategy to study the function of β-...	Mutations in the Lactococc...	2.746582836653309
Accumulation of large non-circular forms of the chromo...	Accumulation of large non...	1.3498625590380235
Aggregation and retention of human urokinase type pla...	Control of dinucleoside po...	0.8383768352211626
Binding of the baculovirus very late expression factor 1 ...	Aggregation and retention ...	0.8295018016940202
Control of dinucleoside polyphosphates by the FHIT-ho...	A protein knockdown strat...	0.5625935665060979
Fragile X.txt	Binding of the baculovirus ...	0.4869781079028116
Functional interaction between RNase III and the Esch...	Natural antisense RNA in...	0.34262531001833013
Genetic heterogeneity in response to adenovirus gene ...	Loss of cellular adhesion t...	0.18518883714970646
Lambda Red-mediated recombinogenic engineering o...	Fragile X.txt	0.1767104193936237
Loss of cellular adhesion to matrix induces p53.txt	A bovine papillomavirus-1 ...	0.1561777616181132
Mutations in the Lactococcus lactis LI.LtrB group II intro...	The p53-inhibitor Pifithrin-...	0.0773267960408133
Natural antisense RNA inhibits the expression of BCM...	Screening for sequence-s...	0.05777477944976173
Screening for sequence-specific RNA-BPs by compreh...	Genetic heterogeneity in r...	0.042404074547111544
The kinase MSK1 is required for induction of c-fos by.txt	Lambda Red-mediated re...	0.026386696877389506
The p53-inhibitor Pifithrin-α inhibits Firefly Luciferase a...	The kinase MSK1 is requir...	0.024943718343761388
Viable nonsense mutants for the essential gene SUP4...	Functional interaction bet...	0.02275731459739845

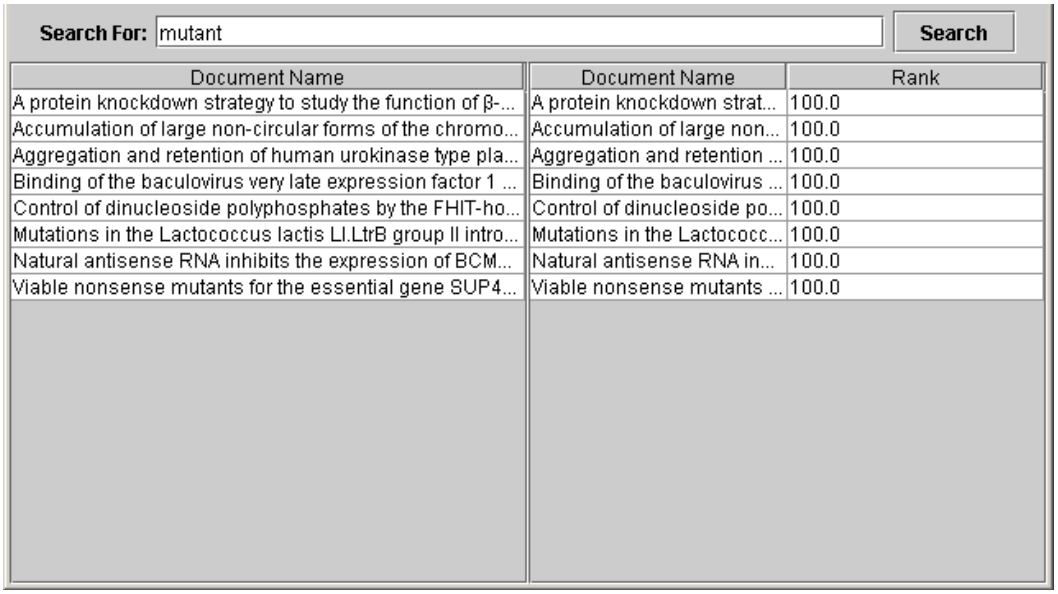
Αυτό έχει σαν αποτέλεσμα να αυξηθεί το βάρος που έχει μια λέξη στο Query και να τροποποιηθεί η τελική διάταξη των αποτελεσμάτων. Συγκρίνοντας τα δύο προηγούμενα αποτελέσματα βλέπουμε πως αλλάζει η θέση των κειμένων 4 και 5 και να έρχεται πρώτο το κείμενο "Control of dinucleoside" το οποίο στην προηγούμενη αναζήτηση έλαβε την $5^{η}$ θέση διότι το Keyword mutant είναι ποίο σημαντικό για το κείμενο "Aggregation and retention" από ότι στο κείμενο "Control of dinucleoside…"

