HY-464:
Συστήματα Ανάκτησης Πληροφορίας
Information Retrieval Systems

Πανεπιστήμιο Κρήτης, Άνοιξη 2006

Φροντιστήριο 2

Θέμα : Retrieval Models

Ημερομηνία : 9 Μαρτίου 2006

Outline

- Previous Semester Exercises
  - Set Theory
  - Vector Model
  - Extended Boolean Model
- IR System Implementation
  - Approach
  - Boolean Model
  - Vector Model

Set Theory Exercise - Εκφώνηση

- Υποθέστε ένα μοντέλο ανάκτησης στο οποίο τα έγγραφα και οι επερωτήσεις είναι υποσύνολα του λεξιλογίου K. Έστω οι ακόλουθες τρεις συναρτήσεις κατάταξης:
  - $R1(d,q) = |d \cap q|/|q|$
  - $R2(d,q) = |d \cap q|/|d|$
  - $R3(d,q) = |d \cap q|/|d \cup q|$
- Σχολιάστε τις διαφορές των διατάξεων που προκύπτουν από αυτές τις τρεις συναρτήσεις και δικαιολογήστε την απάντηση σας.

Set Theory Exercise - Λύση

q = “a b”	$R_1(d,q) = \|d \cap q\|/\|q\|$	$R_2(d,q) = \|d \cap q\|/\|d\|$	$R_3(d,q) = \|d \cap q\|/\|d \cup q\|$
d_1 = “a”	1/2	1/1=1	1/2
d_2 = “a c”	1/2	1/2	1/3
d_3 = “a c d”	1/2	1/3	1/4
d_4 = “a b c”	2/2=1	2/3	2/3
Προκύπτουσα διάταξη εγγράφων	$<d_4, \{d_1,d_2,d_3\}>$	$<d_1, d_4, d_2\ d_3>$	$<d_4, d_1, d_2\ d_3>$

$|(a) \cap (a\ b)|/|(a\ b)| = |(a)|/|(a\ b)| = 1/2$

Set Theory - Συμπεράσματα

- Κάθε συνάρτηση δίνει διαφορετική διάταξη
- R_1
  - Το q είναι πάντα το ίδιο επομένως δεν λαμβάνονται υπόψη οι λέξεις του κάθε εγγράφου που δεν ταιριάζουν με την επερώτηση
- R_2
  - Ο παρανομαστής είναι πάντα διαφορετικός άρα λαμβάνει υπόψη το μέγεθος του κάθε αρχείου και κατ' επέκταση το ποσοστό του εγγράφου στο οποίο δεν έχουμε ταίριασμα
- R_3
  - Λαμβάνει υπόψη όχι μόνο το ποσοστό του εγγράφου στο οποίο δεν έχουμε ταίριασμα αλλά και το ποσοστό της επερώτησης στο οποίο δεν έγινε ταίριασμα.

Vector Model Exercise - Εκφώνηση

- Θεωρείστε μια συλλογή κειμένων που περιέχει τα ακόλουθα 5 έγγραφα:
  - Έγγραφο 1: «New York Times»
  - Έγγραφο 2: «New Times»
  - Έγγραφο 3: «Financial Times»
  - Έγγραφο 4: «High High Times»
  - Έγγραφο 5: «New Financial Times»

1. Δώστε τη διανυσματική παράσταση του κάθε εγγράφου με βάρη TF-IDF (για ευκολία θεωρήστε ότι IDF=N/DF και όχι IDF=log(N/DF)). Θεωρείστε ότι η θέση της κάθε λέξης στα διανύσματα γίνεται αλφαβητικά.
2. Θεωρείστε την επερώτηση q_1=«high financial». Υπολογίστε το TF-IDF διάνυσμα αυτής της επερώτησης και δώστε την διάταξη των εγγράφων που θα επιστρέψει ένα σύστημα που βασίζεται στο διανυσματικό μοντέλο.
3. Θεωρείστε τις επερωτήσεις q_3=«high AND financial» q_4=«high OR financial» και δώστε τις απαντήσεις που θα επιστρέψει ένα σύστημα που βασίζεται στο Extended Boolean μοντέλο.

