ePublishing

# Ανοικτή Εκπαίδευση: το περιοδικό για την Ανοικτή και εξ Αποστάσεως Εκπαίδευση και την Εκπαιδευτική Τεχνολογία

Τόμ. 8, Αρ. 1 (2012)

Ανοικτή Εκπαίδευση

## Αναζήτηση στο Διαδίκτυο με Χρήση Μεθόδων Γραμμικής Άλγεβρας

*Ανδρέας Αρβανιτογεώργος*

doi: 10.12681/jode.9793

**Βιβλιογραφική αναφορά:**

https://epublishing.ekt.gr | **e-Εκδότης:** ΕΚΤ | **Πρόσβαση:** 12/10/2024 09:43:15

*Open Education - The Journal for Open and Distance Education and Educational Technology*
*Volume 8, Number 1, 2012 Section one. © Open Education ISSN: 1791-9312*

# Αναζήτηση στο Διαδίκτυο με Χρήση Μεθόδων Γραμμικής Άλγεβρας

# Internet Search using methods of Linear Algebra

**Ανδρέας Αρβανιτογεώργος**
Πανεπιστήμιο Πατρών
Επίκουρος Καθηγητής
arvanito@math.upatras.gr

**Περίληψη**
Τα τεχνολογικά προβλήματα που παρουσιάζονται στο σύγχρονο κόσμο είναι ιδιαίτερα περίπλοκα και προκειμένου να κατανοηθούν πληρέστερα, συχνά απαιτούν μια μαθηματική περιγραφή. Ένα χαρακτηριστικό παράδειγμα είναι η αναζήτηση ιστοσελίδων στο διαδίκτυο με τέτοιον τρόπο, ώστε να παρουσιάζονται στο χρήστη οι ιστοσελίδες που τον αφορούν περισσότερο. Ο τεράστιος όγκος της διαθέσιμης πληροφορίας απαιτεί μια συντονισμένη μαθηματική προσέγγιση και μια τέτοια μαθηματική μέθοδος χρησιμοποιείται από τη μηχανή αναζήτησης Google. Στην παρούσα εργασία παρουσιάζουμε τη θεμελίωση της μεθόδου αυτής χρησιμοποιώντας γραμμική άλγεβρα, καθώς και μερικές δυσκολίες που προκύπτουν κατά την εφαρμογή της.

**Abstract**
The success of several internet search machines derives in large part from various information retrieval (IR) methods that have been developed. The most frequently cited Web IR methods are the HITS (Hypertext Induced Topic Search), PageRank, and SALSA (Stochastic Approach for Link Structure Analysis). These methods use a vector space model, in particular finding an eigenvector with corresponding eigenvalue 1 of an 8 billion by 8 billion matrix. In the present work we give the mathematical formulation of the PageRank method.

**Keywords**
*information retrieval, pagerank, linear algebra*

## 1. Εισαγωγή

Η αναζήτηση πληροφοριών από το διαδίκτυο αποτελεί εδώ και καιρό μια βασική δραστηριότητα των περισσότερων ανθρώπων. Οι πληροφορίες αυτές βρίσκονται εντελώς διάσπαρτες και εκτός του ότι δεν υπόκεινται σε κάποια διαδικασία κρίσης, πολλές από αυτές είναι άχρηστες στο χρήστη. Για το σκοπό αυτό έχουν αναπτυχθεί διάφορες μέθοδοι, οι οποίες έχουν ως στόχο να παρουσιάζουν, ανάλογα με τις λέξεις-κλειδιά που χρησιμοποιεί κάποιος, τις ιστοσελίδες σε φθίνουσα σειρά ενδιαφέροντος. Οι μέθοδοι αυτές είναι γνωστές ως "μέθοδοι ανάκτησης πληροφορίας" (information retrieval methods) και οι περισσότερες από αυτές χρησιμοποιούν ως κεντρική μαθηματική δομή την έννοια του διανυσματικού χώρου, η οποία αποτελεί το βασικό αντικείμενο μελέτης της γραμμικής άλγεβρας. Παραπέμπουμε στο άρθρο (Berry, et al

1999) για μια ανασκόπηση των μεθόδων γραμμικής άλγεβρας στην ανάκτηση πληροφορίας.