Εναλλακτική προσέγγιση του Vector Model

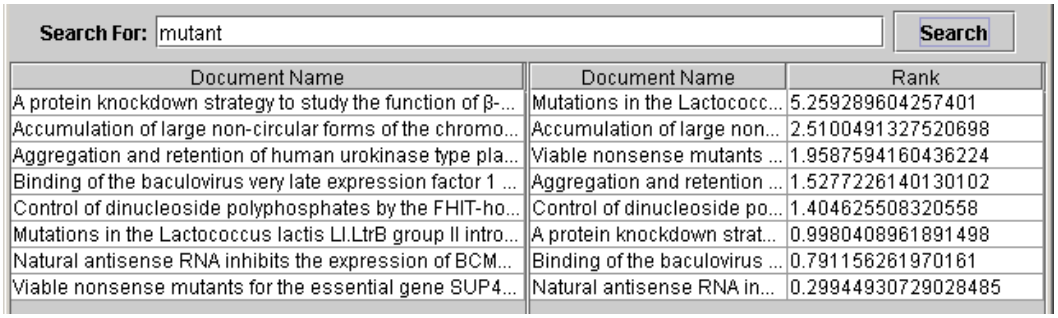
Τέλος παρουσιάζουμε μια εναλλακτική προσέγγιση του Vector Model που μας δίνει ενδιαφέροντα αποτελέσματα. Συγκεκριμένα εκτελώντας το πρόγραμμα **Alternate** η εφαρμογή εκτελείται όχι για κάθε keyword στο σύστημα αλλά για κάθε keyword στο Query. Αυτό έχει ως αποτέλεσμα να αυξάνεται εντυπωσιακά η ταχύτητα του σε βάρος της ποιότητας της βαθμολόγησης. Βλέπουμε πως ζητώντας μία λέξη οι βαθμολογίες που δίνει αγγίζουνε το 100% δηλαδή μας λέει πως οπωσδήποτε υπάρχει η λέξη αυτή μέσα στο κείμενο και είναι σημαντική για το κείμενο αυτό. Παρόλα αυτά εάν αντιπαραθέσουμε την μέθοδο αυτή με την **Extra** βλέπουμε πως η διάταξη είναι λανθασμένη και πως το κείμενο που έλαβε την $1^{η}$ θέση (Mutations in the Lactococc..) στο οποίο η λέξη mutant είναι πραγματικά σημαντική με βάρος (0.056) στο **Alternate** να λαμβάνει την ίδια θέση με άλλα κείμενα στα οποία η λέξη δεν είναι εξίσου σημαντική (έχουμε μικρότερο βάρος στο κείμενο). Ένα άλλο χαρακτηριστικό αυτής της μεθόδου είναι ότι χρησιμοποιώντας τον τύπο του Buckley και Salton για το Query Term Weight

$$w_{i,q} = \left(0.5 + \frac{0.5\, freq_{i,q}}{\max_l \ freq_{l,q}} \right) \times \log \frac{N}{n_i}$$

καταφέρνει και επιστρέφει κείμενα που όντως περιέχουν τη λέξη-κλειδί δηλαδή δεν παρουσιάζει το μειονέκτημα με το fudge-factor 0.5 που παρουσιάζει το αρχικό μας μοντέλο (η μορφή που περιγράφεται στο βιβλίο).



Εικόνα 1 - Alternate



Εικόνα 2 - Extra

Κείμενα

Στα πλαίσια της εργασία μας το σύνολο κείμενων αποτελείται από άρθρα του χώρου της μοριακής βιολογίας που αποκτήθηκαν από το δικτυακό χώρο http://www.biomedcentral.com και βρίσκονται στον κατάλογο documents

A bovine papillomavirus-1 based vector restores the function of the low-density lipoprotein receptor in .txt
A comparison of efficacy and toxicity between electroporation and adenoviral gene transfer.txt
A protein knockdown strategy to study the function of β-catenin in tumorigenesis.txt
Accumulation of large non-circular forms of the chromosome in recombination-defective mutants of Escheri.txt
Aggregation and retention of human urokinase type plasminogen activator in the yeast endoplasmic reticul.txt
Binding of the baculovirus very late expression factor 1 (VLF-1) to different DNA structures.txt
Control of dinucleoside polyphosphates by the FHIT-homologous HNT2 gene, adenine biosynthesis and heat s.txt
Fragile X.txt
Functional interaction between RNase III and the Escherichia coli ribosome.txt
Genetic heterogeneity in response to adenovirus gene therapy.txt
Interactions within the mammalian DNA methyltransferase family.txt
Lambda Red-mediated recombinogenic engineering of.txt
Loss of cellular adhesion to matrix induces p53.txt
Mutations in the Lactococcus lactis Ll.LtrB group II intron that retain mobility in vivo.txt
Natural antisense RNA inhibits the expression of BCMA, a tumour necrosis factor receptor homologue.txt
Phage annealing proteins promote oligonucleotide-directed mutagenesis in Escherichia coli and mouse ES c.txt
Screening for sequence-specific RNA-BPs by comprehensive UV crosslinking.txt
The kinase MSK1 is required for induction of c-fos by.txt
The p53-inhibitor Pifithrin-α inhibits Firefly Luciferase activity in vivo and in vitro.txt
USF2.txt
Viable nonsense mutants for the essential gene SUP45 of Saccharomyces cerevisiae.txt