Vector Model Exercise – Ερώτημα 1ο

	Financial	High	New	Times	York	$MAX_k\{FREQ_{ij}\}$
D_1			1	1	1	1
D_2			1	1		1
D_3	1			1		1
D_4		2		1		2
D_5	1		1	1		1
DF	2	1	3	5	1	
IDF	5/2	5/1=5	5/3	5/5=1	5/1=5	

- Term Occurrence Table
- $FREQ_{ij}$ = το πλήθος των εμφανίσεων του όρου i στο έγγραφο j
- N=5
- IDF=N/DF
- MAXk{FREQij} = συχνότητα της λέξης με τη μέγιστη συχνότητα στο κείμενο

Vector Model Exercise – Ερώτημα 1ο

	Financial	High	New	Times	York	$MAX_k\{FREQ_{ij}\}$
D_1			1/1*(5/3)=1,66	1/1*(5/5)=1	1/1*(5/1)=5	1
D_2			1/1*(5/3)=1,66	1/1*(5/5)=1		1
D_3	1/1*(5/2)=2,5			1/1*(5/5)=1		1
D_4		2/2*(5/1)=5		1/2*(5/5)=0,5		2
D_5	1/1*(5/2)=2,5		1/1*(5/3)=1,66	1/1*(5/5)=1		1
DF	2	1	3	5	1	
IDF	5/2=2,5	5/1=5	5/3=1,66	5/5=1	5/1=5	

- Term Weight Table
- $TF_{ij}=FREQ_{ij}/MAX_k\{FREQ_{ij}\}$
- $W_{ij}=TF_{ij}*IDF_i$

*Συχνότητα/MAXk{FREQij}*IDF*

Vector Model Exercise – Ερώτημα 2ο

	Financial	High	New	Times	York
Q2 = high financial	1*5/2=2,5	1*5/1=5			
IDF	5/2=2,5	5/1=5	5/3=1,66	5/5=1	5/1=5

- q1*d1=(5/2,5,0,0,0)*(0,0,5/3,1,5)=0
- q1*d2=(5/2,5,0,0,0)*(0,0,5/3,1,0)=0
- q1*d3=(5/2,5,0,0,0)*(5/2,0,0,1,0)=25/4
- q1*d4=(5/2,5,0,0,0)*(0,5,0,1/2,0)=25
- q1*d5=(5/2,5,0,0,0)*(5/2,0,5/3,1,0)=25/4

- Άρα η διάταξη των εγγράφων που θα επιστρέψει η ερώτηση Q_1 είναι:
  - D_4, D_3, D_5

Extended Boolean Model Exercise – Ερώτημα 3ο

	Financial	High	New	Times	York
D_1			1/1*(5/3)=1,66	1/1*(5/5)=1	1*(5/1)=5
D_2			1/1*(5/3)=1,66	1/1*(5/5)=1	
D_3	1/1*(5/2)=2,5			1/1*(5/5)=1	
D_4		2/2*(5/1)=5		1/2*(5/5)=0,5	
D_5	1/1*(5/2)=2,5		1/1*(5/3)=1,66	1/1*(5/5)=1	

- Κανονικοποίηση των διανυσμάτων maxIDFi=5:
  - d1'=(0,0,5/3,1,5)/5=(0,0,1/3,1/5,1)
  - d2'=(0,0,5/3,1,0)/5=(0,0,1/3,1/5,0)
  - d3'=(5/2,0,0,1,0)/5=(1/2,0,0,1/5,0)
  - d4'=(0,5,0,1/2,0)/5=(0,1,0,1/10,0)
  - d5'=(5/2,0,5/3,1,0)/5=(1/2,0,1/3,1/5,0)

Extended Boolean Model Exercise – Ερώτημα 3ο

- Q_3="high **AND** financial"
- Sim(q3,d1')=1-sqrt(((1-0)^2+(1-0)^2)/2)=0
- Sim(q3,d2')=1-sqrt(((1-0)^2+(1-0)^2)/2)=0
- Sim(q3,d3')=1-sqrt(((1-1/2)^2+(1-0)^2)/2)=0.21
- Sim(q3,d4')=1-sqrt(((1-0)^2+(1-1)^2)/2)=0.29
- Sim(q3,d5')=1-sqrt(((1-1/2)^2+(1-0)^2)/2)=0.21
- Άρα η διάταξη των εγγράφων που θα επιστρέψει η ερώτηση Q_3 είναι:
  - D_4, D_3, D_5

