ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

Μοντέλα Αναπαράστασης Κειμένων με χρήση Γράφων και Εφαρμογές: Επισκόπηση

Πτυχιακή εργασία

Δημακαράκος Θεόδωρος

Αθήνα, 2018

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ

Τριμελής Εξεταστική Επιτροπή

Ηρακλής Βαρλάμης (Επιβλέπων)
Επίκουρος Καθηγητής, Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

~

Δημήτριος Μιχαήλ
Επίκουρος Καθηγητής, Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

~

Δημοσθένης Αναγνωστόπουλος
Καθηγητής, Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Ο Δημακαράκος Θεόδωρος

δηλώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

Στην Τζοβάννα Νατέλι

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Περίληψη

Το κύριο μέλημα της παρούσας εργασίας είναι να περιγράψει τους τρόπους που δύναται να αναπαρασταθεί το κείμενο σε μορφή δικτύου, σύμφωνα με τη θεωρία γράφων και σε ποιες πρακτικές εφαρμογές μπορεί αυτή να συνεισφέρει και με ποιόν τρόπο. Αρχικά δίνονται ορισμοί στα πεδία της θεωρίας συνόλων, της θεωρίας πινάκων και της θεωρίας γράφων. Αφού εξηγηθούν τα βασικά στοιχεία που θα βοηθήσουν στην κατανόηση όσων ακολουθούνε, οικοδομούνται οι βάσεις και διανοίγονται οι δυνατότητες για την μετατροπή της γραπτής γλώσσας σε μορφή γράφου πάνω στις οποίες θα στηριχθεί η εργασία. Περιγράφονται εργαλεία από την ίδια τη θεωρία γράφων, και δάνεια από άλλα πεδία, που ενδείκνυνται για την εξαγωγή σημαντικών πληροφοριών σχετικά με το κείμενο. Πριν παρουσιαστούν εφαρμογές από τη βιβλιογραφία γίνεται μια ενδεικτική περιγραφή των αποτελεσμάτων που εξήγαγε η γλωσσολογία από τη μελέτη της γλώσσας, και κυρίως της γραπτής, όταν την αντιμετώπισε με τη μορφή δικτύου. Με βάση όλα τα παραπάνω, ανοίγει το δεύτερο μέρος της εργασίας που έχει μορφή βιβλιογραφικής επισκόπησης σύγχρονων ερευνών πάνω στην πρακτική αξία των μοντέλων γράφων αναπαράστασης κειμένου. Η καινοτομία της χρήσης γράφων σε διάφορες εφαρμογές επεξεργασίας φυσικής γλώσσας και ανάκτησης πληροφοριών βρίσκεται στο γεγονός πως παραβιάστηκε η έως τότε υπόθεση της αναμεταξύ των όρων ανεξαρτησίας. Στο κυρίως σώμα της εργασίας, επισκοπούνται τρεις διαφορετικές, όμως στενά συνδεδεμένες, διεργασίες και οι εφαρμογές τους. Η πρώτη που περιγράφεται είναι η εξαγωγή φράσεων-κλειδιών από κείμενα και αποτελεί την πιο, βιβλιογραφικά, πλούσια όσον αφορά την έρευνα πάνω στο θέμα. Στη συνέχεια, η εργασία προχωράει στην αυτόματη περίληψη και εξετάζονται εξορυκτικές περιλήψεις ενός ή πολλών εγγράφων. Τέλος, αναπτύσσεται η ανάκτηση πληροφοριών, όπου περιγράφεται λεπτομερειακά η διαδικασία της. Σε όλες τις παραπάνω διαδικασίες γίνεται παράθεση των αποτελεσμάτων των ερευνών και μια μικρή ερμηνευτική σύγκριση.

Λέξεις κλειδιά: **Γράφοι Κειμένων, PageRank, Ανάκτηση Πληροφοριών, Εξαγωγή φράσεων-κλειδιών, Αυτόματη Μετάφραση, Βιβλιογραφική Επισκόπηση**

Abstract

The main concern of this thesis is to describe the ways in which a text can be represented in network form, in accordance with graph theory and, in addition, in which way and on what practical applications could it contribute. Primarily, several definitions are given in the fields of set theory, matrix theory and graph theory. Once the basic elements that would assist for the clearer comprehension of the subsequent are laid out, a foundations are constructed and possibilities open up for the transformation of written language in graph form, upon which this thesis will be built. Tools from graph theory as well as from other relevant fields are being described, appropriate of extracting relevant information about the text. Just before any research on applications is presented, there occurs a description of the results obtained from computational linguistics when it studied language from the perspective of the network. Based on all of the above, the second section of the thesis is being started, with the form of a survey on modern research upon the practical value of graph-based representations of text. The novelty of applying graphs in several natural language processing and information retrieval application lies in the fact that it violates the assumption of words independence. In the main body of the essay there are being surveyed three different, but closely linked, processes and their applications. The first described is keyphrase extraction from texts and it also is the richest in research material. Subsequently, automatic summarization and more specifically its extractive type for single and multiple documents. Finally, information retrieval is described in details along with its applications. The research results are quoted on all of the above processes, and there is an attempt for interpretive comparisons.

Keywords: Text Garphs, PageRank, Information Retrieval, Keyphrase Extraction, Automatic Summarization, Survey

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Εισαγωγικά. Θεωρία Γράφων – Ορισμοί

Η παρακάτω διατύπωση της θεωρίας γράφων, όπως και η παράθεση εννοιών που αυτή προϋποθέτει είναι σκόπιμη και συμμορφώνεται στην ύλη της εργασίας. Κατά συνέπεια, η σημειογραφία, βασισμένη κυρίως στο (Bollobas 1998), η έκταση και η ανάλυση του περιεχομένου της δεν θεωρούνται ολοκληρωμένες.

Ορισμοί

Ε.1 Σχετικοί ορισμοί θεωρίας συνόλων, επιγραμματικά.

- Στη θεωρία συνόλων ένα *σύνολο* είναι μια καλά προσδιορισμένη συλλογή η οποία χαρακτηρίζεται πλήρως από τα στοιχεία της.
- Ένα σύνολο Α ονομάζεται *υποσύνολο* του Β, όταν όλα τα στοιχεία του Α περιλαμβάνονται στο Β άλλα το αντίθετο δεν ισχύει.
- *Ξένα* ονομάζονται δύο σύνολα που δεν έχουν κανένα κοινό στοιχείο.
- *Πληθικότητα* ενός συνόλου είναι ο αριθμός των στοιχείων του και συμβολίζεται |Α| για ένα σύνολο Α.
- *Πεπερασμένο* ονομάζεται ένα σύνολο, όταν έχει πεπερασμένη πληθικότητα.
- Ένα σύνολο χαρακτηρίζεται *διατεταγμένο* εάν η διάταξη των στοιχείων του θεωρείται σημαντική.
- Η *διατεταγμένη ακολουθία* ή απλά *ακολουθία* ισοδυναμεί με ένα διατεταγμένο σύνολο.
- Ένα *διατεταγμένο ζεύγος* ισοδυναμεί με ένα διατεταγμένο σύνολο που αποτελείται από ένα ζεύγος αντικειμένων.
- Το *διατεταγμένο καρτεσιανό γινόμενο* δύο συνόλων Α, Β συμβολίζεται $A \times B$ και είναι το σύνολο όλων των διατεταγμένων ζευγών (a, b), $a \in A$, $b \in B$.
- Το *μη-διατεταγμένο καρτεσιανό γινόμενο* δύο συνόλων Α, Β συμβολίζεται $A \otimes B$ και είναι το σύνολο όλων των μή-διατεταγμένων ζευγών {a, b}, $a \in A$, $b \in B$.

Ε.2 Ορισμοί θεωρίας γράφων.

Ε.2.1 Κάποιες πρώτες έννοιες της θεωρίας γράφων. Γράφος, Κόμβοι, Ακμές.

Ένας *γράφος G* είναι ένα διατεταγμένο ζεύγος ξένων συνόλων *(V, E)*, έτσι ώστε το *E* να είναι υποσύνολο του συνόλου των μη διατεταγμένων ζευγών των στοιχείων του *V (διατύπωση Bollobas 1998),* δηλαδή $E \subset V \otimes V \subset V \times V$.

Το σύνολο V ονομάζεται *σύνολο των κόμβων* (*vertex set*) και το Ε *σύνολο των ακμών* (*edge set*) και αντίστοιχα, περιλαμβάνουν στοιχεία που ονομάζονται *κόμβοι* και *ακμές*.

Ένας γράφος μπορεί να χαρακτηρισθεί *πεπερασμένος* ή *άπειρος*· πεπερασμένος όταν τα σύνολα V και Ε είναι πεπερασμένα σύνολο και άπειρος όταν δεν είναι πεπερασμένος.
(Σημειώνεται ότι η παρούσα εργασία θα πραγματευθεί μόνο *πεπερασμένους* γράφους.)

Μία *ακμή* {x, y} ενώνει τους κόμβους x και y και συμβολίζεται xy ή yx.

Το σύνολο των κόμβων που ενώνονται μέσω μίας ακμής με έναν κόμβο x ονομάζεται *γειτονιά του κόμβου x (neighborhood of x) Γ(x),* επομένως κάποιος κόμβος που ανήκει στο *Γ(x)* ονομάζεται γείτονας (neighbor) του x και λέγεται πως έχουν σχέση γειτνίασης (adjacency relation), $x \sim y$.

Για την αναφορά στο σύνολο κόμβων ενός συγκεκριμένου γράφου *G* χρησιμοποιείται ο συμβολισμός *V(G)* και όμοια για το σύνολο ακμών ενός γράφου *G* ο *E(G)*.

Όταν δεν υπάρχουν περισσότεροι από έναν γράφο προς αναφορά μπορούν να χρησιμοποιηθούν τα απλοποιημένα σύμβολα V και E αντίστοιχα (όπως παραπάνω).