Για συλλογές που αποτελούνται από σχετικά μικρό αριθμό αρχείων, η πιο συνηθισμένη μέθοδος ανάκτησης πληροφορίας είναι η μέθοδος LSI (Latent Semantic Indexing) (Bonato, 2008). Η μέθοδος όμως αυτή έχει περιορισμένη αποτελεσματικότητα όταν εφαρμοστεί στον τεράστιο όγκο πληροφοριών του παγκόσμιου ιστού (World Wide Web). Το 1998 οι Larry Page και Sergey Brin, θεμελιωτές της μηχανής αναζήτησης Google, στην εργασία (Brin, et al 1998). ανέπτυξαν έναν αλγόριθμο (γνωστός ως PageRank algorithm) σύμφωνα με τον οποίο, σε κάθε ιστοσελίδα ενός δικτυακού τόπου (Web) δίνεται ένας δείκτης βαρύτητας, μέσω μια διαδικασίας "ψηφοφορίας" από άλλες ιστοσελίδες. Αποτέλεσμα αυτού είναι ότι όταν ο χρήστης αναζητήσει μια πληροφορία στον δικτυακό τόπο, οι πιο σχετικές πληροφορίες εμφανίζονται στην οθόνη του σε πρώτη κατάταξη. Για εκτενή περιγραφή της μεθόδου παραπέμπουμε στα άρθρα (Bryan , et al 2006), (Langville et al, 2005) και στο βιβλίο (Bonato, 2008). Άλλοι αλγόριθμοι αναζήτησης είναι οι HITS (Hypertext Induced Topic Search) (Kleinberg, 1999) και SALSA (Stochastic Approach for Link Structure Analysis) (Lempel, et al 2000).

Στην παρούσα εργασία θα παρουσιάσουμε τις μαθηματικές αρχές του προσδιορισμού του δείκτη βαρύτητας των ιστοσελίδων ενός δικτυακού τόπου, καθώς και διάφορες μαθηματικές δυσκολίες που προκύπτουν από τη διαδικασία αυτή. Για μια βαθύτερη κατανόηση συστημάτων κατάταξης από πολλές απόψεις παραπέμπουμε στο βιβλίο (Langville, et al 2010).

Είναι ενδιαφέρον και σε ένα βαθμό απροσδόκητο, το γεγονός ότι ένα πρακτικό πρόβλημα του διαδικτύου οδηγεί όχι μόνο σε ενδιαφέροντα μαθηματικά προβλήματα, αλλά και σε αποδείξεις συγκεκριμένων θεωρημάτων, κάτι το οποίο αναδεικνύει την διεπιστημονική αξία των μαθηματικών.

## 2. Χαρακτηριστικά των μηχανών αναζήτησης.

Ένα από τα πιο βασικά χαρακτηριστικά μιας μηχανής αναζήτησης στο διαδίκτυο (π.χ. Google, Yahoo, Alta Vista) είναι η δυνατότητά της να ταξινομεί όλες τις διαθέσιμες ιστοσελίδες με τέτοιον τρόπο, ώστε να παρουσιάζονται στο χρήστη εκείνες οι σελίδες που του είναι πιο χρήσιμες. Αυτή τη στιγμή υπάρχουν στο διαδίκτυο περίπου 25 δισεκατομμύρια ιστοσελίδες, συνεπώς είναι καίριας σημασίας το ερώτημα το πώς μια μηχανή αναζήτησης θα παρουσιάσει τις ιστοσελίδες σε φθίνουσα σειρά ενδιαφέροντος. Για παράδειγμα, ας φανταστούμε τη δυσκολία που θα είχαμε αν αναζητούσαμε μια πληροφορία σε μια βιβλιοθήκη, η οποία να περιέχει 25 δισεκατομμύρια πηγές, αλλά να μην έχει ούτε κεντρική οργάνωση ούτε βιβλιοθηκάριους.

Μια μηχανή αναζήτησης έχει συνήθως τρεις αποστολές:

α) Πλοήγηση στο διαδίκτυο για αναζήτηση όλων των ιστοσελίδων που έχουν ελεύθερη πρόσβαση.

β) Ταξινόμηση των δεδομένων του α) με τρόπο ώστε να μπορούν να αναζητηθούν χρησιμοποιώντας λέξεις ή φράσεις-κλειδιά.

γ) Δημιουργία ενός "δείκτη βαρύτητας" για κάθε ιστοσελίδα του δικτυακού τόπου, ώστε όταν εντοπιστούν διάφορες ιστοσελίδες σχετικές με κάποια αναζήτηση, οι σελίδες να εμφανίζονται στην οθόνη σε φθίνουσα σειρά βαρύτητας.