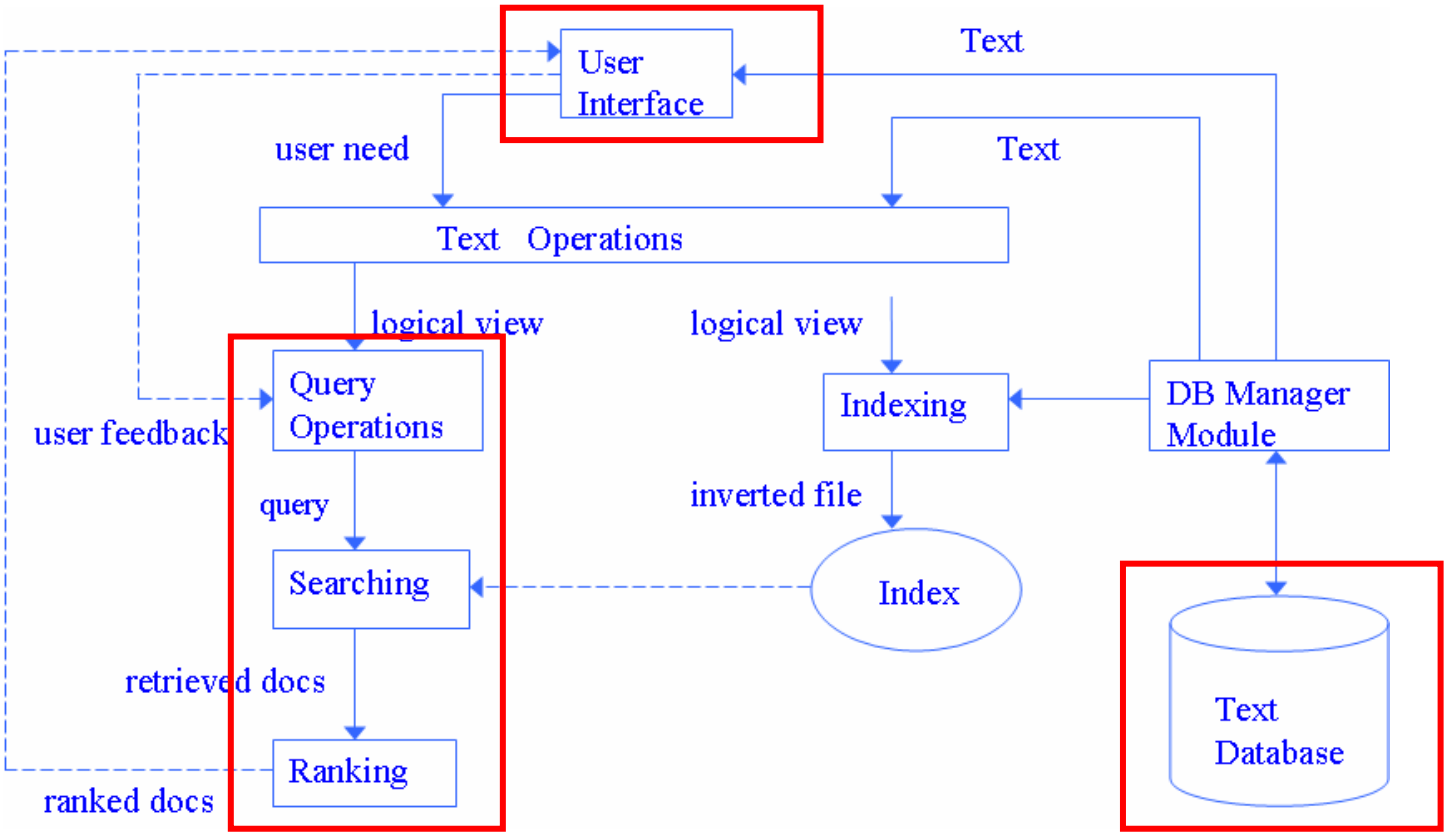
$$sim(q_{AND}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