Ένας γράφος ονομάζεται *υπο-γράφος* ενός άλλου αν το σύνολο κόμβων και το σύνολο ακμών του πρώτου είναι υποσύνολα των αντίστοιχων συνόλων του δευτέρου. Δηλαδή,

$V' \subset V \wedge E' \subset E \rightarrow G' \subset G$, *όπου γράφοι G(V, E), G'(V', E')*

Όταν το σύνολο *E* είναι υποσύνολο του συνόλου <u>των διατεταγμένων ζευγών</u> των στοιχείων του *V* τότε ονομάζεται *κατευθυνόμενος* γράφος ή διγράφος (γλωσσικό δάνειο από την αγγλική digraph που αποτελεί σύνθεση των λέξεων directed και graph) και τα στοιχεία του ονομάζονται *κατευθυνόμενες ακμές.*

Στη περίπτωση αυτή, αφού έχει σημασία η διάταξη των στοιχείων, οι κατευθυνόμενες ακμές (x, y) και (y, x) ονομάζονται *ανεστραμμένες* και δε ταυτίζονται.

Ως προς την ακμή (x, y) το x ονομάζεται *κεφάλι* και το y *ουρά* και συμβολίζεται $\vec{xy}$.

Σταθμισμένος γράφος ή *γράφος με βάρη (weighted graph)* ονομάζεται ο γράφος του οποίου οι ακμές συσχετίζονται με μία αριθμητική τιμή, ("με ένα *βάρος") a:* $E \rightarrow A$. Είναι δυνατό να συσχετισθούν και οι κόμβοι με βάρη, χάριν ορισμού όμως θα τους υποθέσουμε ως ακμές *{v, v}* με βάρη $\alpha: V \rightarrow A \cong \{v, v\} \rightarrow A, \forall v \in V$ Στη περίπτωση που το βάρος ακμής ανάμεσα σε δύο κόμβους ισούται με 1 και το βάρος κάθε κόμβου ισούται με μηδέν τότε ο γράφος ονομάζεται αστάθμητος *(unweighted).*

E.2.2 Τα μέτρα του γράφου.

Τα μέτρα της πρώτης κατηγορίας που πραγματευονται εδώ, δεν αναγνωρίζονται, συνήθως, ως μέτρα με την αυστηρή έννοια της θεωρίας μέτρου αφού αποτελούν πληθικότητες, οι οποίες προϋποτίθενται από τη θεωρίας, μπορούν, όμως, να θεωρηθούν μέτρα ενός γράφου, με την έννοια ότι τον χαρακτηρίζουν ποσοτικά.

E.2.2.1 Η τάξη (order) και το μέγεθος (size)

Τα δύο πρωταρχικά μέτρα είναι η πληθικότητα του συνόλου των κόμβων και του συνόλου των ακμών. Το πρώτο μέτρο συμβολίζεται $|V(G)|$ ή απλοποιημένα $|G|$ και ονομάζεται *τάξη* του γράφου (graph order), το δεύτερο συμβολίζεται *e(G)* και ονομάζεται *μέγεθος* του γράφου (graph size). Η τάξη οριοθετεί τις δυνατές τιμές του μεγέθους στο εύρος [0, $\binom{|G|}{2}$],

όπου $\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1} = \frac{n!}{k!(n-k)!}, k \le n$ και ονομάζεται συνδυασμός n ανά k.

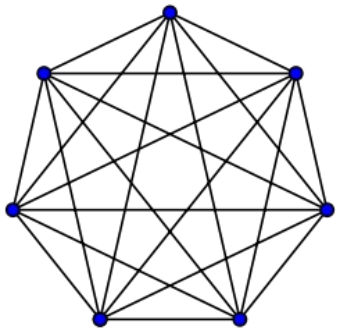
Ένας γράφος με τάξη $|G|$ = n και μέγεθος $e(G) = \binom{n}{2}$ ονομάζεται *πλήρης n-γράφος (complete n-graph) και συμβολίζεται* K^n. (σχήμα Ε.1*)*

Ένας γράφος με τάξη $|G|$ = n και μέγεθος $e(G)$ = 0 ονομάζεται *κενός n-γράφος και συμβολίζεται* E^n.

Ένας γράφος Η ονομάζεται συμπληρωματικός ενός άλλου γράφου *G*, αν

$V(H) = V(G)$ *και* $E(H) = E(K^{|G|})\backslash E(G)$

και συμβολίζεται ως $\overline{G}$. Κατ'επέκταση, $G + \overline{G} = K^{|G|}$ και $\overline{K^n} = E^n$.



Σχήμα Ε.1: Πλήρης 7-Γράφος ή γράφος K^7. (απο τον David Benbennick)

Ε.2.2.2 Ο Βαθμός (Degree)

Τα προηγούμενα μέτρα είχαν σχέση με *ολόκληρο το γράφο* και μπορούν να ονομαστούν *καθολικά (global)*, το επόμενο μέτρο είναι ο *βαθμός (degree)* και σχετίζεται με ένα μόνο κόμβο, ή μάλλον με τον κόμβο και τους γειτονικούς του κόμβους (neighboring nodes ή neighbors)· τέτοιου είδους μέτρα θα ονομάζονται *τοπικά (local)*. Ο βαθμός είναι ο αριθμός των *εφαπτόμενων (incidents),* ως προς τον εξεταζόμενο κόμβο, ακμών και συμβολίζεται *d(x), x ∈ V.* Ο βαθμός έχει, εξίσου, σχέση με την έννοια της πληθικότητας αφού ορίζεται από τη πληθικότητα του συνόλου των εφαπτόμενων ακμών (*Γ(x)* οι εφαπτόμενες ακμές του *x*, *d(x) = |Γ(x)|*). Επιπλέον, ο ελάχιστος βαθμός (minimal degree) που περιλαμβάνεται σε ένα γράφο συμβολίζεται δ(G) και ο μέγιστος βαθμός (maximal degree) Δ(G) και μπορούν να θεωρηθούν καθολικά μέτρα.

Ένας κόμβος χαρακτηρίζεται από το βαθμό του. Όταν έχει βαθμό d(u) = 0 τότε ονομάζεται *απομονωμένος ή μονήρης κόμβος (isolated vertex)*, όταν d(u) = 1 ονομάζεται *εξωτερικός κόμβος ή φύλλο (leaf)* κυρίως όταν μιλάμε για γράφους-δένδρα και, τελικά, όταν $d(u)=|V|-1$ ονομάζεται *κυρίαρχος κόμβος (dominating vertex)*.

Υπάρχουν τρεις τρόποι αναγωγής του βαθμού σε καθολική περιγραφή για το γράφο. Το πρώτο καθολικό μέτρο είναι ο *μέσος βαθμός (average degree)* και υπολογίζεται ως

$$\bar{d}(G)=\frac{1}{|V|}\sum_{u_i\in V} d_G(u_i) \quad \text{ή ισοδύναμα} \quad \bar{d}(G)=2\frac{|E(G)|}{|V(G)|}$$

Το δεύτερο είναι η *κατανομή βαθμού (degree distribution)* στο γράφο

$$P(d(u)=k)=p_k, p_k\in[0,1]$$

η οποία δίνει τη πιθανότητα ενός τυχαία επιλεγμένου κόμβου να έχει βαθμό k. Η κατανομή βαθμού είναι κατανομή πιθανότητας του βαθμού μέσα στο γράφο και η αναμενόμενη τιμή (expected value) της κατανομής προσεγγίζει το μέσο βαθμό.

Το τελευταίο είναι η *ακολουθία βαθμών (degree sequence)* όπου ορίζεται ως μία ακολουθία των βαθμών των κόμβων του γράφου από το μικρότερο στο μεγαλύτερο και συμβολίζεται

$$(d(x_i))_1^n$$

Ένας γράφος του οποίου όλοι οι κόμβοι έχουν τον ίδιο βαθμό k, ονομάζεται k-κανονικός γράφος (k-regular graph) και ακολουθεί μια εκφυλισμένη κατανομή βαθμού (degenerate degree distribution) όπου $P(\bar{d}(u)=k)=1$.

Ε.2.2.3 Συμπληρωματικές έννοιες. Διαδρομή, Κύκλος, Περίπατος, Μονοπάτι, Κύκλωμα, Τρίγωνο. Συστάδες κόμβων

- Μία *διαδρομή (path)* είναι ένας γράφος *P* της μορφής
  $V(P) = \{x_0, x_1, \dots, x_n\}$, $E(P) = \{x_0x_1, x_1x_2, \dots, x_{n-1}x_n\}$
  ή ισοδύναμα μια ακολουθία
  $P = \{e_1, e_2, \dots, e_n\}$, *όπου* $e_i = x_{i-1}x_i$
  συνεπώς, οι ακμές και οι κόμβοι δεν επαναλαμβάνονται.
  Η παραπάνω διαδρομή συμβολίζεται $x_0x_1...x_n$. Οι κόμβοι x_0 και x_n ονομάζονται *τελικοί κόμβοι (endvertices) του P* και το μέγεθος e(P) ονομάζεται *μήκος* του P. Οι διαδρομές ανάμεσα σε δύο κόμβους x_0, x_n συμβολίζονται x_0-x_n, όπου $x_0 \neq x_n$.
  Το *μέγεθος διαδρομής (path length) μεταξύ δύο κόμβων i και j* είναι τα ελάχιστα βήματα που πρέπει να διανυθούν από τον κόμβο i στον j. Το *μέση μέγεθος διαδρομής (average path length) ενός κόμβου* είναι ο μέσος όρος το μέγεθος διαδρομής προς κάθε άλλο κόμβο του γράφου. Το *μέσο μέγεθος διαδρομής του γράφου* είναι ο μέσος όρος της μέσης απόστασης διαδρομής κάθε κόμβου.

- Μία διαδρομή x_0-x_0 , όπου ο πρώτος και ο τελευταίο κόμβος είναι ο ίδιος ονομάζεται *κύκλος (cycle)*.

- Ένας περίπατος (walk) *W* είναι μία εναλλασσόμενη ακολουθία μεταξύ στοιχείων του συνόλου κόμβων και ακμών της μορφής
  $W = \{x_0, e_1, x_1, e_2, \ldots, e_n, x_n\}$, *όπου* $e_i = x_{i-1}x_i$
  όπου κόμβοι και ακμές μπορούν να επαναληφθούν.

- Ένας περίπατος όπου οι ακμές του δεν επαναλαμβάνονται ονομάζεται *μονοπάτι (trail)* και ένα μονοπάτι στο οποίο ο πρώτος και ο τελευταίος κόμβος είναι ο ίδιος, αντιστοίχως με το κύκλο, ονομάζεται *κύκλωμα (circuit)*.