Στην παρουσίαση αυτή θα επικεντρωθούμε στο (γ). Συγκεκριμένα, θα παρουσιάσουμε έναν αλγόριθμο προσδιορισμού ενός δείκτη βαρύτητας ιστοσελίδων (PageRank algorithm), ο οποίος χρησιμοποιήθηκε για πρώτη φορά από τη μηχανή

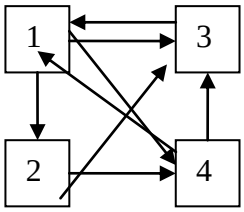
αναζήτησης Google. Αποτέλεσμα του αλγορίθμου αυτού είναι η συγκεκριμένη μηχανή αναζήτησης, να παρουσιάζει τις ιστοσελίδες που αφορούν το χρήστη σε σειρά φθίνουσας σημασίας. Ενδιαφέρον παρουσιάζει η μαθηματική πλευρά του προσδιορισμού του δείκτη βαρύτητας ιστοσελίδων και αυτό είναι το σημείο που θα επικεντρωθούμε. Αν και τα απαιτούμενα μαθηματικά είναι απλή γραμμική άλγεβρα, έχει αξία να παρατηρήσουμε πώς συγκεκριμένα προβλήματα που αφορούν την τεχνολογία μεταφέρονται σε σαφή μαθηματικά ερωτήματα, των οποίων οι απαντήσεις οδηγούν στη διατύπωση και απόδειξη προτάσεων.

## 3. Προσδιορισμός του δείκτη βαρύτητας ιστοσελίδων.

Θα παρουσιάσουμε έναν τρόπο "βαθμολόγησης" κάθε ιστοσελίδας ενός δικτυακού τόπου (Web) αντιστοιχίζοντας σε κάθε ιστοσελίδα έναν μη αρνητικό αριθμό, ο οποίος θα ονομάζεται *δείκτης βαρύτητας* (δ.β.) της ιστοσελίδας. Η βασική ιδέα του προσδιορισμού του δείκτη αυτού, είναι ότι αυτός θα προκύπτει από τον αριθμό των αναφορών που δέχεται η συγκεκριμένη ιστοσελίδα από άλλες ιστοσελίδες, αλλά και από το πόσο σημαντικές είναι οι ιστοσελίδες αυτές. Με άλλα λόγια, ο δ.β. θα προσδιοριστεί από μια διαδικασία "ψηφοφορίας" μεταξύ των ιστοσελίδων.

Ας υποθέσουμε ότι ένας δικτυακός τόπος αποτελείται από $n$ ιστοσελίδες, όπου η κάθε μια προσδιορίζεται από έναν ακέραιο $k, 1 \le k \le n$. Στο παρακάτω Παράδειγμα 1 παρουσιάζεται ένας δικτυακός τόπος που περιέχει $n = 4$ ιστοσελίδες. Κάθε βέλος από την ιστοσελίδα Α στην ιστοσελίδα Β δηλώνει ότι η Α αναφέρει την Β. Συμβολίζουμε με $x_k$ τον δ.β. της ιστοσελίδας $k$. Συνεπώς, όταν η ιστοσελίδα $j$ είναι πιο σημαντική από την ιστοσελίδα $k$, θα πρέπει να ισχύει $x_k \ge 0$ και $x_j > x_k$. Η πιο απλοϊκή προσέγγιση είναι να ορίσουμε ως $x_k$ να είναι ο συνολικός αριθμός των αναφορών που δέχεται η ιστοσελίδα $k$.

**Παράδειγμα 1**. Θεωρούμε τον δικτυακό τόπο του παρακάτω σχήματος. Τα βέλη δείχνουν τις αναφορές των ιστοσελίδων μεταξύ τους, π.χ. η ιστοσελίδα 1 αναφέρεται από τις ιστοσελίδες 3 και 4. Άρα οι δείκτες βαρύτητας είναι $x_1 = 2, x_2 = 1, x_3 = 3$ και $x_4 = 2$, το οποίο σημαίνει ότι η πιο σημαντική είναι η ιστοσελίδα 3 και ακολουθούν οι ιστοσελίδες 1 και 4 (σε ισοβαθμία), με την ιστοσελίδα 2 να είναι η λιγότερο σημαντική.




Η παραπάνω όμως προσέγγιση δεν λαμβάνει υπόψη μια σημαντική παράμετρο. Συγκεκριμένα, θα πρέπει η αναφορά που δέχεται η ιστοσελίδα $k$ από μια σημαντική ιστοσελίδα να συνεισφέρει στον δείκτη βαρύτητας της $k$ περισσότερο από το αν

δεχόταν αναφορά από μια ασήμαντη ιστοσελίδα. Στο Παράδειγμα 1 οι σελίδες 1 και 4 έχουν δύο αναφορές, αλλά η δεύτερη αναφορά της σελίδας 1 προέρχεται από τη σελίδα 3, η οποία φαίνεται να είναι η πιο σημαντική, ενώ η δεύτερη αναφορά της σελίδας 4 προέρχεται από τη σελίδα 2, ο οποία δεν είναι ιδιαίτερα σημαντική.