Extended Boolean Model Exercise – Ερώτημα 3ο

- Q_4=”high **OR** financial”
- Sim(q4,d1’)=sqrt((0^2+0^2)/2)=0
- Sim(q4,d2’)=sqrt((0^2+0^2)/2)=0
- Sim(q4,d3’)=sqrt(((1/2)^2+0^2)/2)=1/(2sqrt(2))
- Sim(q4,d4’)=sqrt((0^2+1^2)/2)=1/(2sqrt(2))
- Sim(q4,d5’)=sqrt(((1/2)^2+0^2)/2)=1/(2sqrt(2))
- Άρα η διάταξη των εγγράφων που θα επιστρέψει η ερώτηση Q_4 είναι:
  - D_4, D_3, D_5

$$sim(q_{OR}, d) = \sqrt{\frac{x^2 + y^2}{2}}$$

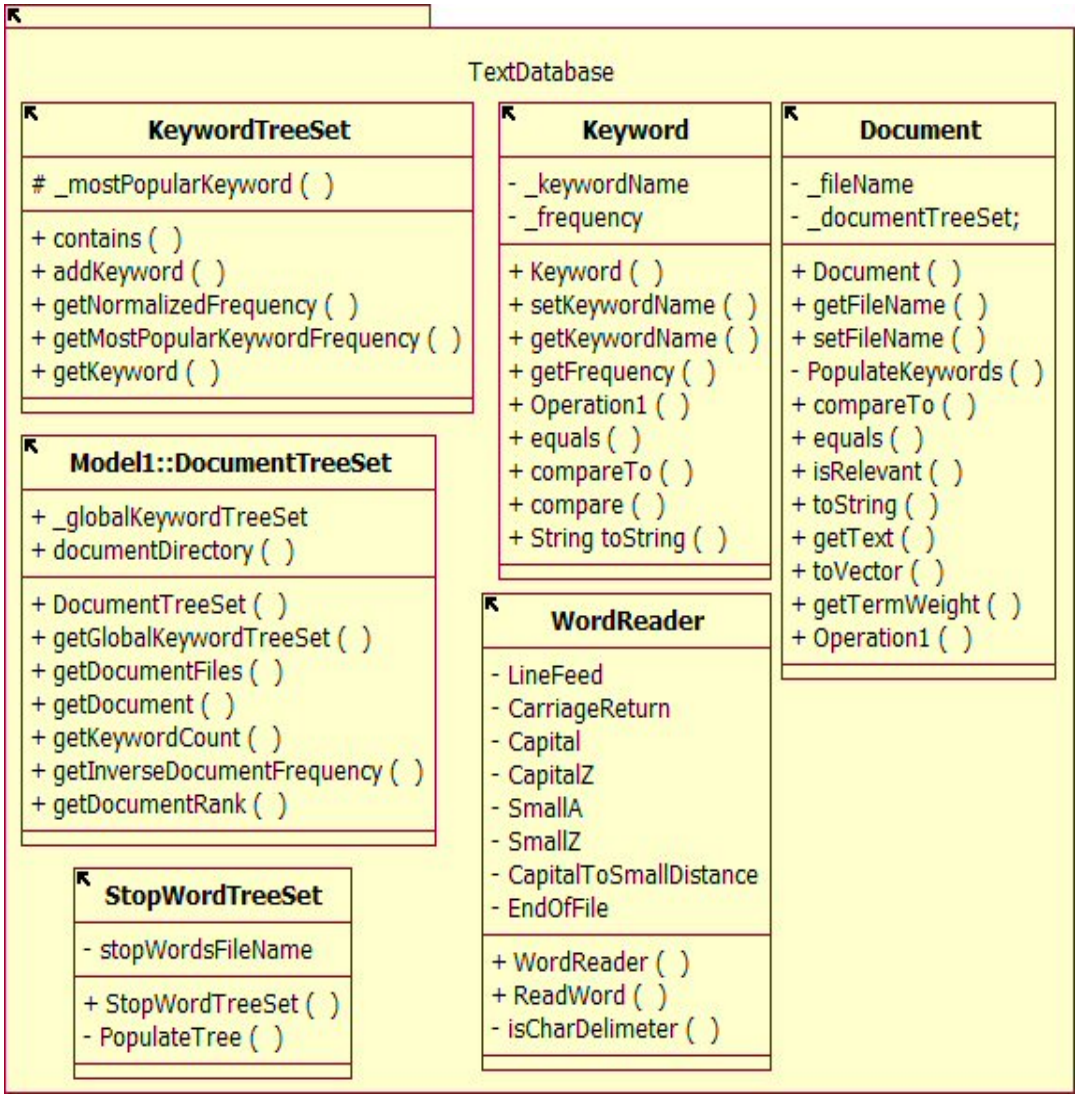
IR System Implementation



- A quick starter's approach
- This sketch is retrieval model independent