- Ένα *τρίγωνο* ισοδυναμεί με έναν κύκλο *C* με τάξη $|V(C)| = 3$ (ή και με ένα *3-πλήρη γράφο*).

- Η *συστάδα κόμβων (node-cluster ή clique)* είναι ένας υπο-γράφος του γράφου ο οποίος είναι πλήρης. Υπάρχουν δύο μέτρα συσταδοποίησης. Ο *τοπικός συντελεστής συσταδοποίησης (local clustering coefficient)* ενός γράφου είναι ο βαθμός στον οποίο ένας κόμβος του γράφου τείνει να αποτελέσει συστάδα κόμβων με τους γείτονες του. Υπολογίζεται διαιρώντας τον αριθμό τριγώνων του γράφου που περιλαμβάνουν τον κόμβο με τον αριθμό τριάδων (υπο-γράφοι τριών κόμβων και δύο ακμών) του γράφου που περιλαμβάνουν τον κόμβο και ο κόμβος να εφάπτεται με τις δύο ακμές τους. Ο *καθολικός συντελεστής συσταδοποίησης (global clustering coefficient)* είναι ο μέσος όρος του τοπικού συντελεστή όλων των κόμβων.

Ε.3 Πίνακες

Ε.3.1 Βασικά στοιχεία της θεωρίας Πινάκων

Ο *πίνακας* είναι μια μαθηματική αναπαράσταση ορθογώνια διατεταγμένων, σε *σειρές* και *στήλες*, μαθηματικών αντικειμένων, τα οποία ονομάζονται *στοιχεία* του πίνακα. Τα στοιχεία έχουν μια μοναδική *θέση,* στην σειρά *i* και στη στήλη *j* του πίνακα, που συμβολίζεται *(i, j)* ή $\alpha_{i,j}$ για πίνακα *A*. Οι ίδιες οι σειρές και στήλες μπορούν να αντιστοιχηθούν σε αντικείμενα έτσι ώστε το στοιχείο να προσδιορίζει τη σχέση ανάμεσα στα αντικείμενα που αντιστοιχούν στη θέση του, *π.χ.* η σχέση "η απόσταση σε cm. ανάμεσα στα αντικείμενα *A* και *B*" προσδιορίζεται από τη τιμή 2. (*Σχημα Ε.2)* Αυτές οι σχέσεις θα καθορίζονται περιγραφικά και δε θα συμβολίζονται στον ίδιο τον πίνακα.

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 2 & 3 & 8 & 6 \\ 2 & 0 & 1 & 6 & 4 \\ 3 & 1 & 0 & 5 & 3 \\ 8 & 6 & 5 & 0 & 2 \\ 6 & 4 & 3 & 2 & 0 \end{pmatrix} \end{array}$$

Σχήμα Ε.2: Πίνακας αποστάσεων ανάμεσα στα σημεία A, B C, D, E

Το *πλάτος* του πίνακα ισοδυναμεί με τον αριθμό των σειρών του και το *μήκος* με τον αριθμό των στηλών του.

Η *διάσταση* ορίζεται από το πλάτος και το μήκος του πίνακα και συμβολίζεται $m \times n$, για πλάτος m και μήκος n.

Στη περίπτωση που απομονώνεται μια σειρά του πίνακα η καινούρια αυτή δομή ονομάζεται *διάνυσμα σειράς*, και αντίστοιχα για τις στήλες *διάνυσμα στήλης*, και, γενικά, οποιαδήποτε δομή που ισοδυναμεί με έναν μονοδιάστατο πίνακα ονομάζεται απλά *διάνυσμα*.

Όταν τα στοιχεία ενός πίνακα A με διαστάσεις $m \times n$ ανήκουν στο σύνολο των πραγματικών αριθμών τότε αυτός *ανήκει στο σύνολο των $m \times n$ πραγματικών πινάκων* και συμβολίζεται $A \in \mathbb{R}^{m \times n}$. Αντίστοιχα και για κάθε άλλο "τύπο" στοιχείων του πίνακα. Ο πίνακας του σχήματος 1.3 ανήκει στο σύνολο των 6×6 *πραγματικών πινάκων*, όπως και στο σύνολο 6×6 ακεραίων πινάκων και σε άλλους. Αν δύο πίνακες *ορισμένοι* σε διαφορετικά σύνολα πινάκων έχουν τις προϋποθέσεις για να ανήκουν στο ίδιο σύνολο πινάκων Χ τότε *ανάγονται* στο Χ.

Ε.3.2 Υπολογισμοί σε πίνακες

Ώστε να παρουσιαστούν οι βασικές υπολογιστικές πράξεις που τροποποιούν πίνακες θα υποτεθούν δύο πραγματικοί πίνακες A, B των οποίων οι διαστάσεις θα οριστούν ανά περίπτωση.[1] Αρχικά, κατ'αντιστοιχία με την αριθμητική, θα οριστεί η *πρόσθεση* : η πρόσθεση δύο πινάκων Α και Β προϋποθέτει ότι Α, Β έχουν τις *ίδιες διαστάσεις* (και ανήκουν ή ανάγονται στο *ίδιο σύνολο*, εδώ στο σύνολο των πραγματικών πινάκων), αφού ισχύουν τα παραπάνω, $A, B \in \mathbb{R}^{m \times n}$ το αποτέλεσμα της πρόσθεσης είναι ένας νέος πίνακας $C \in \mathbb{R}^{m \times n}$ ο οποίος αποτελείται από στοιχεία τιμής ίσης με τη πρόσθεση των αντίστοιχων στοιχείων των Α και Β. Η πρόσθεση των πινάκων A και B συμβολίζεται $A + B$.

$$A+B = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} & \dots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} & \dots & a_{2n}+b_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & a_{m3}+b_{m3} & \dots & a_{mn}+b_{mn} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \dots & c_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & c_{m3} & \dots & c_{mn} \end{bmatrix} = C$$

Σχήμα Ε.3: Πρόσθεση πινάκων

Κατ'επέκταση, ανάλογα με τη πρόσθεση δύο πινάκων, μπορούν να προστεθούν και περισσότεροι, π.χ. Α + Β + Γ + … + Ζ, τυπικά ίσο με Ϲ + Γ + … + Ζ και ούτω καθεξής προς κατάληξη στη πρόσθεση δύο πινάκων. Στη πρόσθεση πινάκων ισχύουν, επίσης, η *προσεταιριστική* ιδιότητα, δηλαδή (Α + Β) + Γ = Α + (Β + Γ) και η *αντιμεταθετική* ιδιότητα όπου Α + Β = Β + Α.

Ακολουθεί ο βαθμωτός (scalar) πολλαπλασιασμός όπου ένας αριθμός k πολλαπλασιάζεται με έναν πίνακα Α, το αποτέλεσμα είναι ένας νέος πίνακας $C = kA$ ο οποίος έχει τις ίδιες διαστάσεις με τον Α και αποτελείται από στοιχεία τιμής ίσης με τον πολλαπλασιασμό των αντίστοιχων του Α επί τον k.

$$kA = \begin{bmatrix} ka_{11} & ka_{12} & ka_{13} & \dots & ka_{1n} \\ ka_{21} & ka_{22} & ka_{23} & \dots & ka_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ ka_{m1} & ka_{m2} & ka_{m3} & \dots & ka_{mn} \end{bmatrix}$$

Σχήμα Ε.4: βαθμωτός πολλαπλασιασμός

1 Στη πραγματικότητα, οι πράξεις ανάμεσα σε πίνακες έχουν σχέση και με το είδος των στοιχείων τους, πέρα από τις συνθήκες που ισχύουν γενικά για τους πίνακες· δηλαδή, όταν πλέον η πράξη "περάσει" στα ίδια τα στοιχεία έχει τη σημασία που ορίζει το σύνολο στο οποίο ανήκουν.

Αφού παρουσιάστηκε ο βαθμωτός πολλαπλασιασμός είναι ευλογοφανές να παρουσιαστεί ο πολλαπλασιασμός μεταξύ πινάκων. Προϋπόθεση για το πολλαπλασιασμό δύο πινάκων είναι ο πρώτος πίνακας Α να έχει μήκος ίσο με το πλάτος του δεύτερου πίνακα Β, $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times r}$. Το αποτέλεσμα αυτού του πολλαπλασιασμού είναι ένας πίνακας $C \in \mathbb{R}^{m \times r}$ του οποίου τα στοιχεία υπολογίζονται με τον παρακάτω τύπο

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix} B = \begin{bmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1r} \\ b_{21} & b_{22} & b_{23} & \dots & b_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & b_{n3} & \dots & b_{nr} \end{bmatrix}$$

$$C = AB = \begin{bmatrix} (\mathbf{ab})_{11} & (\mathbf{ab})_{12} & (\mathbf{ab})_{13} & \dots & (\mathbf{ab})_{1r} \\ (\mathbf{ab})_{21} & (\mathbf{ab})_{22} & (\mathbf{ab})_{23} & \dots & (\mathbf{ab})_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (\mathbf{ab})_{m1} & (\mathbf{ab})_{m2} & (\mathbf{ab})_{m3} & \dots & (\mathbf{ab})_{mr} \end{bmatrix} (\mathbf{ab})_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Σχήμα Ε.5: Πολλαπλασιασμός πινάκων mxn, nxr

Σαν παράδειγμα θα υπολογιστεί το τελευταίο στοιχείο του πίνακα C στη θέση (m, r),

$$c_{mr} = a_{m1} b_{1r} + a_{m2} b_{2r} + a_{m3} b_{3r} + \dots + a_{mn} b_{nr} = \sum_{i=1}^{n} a_{mi} b_{ir} = (\boldsymbol{ab})_{\boldsymbol{mr}}$$

Σχήμα Ε.6: Υπολογισμός στοιχείου (m, r) πίνακα C

Η περίπτωση ενός πίνακα Α m×n διαστάσεων που πολλαπλασιάζεται με ένα n-διαστάσεων διάνυσμα Β υπόκειται στο γενικό τύπο πολλαπλασιασμού δύο πινάκων και έχει ως αποτέλεσμα ένα διάνυσμα m-διαστάσεων. Στον πολλαπλασιασμό πινάκων δεν ισχύει η αντιμεταθετική ιδιότητα, οπότε AB ≠ BA, ισχύει όμως η επιμεριστική ιδιότητα όπου A(B+C) = AB + AC και (B+C)A = BA + CA και η προσεταιριστική ιδιότητα , (AB)C = A(BC)

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \\ j & k & l \end{bmatrix} B = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$AB = \begin{bmatrix} ax + by + cz \\ dx + ey + fz \\ gx + hy + iz \\ jx + ky + lz \end{bmatrix}$$

Σχήμα Ε.7: Πολλαπλασιασμός πινάκων 4x3, 3x1

Ε.3.3 Τύποι και χαρακτηρισμοί πινάκων

Οι πίνακες χαρακτηρίζονται σύμφωνα με τη δομή τους ή τη σχέση τους με άλλους πίνακες, όταν, παραδείγματος χάριν, ο πίνακας προκύπτει από κάποιο μετασχηματισμό του αρχικού πίνακα.