Έτσι λοιπόν, ας ορίσουμε προσωρινά τον δείκτη βαρύτητας της ιστοσελίδας $j$ ως το άθροισμα των δεικτών βαρύτητας όλων των ιστοσελίδων οι οποίες αναφέρουν την $j$. Για το παράδειγμά μας, θα είναι $x_1 = x_3 + x_4, x_2 = x_1, x_3 = x_1 + x_2 + x_4$ και $x_4 = x_1 + x_2$. Αν και φαίνεται ότι η διαδικασία αυτή έχει αυτοαναφορικό χαρακτήρα, είναι η κατάλληλη για τον τελικό προσδιορισμό του δ.β.

Θα χρειαστούμε μόνο μια τελική τροποποίηση. Επιθυμούμε να έχουμε έναν αλγόριθμο κατά τον οποίο ο δ.β μιας ιστοσελίδας να μην κερδίζει επιπλέον επιρροή όταν συνδέεται με πολλές άλλες ιστοσελίδες. Έτσι λοιπόν, αν η ιστοσελίδα $j$ έχει $n_j$ συνδέσεις σε άλλες ιστοσελίδες, μία από τις οποίες είναι προς τη σελίδα $k$, τότε η συνεισφορά στον δ.β. της σελίδας $k$ θα είναι $\frac{x_j}{n_j}$ αντί για $x_j$. Με τον τρόπο αυτό κάθε ιστοσελίδα ουσιαστικά λαμβάνει μία μόνο "ψήφο".

Γενικά, έστω $n$ το πλήθος των ιστοσελίδων σε έναν δικτυακό τόπο και $L_k \subset \{1,\ldots,n\}$ το σύνολο των ιστοσελίδων που αναφέρονται στην ιστοσελίδα $k$. Τότε για κάθε $k = 1,\ldots,n$ απαιτούμε να ισχύει η σχέση

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j}, \qquad (1)$$

όπου $n_j$ είναι ο αριθμός των συνδέσεων που "φεύγουν" από την ιστοσελίδα $j$. Σημειώνουμε ότι $n_j > 0$ δεδομένου ότι αν $j \in L_k$, τότε η σελίδα $j$ συνδέεται τουλάχιστον με τη σελίδα $k$.

Εφαρμόζοντας τη διαδικασία αυτή στο Παράδειγμα 1, λαμβάνουμε τελικά ότι οι δείκτες βαρύτητας των ιστοσελίδων 1,2,3 και 4 πρέπει να ικανοποιούν τις εξισώσεις

$$x_1 = \frac{x_3}{1} + \frac{x_4}{2}, x_2 = \frac{x_1}{3}, x_3 = \frac{x_1}{3} + \frac{x_2}{2} + \frac{x_4}{2}, x_4 = \frac{x_1}{3} + \frac{x_2}{2}.$$

Οι παραπάνω εξισώσεις μπορούν να εκφραστούν ισοδύναμα ως $A\vec{x} = \vec{x}$, όπου $\vec{x} = (x_1\ x_2\ x_3\ x_4)^T$ και

$$A = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}. \qquad (2)$$

Ο πίνακας $A$ ονομάζεται *πίνακας σύνδεσης των ιστοσελίδων*. Έτσι λοιπόν το πρόβλημα του προσδιορισμού των δεικτών βαρύτητας των ιστοσελίδων ενός δικτυακού τόπου ανάγεται σε πρόβλημα προσδιορισμού ενός ιδιοδιανύσματος $x$ ενός τετραγωνικού πίνακα με αντίστοιχη ιδιοτιμή 1. Τα ιδιοδιανύσματα του πίνακα (2) με αντίστοιχη ιδιοτιμή 1 είναι όλα τα πολλαπλάσια του διανύσματος $(2\ 4\ 9\ 6)^T$. Αν κανονικοποιήσουμε τις συντεταγμένες ώστε να έχουν άθροισμα μονάδα, παίρνουμε τελικά ότι $x_1 = \frac{12}{31} \cong 0{,}387, x_2 = \frac{4}{31} \cong 0{,}129, x_3 = \frac{9}{31} \cong 0{,}290$ και $x_4 = \frac{6}{31} \cong 0{,}194$. Η

κατάταξη αυτή διαφέρει ελαφρώς από την προηγούμενη. Η ιστοσελίδα 3, αν και αναφέρεται από όλες τις άλλες ιστοσελίδες, δεν είναι η πιο σημαντική στην κατάταξη αυτή. Για να γίνει αυτό κατανοητό, παρατηρούμε ότι η σελίδα 3 αναφέρει μόνο τη σελίδα 1, οπότε υπό μία έννοια η ψήφος της "καταναλώνεται" προς τη σελίδα αυτή. Σε συνδυασμό με την ψήφο της σελίδας 2, τελικά η σελίδα 1 λαμβάνει την υψηλότερη κατάταξη.