- Does not have a crawler (we suppose that the Text Database is filled somehow)

IR System Implementation

- TextDatabase
  - **WordReader**
  - **StopWordTreeSet**
  - **Keyword**
  - **KeywordTreeSet**
  - **Document**
  - **DocumentTreeSet**



IR System Implementation

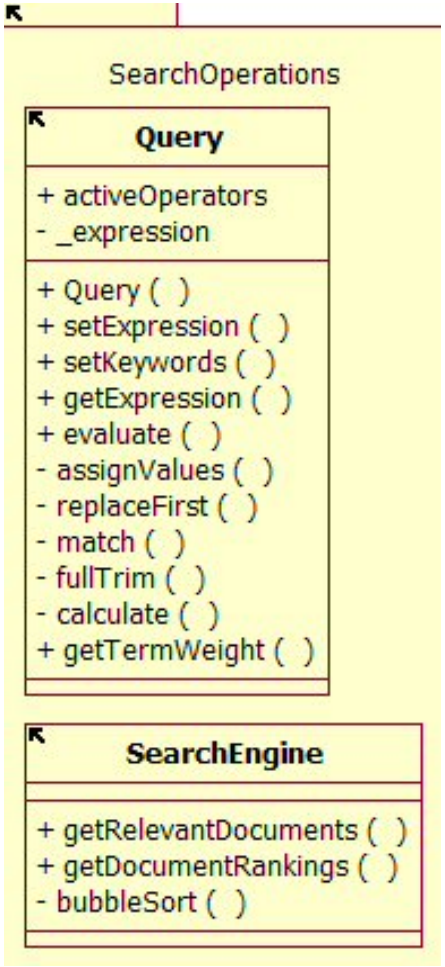
- SearchEngine
- getRelevantDocuments:
  - Boolean model

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

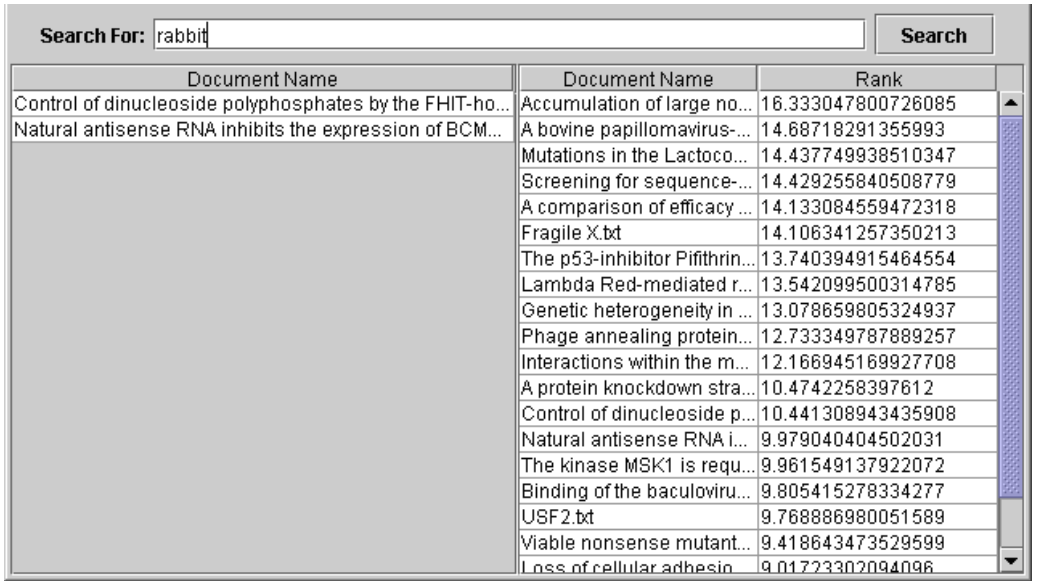
- getDocumentRankings

$$w_{i,j} = f_{i,j} \times idf_i \qquad f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}}$$