Ο ανάστροφος πίνακας ενός πίνακα Α $m \times n$ είναι ο πίνακας A^T $n \times m$ του οποίου οι σειρές και οι στήλες αντιμεταθέτονται έτσι ώστε το στοιχείο στη θέση *{i, j}* στον ανάστροφο έχει τη θέση *{j, i}*. Ο ανάστροφος διαθέτει τις παρακάτω βασικές ιδιότητες, το αποτέλεσμα δύο διαδοχικών αναστροφών ενός πίνακα Α ταυτίζεται με τον αρχικό πίνακα, $(A^T)^T = A$· την επιμεριστική ιδιότητα στη πρόσθεση, όπου $(A + B)^T = A^T + B^T$, όπως και την επιμεριστική ιδιότητα στο πολλαπλασιασμό, όπου $(AB)^T = B^T A^T$, και όχι $A^T B^T$ αφού ο πολλαπλασιασμός πινάκων δε διαθέτει την αντιμεταθετική ιδιότητα· τέλος, την ομογένεια (πρώτου βαθμού) όπου $(\mu A)^T = \mu A^T$ όπου το βαθμωτό μ δεν "επηρεάζεται" από την αναστροφή.

$$A = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} A^T = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}$$

Σχήμα Ε.8: Ανάστροφος πίνακας

Ένας πίνακας ο οποίος έχει το ίδιο μήκος και πλάτος ονομάζεται τετραγωνικός πίνακας, οπότε είναι ένας πίνακας $n \times n$ ή αλλιώς, *τετραγωνικός πίνακας τάξης n*. Σύμφωνα με τις παραπάνω προϋποθέσεις πρόσθεσης δύο τετραγωνικοί πίνακες μπορούν να προστεθούν αν είναι της ίδιας τάξης. Το σύνολο των στοιχείων στις θέσεις *{i, i}*$\in$ *[1, n]* του πίνακα αποτελούν τη *διαγώνιο* και βάσει της διαγωνίου χαρακτηρίζονται διάφοροι τύποι τετραγωνικού πίνακα. Όταν όλα τα στοιχεία ενός τετραγωνικού (αριθμητικού) πίνακα n-τάξης είναι μηδενικά εκτός από τα στοιχεία της διαγωνίου, τότε ο πίνακας ονομάζεται *διαγώνιος πίνακας n-τάξης* και ειδικότερα όταν τα στοιχεία της διαγωνίου ενός διαγώνιου πίνακα n-τάξης είναι ο αριθμός 1 ο πίνακας ονομάζεται *ταυτοτικός πίνακας* n-τάξης και συμβολίζεται I_n. Ο ταυτοτικός πίνακας συμπεριφέρεται στον πολλαπλασιασμό περίπου όπως ο αριθμός 1 στην αριθμητική, όταν πολλαπλασιαστεί με έναν πίνακα Α $m \times n$ έχει ως αποτέλεσμα τον πίνακα Α, δηλαδή $AI_n = A$ αλλά και $I_m A = A$.

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \\ j & k & l \end{bmatrix} I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$AI_3 = \begin{bmatrix} 1a+0b+0c & 0a+1b+0c & 0a+0b+1c \\ 1d+0e+0f & 0d+1e+0f & 0d+0e+1f \\ 1g+0h+0i & 0g+1h+0i & 0g+0h+1i \\ 1j+0k+0l & 0j+1k+0l & 0j+0k+1l \end{bmatrix} =$$

$$= I_4 A = \begin{bmatrix} 1a+0d+0g+0j & 1b+0e+0h+0k & 1c+0f+0i+0l \\ 0a+1d+0g+0j & 0b+1e+0h+0k & 0c+1f+0i+0l \\ 0a+0d+1g+0j & 0b+0e+1h+0k & 0c+0f+1i+0l \\ 0a+0d+0g+1j & 0b+0e+0h+1k & 0c+0f+0i+1l \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \\ j & k & l \end{bmatrix}$$

Σχήμα Ε.9: Πολλαπλασιασμός με μοναδιαίο πίνακα

Όταν τα στοιχεία *κάτω* από τη διαγώνιο είναι μηδενικά, ο πίνακας ονομάζεται άνω-τριγωνικός και αντίστοιχα όταν τα στοιχεία πάνω από τη διαγώνιο είναι μηδενικά, ο πίνακας ονομάζεται κάτω-τριγωνικός. Ο διαγώνιος πίνακας μπορεί να θεωρηθεί ως άνω- και κάτω-τριγωνικός συγχρόνως.

Με αυτά δεδομένα, ένας τετραγωνικός πίνακας Α ονομάζεται *συμμετρικός* όταν ισούται με τον ανάστροφό του, δηλαδή όταν $A = A^T$ και λοξά-συμμετρικό όταν ισούται με το αρνητικό

του ανάστροφου, $A = -A^T$. Επιπλέον, ένας τετραγωνικός πίνακας Α είναι *αντιστρέψιμος* όταν υπάρχει τετραγωνικός πίνακας Β και το παράγωγο του πολλαπλασιασμού τους, ΑΒ ή ΒΑ είναι ο ταυτοτικός πίνακας Ι· ο πίνακας Β ονομάζεται *αντίστροφος* του Α και συμβολίζεται A^{-1}. Ο τετραγωνικός πίνακας Α του οποίου ο αντίστροφος ταυτίζεται με τον ανάστροφο $A^T = A^{-1}$ ονομάζεται *ορθογώνιος* πίνακας, οπότε $AA^T = I$.

Ε.3.4 Ιδιοδιανύσματα και Ιδιοτιμές

Για να αναδειχτούν οι έννοιες του *ιδιοδιανύσματος (eigenvector)* και *της ιδιοτιμής (eigenvalues)* θα παρουσιαστεί αρχικά ένα απλό παράδειγμα. Υπάρχει τετραγωνικός πίνακας Α *n×n* και διάνυσμα Β *n-τάξης*, όπως στο παράδειγμα παρακάτω στο οποίο φαίνεται πως το αποτέλεσμα του πολλαπλασιασμού τους είναι το ίδιο το διάνυσμα Β. Αυτό σημαίνει πως ο πολλαπλασιασμός του πίνακα Α με το διάνυσμα Β ισοδυναμεί στο βαθμωτό πολλαπλασιασμό του διανύσματος Β με το 1.

$$A = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix} \quad B = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$AB = \begin{bmatrix} 1/3 + 1/3 + 1/3 \\ 1/2 + 0 + 1/2 \\ 1/4 + 1/2 + 1/4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = B$$

Σχήμα Ε.10: Ιδιοτιμή και Ιδιοδιάνυσμα πίνακα

Το διάνυσμα Β είναι ένα *ιδιοδιάνυσμα* του πίνακα Α και ο αριθμός 1 η *ιδιοτιμή* του. Από τα παραπάνω προκύπτει ότι $AB = 1B \Leftrightarrow AB - 1B = 0 \Leftrightarrow (A - 1I)B = 0$ και πιο γενικά $(A - \lambda I)B = 0$ όπου το ιδιοδιάνυσμα αντιστοιχεί με το διάνυσμα Β και η ιδιοτιμή με τον αριθμό λ. Ακόμα γενικότερα, από τον τύπο του πολλαπλασιασμού πίνακα με πίνακα στη περίπτωση διανυσμάτων προκύπτει ότι $w_i = \sum A_{ij} B_j$, αν το w_i και B_j είναι βαθμωτά πολλαπλάσια, δηλαδή υπάρχει σχέση $AB = w_i = \lambda B$ τότε το διάνυσμα Β και ο αριθμός λ μπορούν να οριστούν ως ένα ιδιοδιάνυσμα και μια ιδιοτιμή του Α, αντίστοιχα. Το σύνολο Ε όλων των διανυσμάτων *ν* που ικανοποιούν τη συνθήκη $\nu : (A - \lambda I)\nu = 0$ αποκαλείται *ιδιοχώρος (eigenspace)* του Α.

Ε.4 Αναπαραστάσεις γράφου με πίνακες

Εφόσον, καθ'ορισμού, ένας γράφος *G* προσδιορίζεται από το σύνολο των ακμών και το σύνολο των κόμβων του, μπορεί να ανασυσταθεί αντιπροσωπευτικά σε άλλες αναπαραστάσεις, αρκεί να ληφθούν υπ'όψιν και να διατηρηθούν τα στοιχεία των παραπάνω συνόλων. Συνήθως ένας γράφος αναπαρίσταται για υπολογιστικούς λόγους και μια τέτοια αναπαράσταση γίνετε κυρίως με πίνακες αφού επιτρέπουν υπολογισμούς και μετασχηματισμούς.

Αρχικά, πρέπει να φανούν οι δυνατοί τρόποι αναπαράστασης με πίνακα. Προκειμένου να αποκτήσει δομή ο πίνακας του γράφου πρέπει πρώτα να αντιστοιχηθούν στοιχεία από τα σύνολα του γράφου στις σειρές και τις στήλες του πίνακα. Υπάρχουν τρεις δυνατότητες· πρώτον, οι σειρές να αντιστοιχούν σε κόμβους και οι στήλες σε ακμές ή αντίστροφα, με αναστροφή, δεύτερον, οι κόμβοι να αντιστοιχηθούν και στις σειρές και στις στήλες ή τρίτων, να γίνει το ίδιο με το σύνολο των ακμών.