## 4. Μαθηματική θεμελίωση του προβλήματος

Θα αποδείξουμε ότι αν ο πίνακας $A$ ενός δικτυακού τόπου δεν περιέχει "αιωρούμενες" ιστοσελίδες (dangling nodes) (δηλ. ιστοσελίδες από τις οποίες να μην "φεύγει" κανένα βέλος), τότε αυτός έχει πάντα ως ιδιοτιμή το 1.

**Ορισμός.** *Ένας τετραγωνικός πίνακας ονομάζεται στοχαστικός εάν όλα τα στοιχεία του είναι μη αρνητικά και το άθροισμα των στοιχείων κάθε στήλης του ισούται με 1.*

**Πρόταση 1.** *Εάν ο πίνακας $A$ ενός δικτυακού τόπου δεν περιέχει αιωρούμενες ιστοσελίδες, τότε είναι στοχαστικός.*

**Απόδειξη.** Από τον τρόπο κατασκευής του πίνακα $A$, εάν η ιστοσελίδα $j$ συνδέεται με την ιστοσελίδα $i$ τότε το $(i,j)$- στοιχείο του είναι το $a_{ij} = 1/n_j$, διαφορετικά $a_{ij} = 0$. Συνεπώς, η $j$-στήλη του $A$ περιέχει $n_j$ το πλήθος μη μηδενικά στοιχεία, το κάθε ένα από τα οποία ισούται με $1/n_j$, άρα το άθροισμα των στοιχείων κάθε στήλης είναι 1.

**Πρόταση 2.** *Κάθε στοχαστικός πίνακας έχει το 1 ως ιδιοτιμή.*

**Απόδειξη.** Έστω $A$ ένας $n \times n$ στοχαστικός πίνακας και $e = (1 \cdots 1)^T \in \mathrm{R}^n$. Επειδή ο $A$ είναι στοχαστικός έχουμε ότι $A^t e = e$, δηλ. το 1 είναι ιδιοτιμή του $A^t$, άρα ως γνωστόν είναι ιδιοτιμή και του $A$.

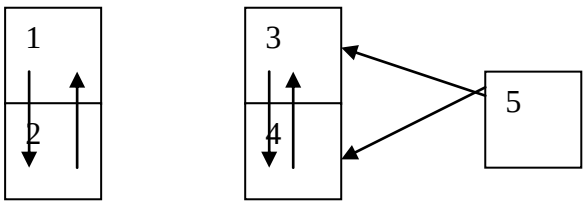
Συμβολίζουμε με $E_1(A)$ τον ιδιόχωρο του πίνακα $A$ που αντιστοιχεί στην ιδιοτιμή 1. Η χρήση του τύπου (1) παρουσιάζει διάφορες δυσκολίες, δύο από τις οποίες είναι οι εξής:
1) Οι δείκτες βαρύτητας των ιστοσελίδων ενός δικτυακού τόπου δεν προσδιορίζονται κατά μοναδικό τρόπο.
2) Η ύπαρξη "αιωρούμενων" ιστοσελίδων σε έναν δικτυακό τόπο.

Αποτέλεσμα της δυσκολίας 2) είναι να υπάρχουν στήλες του πίνακα $A$ με όλα τα στοιχεία μηδέν. Αυτό μπορεί να ξεπεραστεί με διαφόρους τρόπους (βλ. για παράδειγμα (Bryan, et al 2006: 5). Ο πιο απλοϊκός είναι να θέσουμε όλα τα στοιχεία της στήλης αυτής ίσα με $1/n$, οπότε ο πίνακας γίνεται στοχαστικός.

Σχετικά με το 1), προκειμένου να υπάρχει μοναδικό ιδιοδιάνυσμα $x = (x_1 \cdots x_n)^T$ ώστε $\sum x_i = 1$ (το οποίο θα προσδιορίζει τους δείκτες βαρύτητας των ιστοσελίδων κατά μοναδικό τρόπο), θα πρέπει να ισχύει $\dim E_1(A) = 1$. Ο πίνακας (2) ικανοποιεί

αυτή τη συνθήκη αλλά όπως φαίνεται στο παρακάτω παράδειγμα, αυτό δεν ισχύει γενικά.