$$idf_i = \log \frac{N}{n_i} \qquad sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}_j|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$



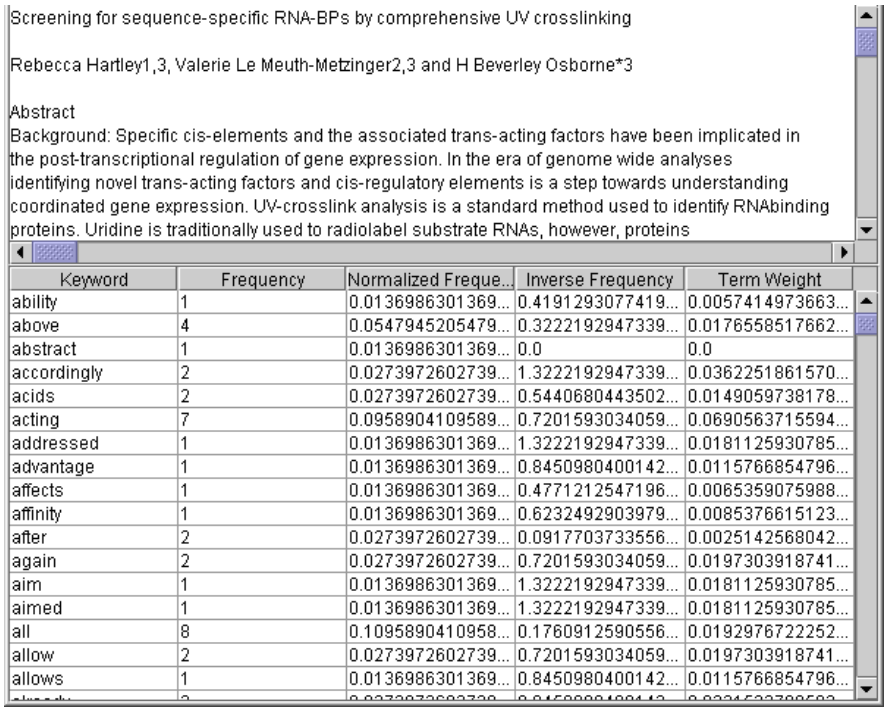
Demonstration



Search For: rabbit

Search

Document Name	Document Name	Rank
Control of dinucleoside polyphosphates by the FHIT-ho...	Accumulation of large no...	16.333047800726085
Natural antisense RNA inhibits the expression of BCM...	A bovine papillomavirus-...	14.68718291355993
	Mutations in the Lactoco...	14.437749938510347
	Screening for sequence-...	14.429255840508779
	A comparison of efficacy ...	14.133084559472318
	Fragile X.txt	14.106341257350213
	The p53-inhibitor Pifithrin...	13.740394915464554
	Lambda Red-mediated r...	13.542099500314785
	Genetic heterogeneity in ...	13.078659805324937
	Phage annealing protein...	12.733349787889257
	Interactions within the m...	12.166945169927708
	A protein knockdown stra...	10.4742258397612
	Control of dinucleoside p...	10.441308943435908
	Natural antisense RNA i...	9.979040404502031
	The kinase MSK1 is requ...	9.961549137922072
	Binding of the baculoviru...	9.805415278334277
	USF2.txt	9.768886980051589
	Viable nonsense mutant...	9.418643473529599
	Loss of cellular adhesio...	9.01723302094096



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...

Implementation Insight

- The main idea in information retrieval systems is the attempt to vectorize the information by quantifying it's structural elements.
- Most vectorization processes run iteratively
- This means that you can take advantage of each iterative cycle and store information that may be reused later.

Questions