Στη περίπτωση που και στήλες και σειρές είναι οι κόμβοι του πίνακα, υπάρχουν τέσσερις τρόπου που μπορούν να αναπαρασταθούν οι σχέσεις τους και από αυτούς προκύπτουν τετραγωνικοί πίνακες NxN. Όταν στον πίνακα εγγράφονται οι βαθμοί των κόμβων του γράφου, αυτός ονομάζεται *πίνακας κόμβων (degree matrix Deg)* και έχει όλα τα στοιχεία μηδενικά εκτός από της διαγωνίου όπου κάθε {i, i} = $d(v_i)$. Ακόμα, μπορούν να αναπαρασταθούν οι σχέσεις γειτονίας ανάμεσα σε δύο κόμβους, με τον πίνακα γειτνίασης. Για μη κατευθυνόμενους γράφους ο *πίνακας γειτνίασης (adjacency matrix Adj)* εγγράφει στη σειρά κάθε κόμβου vi, όσους κόμβους συνδέονται με ακμή μαζί του. Το αποτέλεσμα είναι ένας συμμετρικός πίνακας με τη τη διαγώνιο μηδενική, σε περίπτωση που απαγορεύονται οι αυτοσυνδέσεις. Όσο για τον κατευθυνόμενο γράφο προκύπτει μη συμμετρικός πίνακας. Υπάρχουν ακόμα ο πίνακας ελαχίστων αποστάσεων που εγγράφει την ελάχιστη απόσταση ανάμεσα στους κόμβους και είναι συμμετρικός *(distance matrix)* και ο *Λαπλασιανός πίνακας (Laplacian matrix)* που υπολογίζεται ως Lap = Deg – Adj, δηλαδή η διαγώνιος είναι οι βαθμοί και τα υπόλοιπα στοιχεία η γειτνίαση με αρνητικό πρόσημο. Άλλη μια σημαντική αναπαράσταση είναι ο *πίνακας εφαπτόμενων (incidence matrix)* στον οποίο για κάθε κόμβο εγγράφονται οι εφαπτόμενες σε αυτών ακμές. |V|x|E|. Ένα παράδειγμα φαίνεται παρακάτω στα σχήματα.

$$Adj = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

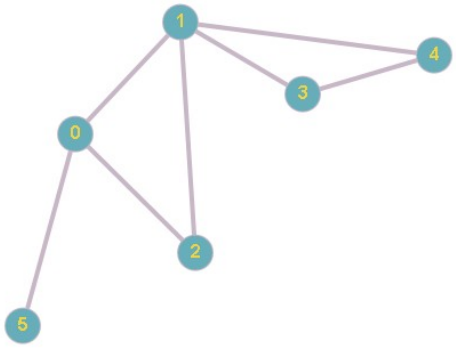
Σχήμα Ε.11: Πίνακας γειτνίασης γράφου

$$Deg = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Σχήμα Ε.12: Πίνακας βαθμών γράφου

$$Inc = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Σχήμα Ε.13: Πίνακας εφαπτόμενων γράφου



Σχήμα Ε.14: Γράφος (παράδειγμα)

ΚΕΦΑΛΑΙΟ Ι. Αναπαράσταση Κειμένων ως Γράφων

Το βασικό κριτήριο που καθορίζει τη φύση του γράφου είναι το επίπεδο στο οποίο θα μετασχηματιστεί και τι σχέσεις θα περιγράφει. Ως προς το επίπεδο, οποιεσδήποτε *σημασιολογικές μονάδες* του κειμένου μπορούν να επιλεχθούν ως όροι οι οποίοι θα αποτελέσουν και τους κόμβους του γράφου. Εν προκειμένω, σημασιολογικές μονάδες θεωρούνται τα μορφήματα, οι λέξεις, οι φράσεις, οι προτάσεις και οι παράγραφοι, οι οποίες θεωρητικά περιέχουν σημασιολογικές πληροφορίες. Όπως και οποιαδήποτε άλλη μονάδα ενδέχεται να έχει νόημα για μια συγκεκριμένη εφαρμογή . Οι όροι μπορούν να εξαχθούν από το κείμενο με την ανάλογη, στο επίπεδο, τμηματοποίηση κατά την προεπεξεργασία. Επίσης η διάκριση μορφήματος και λέξεις από εδώ και πέρα θα υποδηλώνεται όποτε αναφέρεται ότι επιτελείται αποκοπή καταλήξεων (stemming) στις λέξεις. Αν το επίπεδο στο οποίο θα σχηματιστεί ο γράφος αποτελείται από έναν μόνο τύπο σημασιολογικής μονάδας, τότε ο γράφος θα ονομάζεται *ομοιογενής*. Αντίθετα, αν αποτελείται από περισσότερες των δύο τύπων μονάδες ο γράφος θα ονομάζεται *ανομοιογενής*.

Αφού καθοριστεί το επίπεδο μετασχηματισμού, οι ακμές που σχηματίζονται ανάμεσα στους κόμβους θα εξαρτώνται άμεσα από αυτό. Στη περίπτωση ενός ομοιογενούς γράφου στο λεκτικό επίπεδο π.χ. οι ακμές μπορούν να δηλώνουν στατιστικά χαρακτηριστικά, όπως σχέσεις συνεμφάνισης σε παράθυρο (στατικού ή δυναμικού μεγέθους) ή πρόταση· σημασιολογικά χαρακτηριστικά, λ.χ. σημασιολογική ομοιότητα, ιεραρχικές σχέσεις (υπωνυμία)· συντακτικά χαρακτηριστικά κ.α. Οι ακμές μπορούν, ακόμα, να είναι κατευθυνόμενες, όπου περιγράφεται μια σχέση εξάρτησης (π.χ. επίθετο->ουσιαστικό), και να έχουν βάρη. Τα ανομοιογενή στοιχεία συνηθίζεται να σχηματίζουν διμερής γράφους, το οποίο, όμως, δεν αποτελεί κανόνα.

Οι συγκεκριμένες κατασκευές γράφων από κείμενα εξαρτώνται από τις, ρητές και άρρητες, υποθέσεις κατά την ανάπτυξη των συστημάτων, τον σκοπό του καθενός και τη φύση της κάθε εφαρμογής στην οποία τοποθετείται αυτός ο σκοπός. Οι γράφοι κειμένων έχουν πολυάριθμες εφαρμογές, όπως π.χ. για να βρούνε συντακτικά μοτίβα σε χειρόγραφα προς αποκρυπτογράφηση βάσει των συστάδων των γράφων (Sinha et al., 2009 αναφ. Blanco & Lioma 2012). Κατά την έκθεση των σημαντικότερων μοντέλων συστημάτων, παρακάτω, θα φανούν διάφοροι τρόποι παραγωγής γράφου από κείμενα που έχουν εφαρμοστεί στη πράξη. Στην παρούσα εργασία, κατά κύριο λόγο, παρουσιάζονται δομές γράφων για την ποσοτικοποίηση της σημαντικότητας των εκάστοτε μονάδων που εγγράφονται ως κόμβοι, σαν εναλλακτική σε μεθόδους βασισμένες στη συχνότητα[2] των όρων στα κείμενα σε εφαρμογές Επεξεργασίας Φυσικής Γλώσσας και Ανάκτησης Πληροφοριών.

Ο απλούστερος τρόπος υπολογισμού του βαθμού σημαντικότητας ενός κόμβου είναι μέσω του βαθμού (degree) του· η υπόθεση που το στηρίζει είναι πως όσους περισσότερους γείτονες έχει ένας κόμβος τόσο πιο σημαντικός πρέπει να είναι. Σε κατευθυνόμενους γράφους ο παραπάνω υπολογισμός γίνεται με το βαθμό εισόδου του κάθε κόμβου, δηλαδή από τον αριθμό εξαρτημένων από αυτόν άλλων κόμβων, βλ. εξ. (1) και (2). Σε γράφους κειμένων και σε αντιστοιχία με παλιότερα μοντέλα που βασίζοντας στη συχνότητα των όρων, μπορούν να προστεθούν στον παραπάνω υπολογισμό κι άλλα χαρακτηριστικά για την κανονικοποίηση του. Η εργασία των Rousseau & Vazigiannis (2013) πειραματίστηκε με την παραπάνω ιδέα στο πεδίο της ανάκτησης πληροφοριών εμπνευσμένοι από τους αξιωματικούς κανόνες περιορισμού που έθεσαν οι Fang et al. (2004) και Lv & Zhai (2011) για μεθόδους με συχνότητα όρου σε εφαρμογές ανάκτησης. Κατέληξαν πως η κανονικοποίηση κατά το μεγέθος του κειμένου (συγκριτικά μικρότερου βαθμού από τις μεθόδους με συχνότητα όρων) αυξάνει τις επιδόσεις

2 Η λέξη συχνότητα στις περισσότερες εφαρμογές είναι παραπλανητικός και σημαίνει απλά τον αριθμό εμφανίσεων των όρων μέσα στο κείμενο. Ωστόσο, σε μερικές εργασίες διαιρείται όντως με το σύνολο των όρων του κειμένου.

(length normalization constraint), όπως και η αντίστροφος της συχνότητας των κειμένων που περιέχουν τον όρο (IDF, term discrimination constraint), βλ. εξ (3). Άλλα χαρακτηριστικά που αναπτύχθηκαν αξιωματικά είναι α) η μη γραμμική, αλλά κοίλη συσχέτιση ανάμεσα στη συχνότητα και τη βαθμολογία του όρου ώστε, για παράδειγμα η αλλαγή στη βαθμολογία κατά την αύξηση της συχνότητας από το 1 στο 2 πρέπει να είναι πιο ραγδαία από το 100 στο 101 (term-frequency constraint)· β) η κάτω φραγή της συνάρτησης ώστε να διατηρείται η πληροφορία των συχνοτήτων ενός όρου η οποία σχεδόν μηδενίζεται από τη κανονικοποίηση κατά μέγεθος για πολύ μεγάλα κείμενα (lower-bounding constraint). Μια παρόμοια αξιωματική εργασία είναι δυνατό να αναπτυχθεί και για τους περιορισμούς που πρέπει να πληρούν βαθμολογήσεις σημαντικότητας σε γράφους κειμένων για την εκάστοτε εφαρμογή.

$$S_{\beta}(v_i)=d(v_i) \tag{1}$$

$$S_{\beta\sim}(v_i)=\left|\mathrm{In}(v_i)\right| \tag{2}$$

$$S(v_i)=\frac{\left|\mathrm{In}(v_i)\right|}{1-b+b\times\frac{|d|}{avdl}}\times\log\frac{N+1}{df(t)} \tag{3}$$

Βαθμολόγηση βάσει degree με κανονικοποίηση,
Rousseau & Vazigiannis (2013)

Υπάρχει, όμως, και μια πιο ανεπτυγμένη υπόθεση που μπορεί να βοηθήσει στη βαθμολόγηση σημαντικότητας: Η σημαντικότητα ενός κόμβου είναι ανάλογη με τη σημαντικότητα των γειτόνων του. Αυτή η ιδέα διατυπώθηκε, πρώτα, για τη βαθμολόγηση διαδικτυακών σελίδων βάσει των υπερσυνδέσμων τους. Πάνω σε αυτή αναπτύχθηκαν δύο διαφορετικοί αλγόριθμοι, ο HITS και ο PageRank (Kleinberg, 1997· Brin & Page, 1998, αντίστοιχα.). Και οι δύο αλγόριθμοι βασίζονταν σε κατευθυνόμενους γράφους αφού οι ακμές του ορίζονταν ως υπερσύνδεσμοι οι οποίοι έχουν κατεύθυνση.