**Παράδειγμα 2.** Θεωρούμε το διακτυακό τόπο του παρακάτω σχήματος.




Ο αντίστοιχος πίνακας σύνδεσης των ιστοσελίδων είναι ο $A=\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$, όπου $A_1=\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ και $A_2=\begin{pmatrix} 0 & 1 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 0 & 0 \end{pmatrix}$. Ο ιδιόχωρος $E_1(A)$ έχει διάσταση 2 (π.χ. $E_1(A)=\operatorname{span}\{x=(1/2,1/2,0,0,0)^t, y=(0,0,1/2,1/2,0)^t\}$.
Επειδή οποιοσδήποτε γραμμικός συνδυασμός των διανυσμάτων $x, y$ είναι ένα στοιχείο του $E_1(A)$, δεν υπάρχει μοναδικός προσδιορισμός του δείκτη βαρύτητας από ένα και μόνο ιδιοδιάνυσμα του $E_1(A)$. Ισχύει λοιπόν το εξής:

**Πρόταση 3.** *Εάν ένας δικτυακός τόπος W με πίνακα σύνδεσης $A$ αποτελείται από $r$ συνεκτικές συνιστώσεις (δηλ. μικρότερους δικτυακούς τόπους), οι οποίες δεν συνδέονται μεταξύ τους, τότε* $\dim E_1(A) \geq r$.

**Απόδειξη.** Έστω ότι ο ιστότοπος W αποτελείται από $n$ ιστοσελίδες οι οποίες δεν συνδέονται μεταξύ τους και έστω $W_1,\ldots,W_r$ οι $r$ συνεκτικές συνιστώσες. Αν $n_i$ είναι ο αριθμός των ιστοσελίδων του δικτυακού τόπου $W_i$, αριθμούμε τις ιστοσελίδες του W ώστε από 1 έως $n_1$ να είναι οι ιστοσελίδες του $W_1$, από $n_1+1$ έως $n_1+n_2$ να είναι οι ιστοσελίδες του $W_2$, από $n_1+n_2+1$ έως $n_1+n_2+n_3$ του $W_3$, κ.ο.κ.. Τότε, ο πίνακας σύνδεσης $A$ του δικτυακού τόπου W θα έχει τη μορφή

$$A=\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & 0 & 0 \\ 0 & \vdots & \ddots & 0 \\ 0 & 0 & 0 & A_r \end{pmatrix},$$

όπου $A_i$ ο πίνακας σύνδεσης του δικτυακού τόπου $W_i$. Κάθε $n_i \times n_i$ πίνακας $A_i$ είναι στοχαστικός, άρα έχει ένα ιδιοδιάνυσμα $v^i \in \mathrm{R}^{n_i}$ με αντίστοιχη ιδιοτιμή 1. Για κάθε

$1 \le i \le r$ θεωρούμε το διάνυσμα $w^i \in \mathrm{R}^n$ με $w^i = (0 \cdots 0\, v^i\, 0 \cdots 0)^t$. Τότε, επειδή $Aw^i = A(0 \cdots 0\, v^i\, 0 \cdots 0)^t = w^i$, τα $w^i$ $1 \le i \le r$ είναι γραμμικώς ανεξάρτητα ιδιοδιανύσματα του πίνακα $A$ με αντίστοιχες ιδιοτιμές 1, συνεπώς $\dim E_1(A) \ge r$.

Δεδομένου ότι το διαδίκτυο αποτελείται από δισεκατομμύρια ιστοσελίδες με πάρα πολλές μη συνεκτικές συνιστώσες από μικρότερους δικτυακούς τόπους, είναι σημαντικό να ξεπεραστεί η δυσκολία ότι $\dim E_1(A) \ge 1$. Εργαζόμαστε ως εξής:

Υποθέτουμε ότι ένας δικτυακός τόπος αποτελείται από $n$ μη αιωρούμενες ιστοσελίδες. Έστω $S$ ο $n \times n$ πίνακας με όλα τα στοιχεία ίσα με $\frac{1}{n}$. Τότε ο $S$ είναι στοχαστικός και $\dim E_1\ S = 1$. Θεωρούμε τον πίνακα

$$M = 1-m\ A + mS,\ 0 \le m \le 1 \quad (3).$$

Θα αποδείξουμε ότι αν $m \in (0,1]$ τότε $\dim V_1\ M = 1$, άρα ο πίνακας $M$ μπορεί να χρησιμοποιηθεί για τη μοναδικότητα του προσδιορισμού του δείκτη βαρύτητας. Εάν $m = 1$ τότε $M = S$ και το μόνο ιδιοδιάνυσμα $x$ με ιδιοτιμή 1 έχει όλες τις συντεταγμένες ίσες με $\frac{1}{n}$ οπότε όλες οι ιστοσελίδες έχουν την ίδια βαρύτητα.

**Ορισμός.** *Ένας πίνακας $M$ ονομάζεται θετικός εάν $m_{ij} > 0$ για κάθε $i, j$.*

*Επειδή ο $M$ είναι στοχαστικός* ισχύει $E_1\ M \ne \varnothing$. Θα χρειαστούμε δύο λήμματα, η απόδειξη των οποίων παρατίθεται στο τέλος της εργασίας.