Ο HITS είναι επαναληπτικός αλγόριθμος όπου βαθμολογεί τους κόμβους βάσει δύο κριτηρίων, τον βαθμό εισόδου και τον βαθμό εξόδου τους. Οι κόμβοι με υψηλό βαθμό εξόδου ονομάζονται κεντρικοί (hubs) και θεωρούνται σελίδες-κατάλογοι οι οποίες αναφέρονται σε πολλές άλλες σελίδες. Από την άλλη μεριά, οι κόμβοι με υψηλό βαθμό εισόδου, που ονομάζονται αυθεντίες (authorities), θεωρούνται σελίδες με σημαντικές πληροφορίες αφού πολλές άλλες σελίδες αναφέρονται σε αυτές. Οι βαθμολογία αυθεντίας και κεντρικότητας κάθε κόμβου υπολογίζεται στις εξισώσεις (4) και (5), όπου συνυπολογίζονται επαναληπτικά η κεντρικότητα και η αυθεντία των γειτόνων του και βασίζεται στην αρχή της αμοιβαίας ενίσχυσης (Principle of Mutual Reinforcement). Ο αλγόριθμος επαναλαμβάνεται έως ότου να υπάρξει σύγκλιση[3] κάτω από ένα ορισμένο κατώφλι. Τέλος, ο βαθμός κεντρικότητας που προκύπτει είναι ανάλογος της σημαντικότητας της σελίδας ως προς τις πληροφορίες που διαθέτει.

3 Σύγκλιση υπάρχει όταν η διαφορά στα αποτελέσματα οποιουδήποτε κόμβου δύο διαδοχικών επαναλήψεων (το ποσοστό σφάλματος) πέφτει κάτω από ένα ορισμένα εκ των προτέρων κατώφλι. (Mihalcea & Tarau, 2004 σ.)

$$HITS_A(v_i)=\sum_{Vj\in \mathrm{In}(v_i)} HITS_H(v_j) \quad (4)$$

$$HITS_H(v_i)=\sum_{Vj\in \mathrm{Out}(v_i)} HITS_A(v_j) \quad (5)$$

Ο PageRank είναι αναδρομικός αλγόριθμος και βαθμολογεί τους κόμβους σύμφωνα με το βαθμό εισόδου τους. Η σημαντικότητα ενός κόμβου υπολογίζεται, όπως παραπάνω για τους κατευθυνόμενους γράφους, από τον βαθμό εισόδου του και ενισχύεται αν οι κόμβοι που δείχνουν σε αυτόν είναι οι ίδιοι σημαντικοί. Οπότε όταν σημαντικές σελίδες αναφέρονται σε άλλες σημαντικές σελίδες τους δίνουν κύρος. Ο PageRank δίνει μια πιθανότητα λ να ακολουθηθούν οι ακμές ενός κόμβου και στη περίπτωση που δεν συμβεί αυτό, επιλέγεται ένας τυχαίος κόμβος του γράφου. Το λ ενσωματώνεται στον αλγόριθμο ώστε να υπάρχει σίγουρη σύγκλιση και ορίζεται συνήθως στα 85%. Όταν ο αλγόριθμος ξεκινάει οι κόμβοι κατέχουν αυθαίρετες βαθμολογίες (π.χ. S=1), μόλις ο αλγόριθμος συγκλίνει κάτω από ένα ορισμένο κατώφλι, το αποτέλεσμα είναι ένα διάνυσμα με τις τελικές βαθμολογίες κάθε κόμβου. Ο υπολογισμός του PageRank κατευθυνόμενου γράφου χωρίς βάρη όπως περιγράφεται εδώ εμφανίζεται στην εξίσωση (7).

Υπάρχουν καλοί λόγοι να υποτεθεί ότι μονάδες μέσα σε ένα κείμενο συμπεριφέρονται παρόμοια με τις σελίδες στο διαδίκτυο όταν οριστούν συγκεκριμένες σχέσεις μεταξύ τους. **(**βλ. Κεφ. ΙΙ**)**. Πάνω σε αυτή τη βάση, ο αλγόριθμος PageRank και HITS μπορούν να βοηθήσουν στον υπολογισμό του βαθμού σημαντικότητας των όρων ενός ή περισσότερων περισσότερων κειμένων. Σε διάφορες εφαρμογές εμφανίζονται παραλλαγές του PageRank για διάφορα είδη γράφων, βλ εξ. (6), (8), (9) και για να πειραχτεί η πιθανότητα ο αλγόριθμος να περάσει σε έναν τυχαίο κόμβο, εισάγοντας έτσι τη προκατάληψη υπέρ ή εναντίον ορισμένων όρων βλ. εξ 10.

(6)
$$S_{PR}(v_i)=\lambda\sum_{j\in\Gamma(v_i)}\frac{1}{|\Gamma(v_j)|}S(v_j)+(1-\lambda)$$

PageRank μη κατευθυνόμενου γράφου χωρίς βάρη

$$S_{\mathrm{PR\sim d}}(v_i)=\lambda\sum_{j\in \mathrm{In}(v_i)}\frac{1}{Out(v_j)}S(v_j)+(1-\lambda) \quad (7)$$

PageRank κατευθυνόμενου γράφου χωρίς βάρη

(8)
$$S_{\mathrm{PR\sim w}}(v_i)=\lambda\sum_{j\in\Gamma(v_i)}\frac{w_{ij}}{\sum_{Vk\in\Gamma(V_j)}w_{jk}}S(v_j)+(1-\lambda)$$

PageRank μη κατευθυνόμενου γράφου με βάρη

$$S_{\mathrm{PR\sim dw}}(v_i)=\lambda\sum_{j\in \mathrm{In}(v_i)}\frac{w_{ji}}{\sum_{Vk\in \mathrm{Out}(V_j)}w_{jk}}S(v_j)+(1-\lambda) \quad (9)$$

PageRank κατευθυνόμενου γράφου με βάρη

(10)
$$S_{\mathrm{PR\sim b}}(v_i)=\lambda\sum_{j\in\Gamma(v_i)}\frac{1}{|\Gamma(v_j)|}S(v_j)+(1-\lambda)\cdot\pi$$

Προκατειλημμένος PageRank μη κατευθυνόμενου γράφου χωρίς βάρη

ΚΕΦΑΛΑΙΟ ΙΙ. Το γλωσσικό δίκτυο

Όπως έχει φανεί, η δομή γράφου μπορεί να χρησιμοποιηθεί για τη περιγραφή οποιασδήποτε κατάστασης η οποία περιέχει ορισμένες μονάδες και σχέσεις μεταξύ αυτών. Όταν η θεωρία γράφων εφαρμόζεται σε πραγματικά αντικείμενα, τότε μπορεί να γίνει λόγος για πολύπλοκα δίκτυα και θεωρία δικτύων αντίστοιχα. Παραπάνω δόθηκαν οι δυνατότητες κατασκευής γράφων από κείμενα και κάποια εργαλεία για την εξαγωγή χρήσιμων πληροφοριών από αυτά για χρήση σε διάφορες εφαρμογές, όπως θα περιγραφούν σε άλλα σημεία της εργασίας. Πέρα από τη χρησιμότητα των τεχνικών δικτύων πάνω σε πραγματικές εφαρμογές, η δομή αυτή μπορεί να βοηθήσει σημαντικά και σε θεωρητικά ζητήματα της γλωσσολογίας. Μια μελέτη τέτοιου τύπου, όπου μαθηματικοποιούνται, με τον ένα ή τον άλλο τρόπο, πτυχές της γλώσσας είναι εργασία του κλάδου της υπολογιστικής γλωσσολογίας. Πολλές φορές, ακόμα, κάποια θεωρητικά αποτελέσματα της γλωσσολογία αναφαίνονται σε υποθέσεις συστημάτων που επιτελούν ένα συγκεκριμένο έργο.

Πριν γίνει λόγος για τα συγκεκριμένα ζητήματα της γλώσσας, πρέπει να περιγραφούν κάποια μοντέλα γράφων, τα οποία κατασκευάζονται με συγκεκριμένο τρόπο και εμφανίζουν συγκεκριμένα χαρακτηριστικά. Εναρκτήριο σημείο για τη μελέτη μοντέλων μπορεί να θεωρηθεί το μοντέλο του Τυχαίου Γράφου (Random Graph Model). Αναπτύχθηκε από τους Erdös & Rényi (1959) και από τον Gilbert (1959) και εδώ θα περιγραφεί, ειδικά, το μοντέλο Erdös-Rényi (E-R). Ένας τυχαίος γράφος G(V, E) τύπου E-R παράγεται όταν για έναν αριθμό n κόμβων και από το σύνολο των πιθανών ακμών ανάμεσα τους επιλέγεται ένα τυχαίο υποσύνολο E. Για την επιλογή κάθε ακμής του συνόλου VxV υπάρχει μια πιθανότητα σύνδεσης p να ανήκει στο E και το αποτέλεσμα είναι ένας γράφος όμοιος με του *σχήματος 2.1*. Σε έναν τυχαίο γράφο η κατανομή βαθμού είναι μια διωνυμική κατανομή (ή και Poisson) με παραμέτρους p και n-1 εξ. (1). Οι γράφοι αυτού του είδους παρουσιάζουν μικρό μέσο μέγεθος διαδρομής και υψηλό συντελεστή συσταδοποίησης. Στον αντίποδα των τυχαίων γράφων είναι οι k-κανονικοί γράφοι με χαμηλό συντελεστή συσταδοποίησης και μεγάλο μέσο μέγεθος διαδρομής.