**Λήμμα 1.** *Εάν ο πίνακας $M$ είναι θετικός και στοχαστικός, τότε οι συντεταγμένες κάθε ιδιοδιανύσματος του $E_1\ M$ είναι είτε όλες θετικές είτε όλες αρνητικές.*

**Λήμμα 2.** *Έστω $v, w$ δύο γραμμικώς ανεξάρτητα διανύσματα του $\mathbf{R}^m$, $m \ge 2$. Τότε υπάρχουν $s, t \in \mathbf{R}$ όχι και οι δύο μηδέν, ώστε οι συντεταγμένες του διανύσματος $x = sv + tw$ να είναι ταυτόχρονα θετικές και αρνητικές.*

Τώρα μπορούμε να αποδείξουμε το κεντρικό αποτέλεσμα.

**Θεώρημα 1.** *Εάν ο πίνακας $M$ είναι θετικός και στοχαστικός τότε* $\dim E_1\ M = 1$.

**Απόδειξη.** Υποθέτουμε ότι υπάρχουν δύο γραμμικώς ανεξάρτητα διανύσματα $v, w \in E_1\ M$ και θα οδηγηθούμε σε άτοπο. Για κάθε $s, t \in \mathbf{R}$ όχι και τα δύο μηδέν, το διάνυσμα $x = sv + tw$ ανήκει στον $E_1\ M$ και από το Λήμμα 1 έχει όλες τις συντεταγμένες του είτε θετικές είτε αρνητικές. Αλλά αυτό αντιβαίνει στο Λήμμα 2, οπότε ο ιδιόχωρος $E_1\ M$ δεν μπορεί να περιέχει δύο γραμμικώς ανεξάρτητα διανύσματα, άρα $\dim E_1\ M = 1$.

Δεδομένου ότι το διαδίκτυο αποτελείται από τουλάχιστον οκτώ δισεκατομμύρια ιστοσελίδες, τίθεται το ερώτημα πώς υπολογίζουμε το ιδιοδιάνυσμα ενός $n \times n$ πίνακα, όπου $n = 8.000.000.000$. Αυτό γίνεται με τη χρήση της μεθόδου της δύναμης (βλ. για παράδειγμα ένα από τα (Meyer, et al 2000), (Strang, 2006) και αποδεικνύεται

ότι με τις υποθέσεις που έχουμε η μέθοδος συγκλίνει. Έχουμε λοιπόν καταλήξει στο ακόλουθο θεώρημα.

**Θεώρημα 2.** *Έστω $M$ ο πίνακας που αντιστοιχεί σε έναν δικτυακό τόπο χωρίς αιωρούμενες ιστοσελίδες, όπως δίνεται στη σχέση (3). Τότε ο $M$ είναι ένας στοχαστικός πίνακας άρα υπάρχει μοναδικό διάνυσμα $\vec{q}$, τέτοιο ώστε $M\vec{q} = \vec{q}$ και $\sum_i q_i = 1$. Το διάνυσμα $\vec{q}$ είναι δυνατόν να υπολογιστεί ως το όριο της επαναληπτικής διαδικασίας $\vec{x}_k = (1-m)Ax_{k-1} + mS$, για οποιοδήποτε αρχικό διάνυσμα $x_0$ με θετικές συντεταγμένες και μέτρου 1 (ως προς το μέτρο $|\vec{x}|_1 = \sum_i x_i$ ).*

Συμπερασματικά, η ανάκτηση πληροφορίας από το διαδίκτυο απαιτεί νέες μαθηματικές τεχνικές, λόγω του όγκου των πληροφοριών που είναι διάσπαρτες σε αυτό. Παραδοσιακές μέθοδοι όπως η LSI εμφανίζουν αδυναμίες. Οι απαραίτητοι υπολογισμοί ανοίγουν νέους δρόμους για εφαρμογή γνωστών μαθηματικών θεωριών, αλλά και διατυπώνουν προτάσεις προς απόδειξη.

## 5. Παράρτημα.

Δίνουμε εδώ τις αποδείξεις των λημμάτων 1 και 2.

**Απόδειξη Λήμματος 1.** Υποθέτουμε ότι υπάρχει ένα $x \in E_1$ $M$ με συντεταγμένες θετικές και αρνητικές. Τότε από τη σχέση $x = Mx$ προκύπτει ότι $x_i = \sum_{j=1}^{n} m_{ij} x_j$ $i = 1,..,n$. Επειδή $m_{ij} > 0$ οι όροι $m_{ij}x_j$ έχουν ανάμεικτα πρόσημα. Χρησιμοποιώντας την τριγωνική ανισότητα έχουμε ότι $|x_i| = |\sum_{j=1}^{n} m_{ij} x_j| < \sum_{j=1}^{n} m_{ij} |x_j|$. Αθροίζουμε από $i = 1$ έως $i = n$ και εναλλάσσουμε τους δείκτες $i$ και $j$. Επειδή ο $M$ είναι στοχαστικός, τότε για κάθε $j = 1,\ldots,n$ προκύπτει ότι $\sum_i m_{ij} = 1$, άρα

$$\sum_{i=1}^{n} |x_i| = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}|x_j| = \sum_{j=1}^{n}\left(\sum_{i=1}^{n} m_{ij}\right)|x_j| = \sum_{j=1}^{n}|x_j|, \text{ άτοπο.}$$