Οι Watts & Strogatz (1998), με τη σειρά τους, ανέπτυξαν ένα τρόπο παραγωγής τυχαίων γράφων με παράμετρο την αταξία (disorder) α του γράφου, το μοντέλο Watts-Strogatz. Το α είναι η πιθανότητα να επανεγγραφεί το ένα άκρο μιας ακμής ενός κανονικού γράφου σε οποιονδήποτε άλλο κόμβο για κάθε ακμή. Για α = 0 το αποτέλεσμα είναι ο κανονικός γράφος και για α=1 μια προσέγγιση του τυχαίου γράφου E-R. Ρύθμισαν την παράμετρο α για να πετύχουν, συγχρόνως, υψηλό συντελεστή συσταδοποίησης και μικρό μέσο μέγεθος διαδρομής βασισμένοι στη παρατήρηση ότι τα πραγματικά δίκτυα συνηθίζεται να εμφανίζουν αυτά τα χαρακτηριστικά και ονομάστηκαν ιδιότητες μικρού κόσμου (small-world properties). Στα δίκτυα με ιδιότητες μικρού κόσμου μπορεί κανείς να διασχίσει σε λίγα βήματα τη διαδρομή ανάμεσα σε δύο κόμβους και διαθέτουν κλίκες από συνδεδεμένους κόμβους, π.χ. όπως στα δίκτυα φίλων. Παράλληλα, ένα άλλο μοντέλο παραγωγής τυχαίων γράφων αναπτύχθηκε στην εργασία των Barabási & Albert (1999). Παρατηρήθηκε πως πολλά πραγματικά δίκτυα είναι μη-κλιμακούμενα (scale-free), δηλαδή οι βαθμοί των κόμβων τους ακολουθούν μια κατανομή νόμου δύναμης (power-law distribution) της μορφής **(**2α**)** με a=1, ε=0 και γ = -3 βλ. εξ (2β). Στο πρώτο στάδιο της παραγωγής του γράφου τύπου B-A δημιουργούνται m_0 συνδεδεμένοι κόμβοι. Κάθε καινούριος κόμβος συνδέεται σε έναν υπάρχοντα κόμβο i με πιθανότητα p_i ανάλογη του βαθμού του, δηλαδή οι κόμβοι με μεγάλο βαθμό ελκύουν περισσότερους καινούριους κόμβους βλ.εξ. (3). Αυτός ο τρόπος επιλογής ονομάζεται προνομιακή προσάρτηση (preferential attachment) και η συγκεκριμένη πιθανότητα παράγει δίκτυα που ακολουθούν κατανομή νόμου δύναμης. Προνομιακή προσάρτηση παρατηρείται σε πολλά πραγματικά δίκτυα, όπως στα δίκτυα υπερσυνδέσμων στο διαδίκτυο όπως φάνηκε παραπάνω ή στα δίκτυα αναφορών στα επιστημονικά άρθρα. Το πρόβλημα που παρουσιάζουν τα δίκτυα με ιδιότητες μικρού κόσμου

κατά την αντιπροσώπευση πραγματικών δικτύων είναι ότι δεν περιγράφουν την την προνομιακή προσάρτηση, έτσι ώστε αν και ο συντελεστής συσταδοποίησης και το μέσο μέγεθος διαδρομής προσεγγίζουν τα πραγματικά δεδομένα, η κατανομή βαθμών τους δεν ταιριάζει. Αντίθετα, το μοντέλο B-A προσεγγίζει την κατανομή όμως δεν μπορεί να παράγει τον παρόμοιο συντελεστή συσταδοποίησης. Για περισσότερες πληροφορίες πάνω στα τυχαία δίκτυα βλ. Dorogovtzev & Mendes (2002) και Chung (2008).

Η γλώσσα ως πολύπλοκο δίκτυο όταν μελετάται στις εμφανίσεις της στην ομιλία ή στη γραφή παρουσιάζει υπό όρους χαρακτηριστικά των τυχαίων γράφων με ιδιότητες μικρού κόσμου και με προνομιακή προσάρτηση. Ένας γράφος γλώσσας μπορεί να δημιουργηθεί από κείμενα, όπως παραπάνω, με κόμβους τις λέξεις των κειμένων και ακμές είτε γειτονίες λέξεων μέσα στο κείμενο, η συνεμφανίσεις μέσα σε ένα ορισμένο παράθυρο ή πρόταση. Η ιδιότητες μικρού κόσμου της γλώσσας και η κατανομή βαθμού των κόμβων του γράφου της περιγραφήκαν στην εργασία των i Cancho & Solé (2001) οι οποίοι κατασκεύασαν έναν γράφο συνεμφάνισης σε παράθυρο μεγέθους 2 μέσα στη πρόταση. Υπολογίζοντας τον συντελεστή συσταδοποίησης και το μέσο μέγεθος διαδρομής αυτού του παραπάνω γράφου ανακάλυψαν ότι τα αποτελέσματα ταίριαζαν με αυτά που παράγουν οι τυχαίοι γράφοι με ιδιότητες μικρού κόσμου. Επίσης, κατασκευάζοντας έναν επιπλέον γράφο όπου οι ακμές σήμαιναν τη γειτονία δύο λέξεων στη πρόταση με πιθανότητα εμφάνισης πάνω από ένα κατώφλι βρήκαν τη κατανομή βαθμού του γράφου η οποία είχε καλή προσαρμογή πάνω στην κατανομή νόμου δύναμης με γ=-2.7 και σε αυτό συμφωνεί με το μοντέλο B-A. Βασιζόμενοι στα προηγούμενα οι Kulig et al. (2015) κατασκευάζουν δίκτυα κειμένων πέντε διαφορετικών γλωσσών με κόμβους λέξεις και ακμές γειτονίες λέξεων. Αρχίζοντας από τη πρώτη λέξη του κειμένου και εξελίσσοντας το δίκτυο με κάθε καινούρια λέξη εξετάζουν τις αυξομειώσει στο μέσο μέγεθος διαδρομής και τον αριθμό των κόμβων κατά τη διαδικασία. Κατά την εξέλιξη του δικτύου υπάρχει ένα σημείο που το μέσο μέγεθος διαδρομής φτάνει στο μέγιστο του, περίπου L=10 για N<100, και στη συνέχεια πέφτει και σταθεροποιείται στο L=5 για N>1000. Το ύφος κάθε συγγράμματος αναφαίνεται κατά τη διάρκεια όπου η διαδρομή φτάνει στα μέγιστα και όταν ο γράφος μεγαλώσει πάρα πολύ τότε σχεδόν όλα τα κείμενα της ίδιας γλώσσας ακολουθούν περίπου τον ίδιο μέσο και μπορεί να υποτεθεί ότι μεγάλα δίκτυα περιγράφουν χαρακτηριστικά της ίδιας της γλώσσας. Επίσης κατά την εξέλιξη του γράφου κειμένου η κατανομή βαθμού αν και είναι κατανομή παρόμοια με των B-A έχει δύο φάσεις, μέχρι τον βαθμό $ct^{-1}(2 + ct)^{3/2}$ (όπου ct είναι ο αριθμός των καινούριων ακμών μεταξύ των κόμβων) το γ = -3/2 και το α = ½ ενώ μετά από αυτό το σημείο το γ = -3 και το α = $1/4(t)^2$ (όπου t το βήμα (το σημείο χρόνου) κατασκευής) (*σχήμα 2.2*). Τέλος, οι i Cancho et al. (2004) ερευνώντας τις ιδιότητες γράφων συντακτικής εξάρτησης (Mel′čuk, 1988) αποφάνθηκαν πως, από τις ιδιότητες που ήδη έχουν περιγραφεί, εμφανίζουν ιδιότητες μικρού κόσμου, είναι μη-κλίμακούμενοι με γ=2.2.

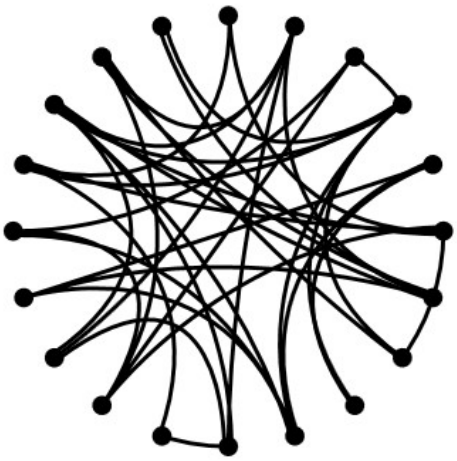
Η παραπάνω έκθεση ήταν ενδεικτική και απλουστευμένη για τη χρήση της θεωρίας γράφων στο πεδίο της γλωσσολογίας. Το πεδίο είναι σχετικά καινούριο όμως έχουν ήδη υπάρξει πολλά αποτελέσματα και αποσαφηνίσεις σε παλαιότερες ιδέες. Για περισσότερα πάνω στο θέμα βλ. Mehler et al. (2016). Όπως θα φανεί, κάποια σημεία από αυτά τα ευρήματα θα χρησιμοποιηθούν σε διάφορες εφαρμογές επεξεργασίας φυσικής γλώσσας και ανάκτησης πληροφοριών που βασίζονται σε γράφους.

$$P(d(u)=k)=\binom{n}{k} p^d(1-p)n-k \quad (1)$$

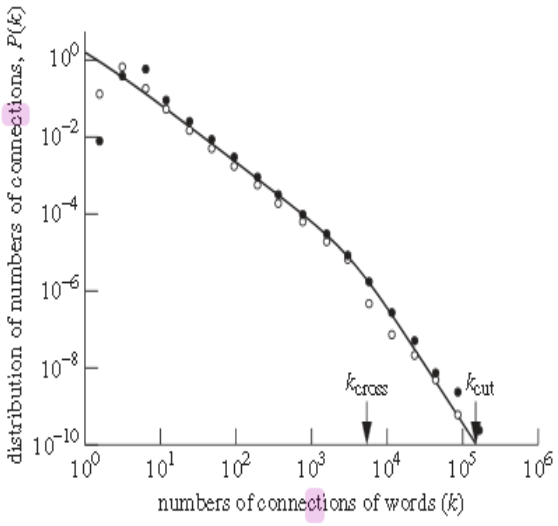
$$p_i=\frac{d_i}{\sum_j d_j} \quad (3)$$

$$f(x)=ax^{\gamma}+\varepsilon \quad (2\alpha)$$

$$P(d(u)=k)=k^{-3} \quad (2\beta)$$



Σχήμα 2.1: Τυχαίος γράφος όπως εμφανίζεται στο Watts & Strogatz (1998)



Σχήμα 2.2: Κατανομή βαθμού σε λογαριθμική κλίμακα για t = 470000 όπως εμφανίζεται στο Dorogotzev & Mendes (2001)

ΚΕΦΑΛΑΙΟ ΙΙΙ. Αυτόματη Εξαγωγή Φράσεων(/Λέξεων)-Κλειδιών (Automatic Keyphrase(/word) Extraction)

ΙΙΙ.1 Εισαγωγή.