Άρα το διάνυσμα $x$ δεν μπορεί να έχει ταυτόχρονα θετικές και αρνητικές συντεταγμένες. Εξετάζουμε τώρα την περίπτωση όπου $x_i \geq 0$ για κάθε $i$ (αλλά όχι όλα τα $x_i$ μηδέν). Τότε επειδή $x_i = \sum_{j=1}^{n} m_{ij} x_j$ και $m_{ij} > 0$ προκύπτει ότι $x_i > 0$. Ανάλογα δείχνουμε ότι αν $x_i \leq 0$, τότε $x_i < 0$ για κάθε $i$.

**Απόδειξη Λήμματος 2.** Λόγω της γραμμικής ανεξαρτησίας τα $v$ και $w$ είναι μη μηδενικά διανύσματα. Έστω $v = v_1,..,v_m$, $w = w_1,..,w_m$ και $d = \sum_{i=1}^{m} v_i$. Εάν $d = 0$, τότε οι συντεταγμένες του $v$ έχουν θετικό και αρνητικό πρόσημο, οπότε για $s = 1$ και $t = 0$ λαμβάνουμε το αποτέλεσμα. Εάν $d \neq 0$, τότε θέτουμε $s = -\frac{\sum_{i=1}^{m} w_i}{d}$,

$t=1$. Επειδή τα $v, w$ είναι γραμμικώς ανεξάρτητα, το διάνυσμα $x = sv + tw$ είναι μη μηδενικό. Επίσης $\sum x_i = 0$, συνεπώς το διάνυσμα $x$ περιέχει θετικές και αρνητικές συντεταγμένες.

Δίνουμε ένα παράδειγμα εφαρμογής του Λήμματος 2.

**Παράδειγμα 3.** Έστω $v = -1,3,2$, $w = 7,-4,-5 \in \mathbf{R}^3$. Τότε $d = 4 \neq 0$, $s = -\frac{-2}{4} = \frac{1}{2}$ και το διάνυσμα $x = \frac{1}{2}\ -1,3,2\ +1\ 7,-4,-5\ = \left(\frac{13}{2}, \frac{-5}{2}, -4\right)$ έχει την επιθυμητή ιδιότητα..

## Βιβλιογραφία

Berry, M.W. and Browne, M. (2005). *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, 2nd ed., SIAM, Philadelphia.

Berry, M.W., Drmac, Z. and Jessup, E.R. (1999). Matrices, vector spaces, and information retrieval, *SIAM Review, 41*, pp. 335-362.

Brin, S. and Page, L. (1998). Anatomy of a large-scale hypertextual web search engine, in *Proceedings of the 7th Intern. World Wide Web Conference*.

Bryan, K. and Leise, T. (2006). The $25,000,000 eigenvector: The linear algebra behind google, *SIAM Review 48(3)* pp. 569-581.

Bonato, A. (2008). *A Course on the Web Graph*, American Math. Society, Graduate Studies in Mathematics 89, Rhode Island.

Dumais, S.T. (1991). Improving the retrieval of information from external sources, *Behavior Research Methods, Instruments and Computers, 23*, pp. 229-236.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment, *J. of the ACM, 46*, pp. 604-632

Langville, A.N. and Meyer, C.D. (2005). A survey of eigenvector methods of web information retrieval, *SIAM Review, 47*, pp. 135-161.

Langville, A.N. and Meyer, C.D. (2005). Deeper inside PageRank, *Internet Math., 1*, pp. 335-380.

Langville, A.N. and Meyer, C.D. (2010). *Η Μέθοδος PageRank της Google και άλλα Συστήματα Κατάταξης*, Πανεπιστημιακές Εκδόσεις Κρήτης.

R. Lempel, P. and Moran, S. (2000). The stochastic approach for link-structure analysis (SALSA) and the TKC effect, in *The Ninth International World Wibe Web Conference*, New York, ACM Press.

Meyer, C.D. (2000). *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia.

Rogers, I. The Google Pagerank algorithm and how it works, http://www.iprcom.com/papers/pagerank

Strang, G. (2006): *Εισαγωγή στη Γραμμική Άλγεβρα*, Εκδόσεις Πανεπιστημίου Πατρών.