Αρχικά, πρέπει να επισημανθεί ότι υπάρχει μια σύγχυση στη διαφοροποίηση των όρων φράση-κλειδί και λέξη-κλειδί. Ο βασικός λόγος για αυτή τη σύγχυση είναι η πρακτική των επιστημονικών περιοδικών να ζητάνε λέξεις-κλειδιά (keywords) από τους συγγραφείς των άρθρων, εννοώντας συγχρόνως λέξεις και μικρές φράσεις, των δύο ή περισσότερων λέξεων, για να περιγράψουν τη θεματική της εργασίας τους. Για τον λόγο αυτόν, στο Turney (1997) αφού διαπιστώνεται η παραπάνω σύγχυση, γίνεται ο διαχωρισμός σε *λέξεις-κλειδιά*, όπου προκύπτει μια λέξη, και σε *φράσεις-κλειδιά (keyphrases)*, όπου η θεματική περιγραφή γίνεται με μικρές φράσεις. Στη συνέχεια, για τεχνικούς λόγους, διατηρείται ο όρος φράσεις-κλειδιά για θεματικές περιγραφές μέσω *και* λέξεων *και* φράσεων, ή καλύτερα, η λέξη, υπό αυτό το πρίσμα, θεωρείται ως φράση.

Αφού γίνει αυτή η διασάφηση, η *παραγωγής φράσεων-κλειδιών* μπορεί να οριστεί ως η διαδικασία στην οποία ένα κείμενο δίνεται στην είσοδο και στην έξοδο προκύπτουν φράσεις που να περιγράφουν τις θεματικές του κειμένου. Η *εξαγωγή φράσεων-κλειδιών* είναι μια υποκατηγορία της παραγωγής στην οποία οι φράσεις-κλειδιά προκύπτουν από σειρές λέξεων που εμφανίζονται στο ίδιο το κείμενο. Όταν ένα υπολογιστικό σύστημα εκτελεί τη διαδικασία τότε αυτή ονομάζεται *αυτόματη*. Έτσι, στο Turney (2002) η αυτόματη εξαγωγή φράσεων-κλειδιών ορίζεται ως "η αυτόματη επιλογή σημαντικών και θεματικών φράσεων από το σώμα ενός κειμένου".

Η αυτόματη εξαγωγή φράσεων-κλειδιών χρησιμοποιείται τόσο ως αυτάρκης πρακτική, όσο και ως βοηθητική για εφαρμογές ανάκτησης πληροφοριών (Information Retrieval) και επεξεργασίας φυσικής γλώσσας (Natural Language Processing). Για παράδειγμα, στο Hulth & Megyesi (2006) γίνεται εξαγωγή φράσεων-κλειδιών για τη βελτίωση της κατηγοριοποίησης κειμένων, ή στο Guan (2016) για την απάντηση βιο-ιατρικών ερωτήσεων με αυτόματες εξορυκτικές ή αφαιρετικές περιλήψεις (extractive/ abstractive).

Η φράση-κλειδί πρέπει να δίνει πληροφορίες για κάποια θεματική στο κείμενο (*informativeness*) και να μπορεί να μπορεί να θεωρηθεί, κατά κάποιο βαθμό, φράση (*phraseness*) (Tomokiyo & Hurst 2003). Έτσι, μια επιτυχημένη εξαγωγή φράσεων-κλειδιών πρέπει να έχει ως έξοδο της ένα σύνολο φράσεων-κλειδιών που να περιγράφουν ικανοποιητικά όλες τις θεματικές του εγγράφου και συγχρόνως να μπορούν να θεωρηθούν ως σωστά δομημένες φράσεις. Το τι αποτελεί σωστά δομημένη φράση εξαρτάται από τον χρήστη και τον σκοπό του συστήματος. (Tomokiyo & Huert, 2003). Αυτά τα δύο χαρακτηριστικά (informativeness, phraseness) μπορούν να ενσωματωθούν στο σύστημα ως κριτήρια εξαγωγής καλών φράσεων-κλειδιών και να μετρηθούν.

Η επιτυχία της αυτόματης εξαγωγής φράσεων-κλειδιών, πέρα από τα παραπάνω κριτήρια, έχει φανεί πως εξαρτάται και από τα χαρακτηριστικά του σώματος εγγράφων που δίνεται στο σύστημα, τα οποία ο σχεδιαστής του συστήματος πρέπει να λάβει υπόψιν. Στην επισκόπηση των Hasan & Ng (2014) αναδείχθηκαν τέσσερις τέτοιοι παράγοντες. α) **Μέγεθος.** Το μέγεθος των εγγράφων που εισάγονται στο σύστημα επηρεάζει τον αριθμό των υποψήφιων (candidate) φράσεων-κλειδιών, π.χ. συνήθως μεγάλα έγγραφα περιλαμβάνουν περισσότερες σχετικές θεματικές, οπότε πρέπει να εξαχθούν περισσότερες φράσεις-κλειδιά για την περιγραφή όλου του εγγράφου. β) **Δομική Συνοχή (Structural Consistency).** Όταν το έγγραφο είναι δομημένο κατά ορισμένο τρόπο, συνηθίζεται να υπάρχει κοινή λειτουργία των διαφόρων

δομικών στοιχείων του. Έτσι, σε ένα επιστημονικό άρθρο, όπου είναι κατά κύριο λόγο γνωστή η δομή της (Περίληψη, Εισαγωγή, Σχετικές Εργασίες, Κυρία Εργασία, Αποτελέσματα), οι σημαντικότερες φράσεις-κλειδιά θα βρίσκονται στη περίληψη και την εισαγωγή. Σε λιγότερο δομημένα έγγραφα ο εντοπισμός φράσεων κλειδιών δεν μπορεί να γίνει βάσει μιας εκ των προτέρων γνώσης της δομής τους, όπως π.χ. στις σελίδες στο διαδύκτιο. γ) **Θεματική Αλλαγή (Topic Change).** Κατά τη ροή του κειμένου υπάρχουν αλλαγές στη θεματική του, με φυγές από και επιστροφές σε θέματα. Η αλλαγή θεματικής μπορεί να σχετίζεται με τη προηγούμενη ή μπορεί να μην σχετίζεται. Ενώ σε κάποια έγγραφα αυτές οι αλλαγές είναι καλά κατανεμημένες ή ορισμένες εκ των προτέρων (π.χ. πρακτικά συνεδρίων), οπότε εύκολα προσδιορίσιμες, σε άλλου είδους εγγράφων η θεματικές αλλαγές είναι δομικά τυχαίες και ακολουθούν κανόνες του λόγου (discourse) (π.χ. έγγραφα συζητήσεων σε chat). δ) **Θεματικές Συσχετίσεις (Topic Correlation).** Μια σημαντική παρατήρηση είναι πως οι φράσεις-κλειδιά συνήθως σχετίζονται μεταξύ τους, και κατά συνέπεια υπάρχει και θεματική συσχέτιση στα έγγραφα. Αυτή η παρατήρηση αληθεύει κυρίως σε κείμενα τυπικών εγγράφων με καθορισμένο σκοπό, όπως επιστημονικές εργασίες, νομικά έγγραφα κ.λπ. Αντίθετα, σε άτυπα κείμενα, όπως αυτά των καθημερινών συζητήσεων, συνηθίζεται η θεματική τους να είναι αποσπασματική και η θεματική αλλαγή ραγδαία.

Βάσει των παραπάνω, θα περιγραφεί η δομή του συστήματος.

ΙΙΙ.2 Δομή του Συστήματος Αυτόματης Εξαγωγής Φράσεων-Κλειδιών.

Ένα σύστημα αυτόματης εξαγωγής φράσεων-κλειδιών, κατά κύριο λόγο, μπορεί να χωριστεί σε ορισμένα υποσυστήματα τα οποία επιτελούν συγκεκριμένες διαδικασίες και αποτελούν συστατικά στοιχεία του (components). Τα συστατικά στοιχεία αυτά είναι *α) η εξαγωγή των υποψήφιων φράσεων ή/και λέξεων από το κείμενο (candidate keyphrases) και β) η εξ αυτών επιλογή των θεματικά πιο περιγραφικών φράσεων του εγγράφου ως φράσεις-κλειδιά* (Hasan & Ng, 2014).

Η εξαγωγή λίστας υποψηφίων φράσεων και λέξεων είναι πολύ σημαντική διαδικασία εφόσον επηρεάζει άμεσα την πιστότητα (precision) και την ανάκλαση (recall) κατά την εκτίμηση (evaluation) του συστήματος. Έτσι ώστε, επιλέγοντας μια πολύ μικρή λίστα υποψηφίων παραβλέπονται κάποιες φράσεις κλειδιά, ενισχύοντας την πιστότητα του αποτελέσματος, ενώ με μια πολύ μεγάλη λίστα θα αυξηθεί η πιθανότητα λαθών, ευνοώντας την ανάκλαση (Nguyen & Luong, 2010). Ως προς την εξαγωγή της λίστας χρησιμοποιούνται ευρετικοί κανόνες για την απλοποίηση της διαδικασίες. Τέτοια ευρετικά είναι η αφαίρεση διακοπτουσών λέξεων (stopwords removal), η χρησιμοποίηση συντακτικών φίλτρων (syntactic filters), ο περιορισμός εξαγωγής μόνο ονοματικών φράσεων ή n-grams που ικανοποιούν συγκεκριμένα λεξικο-συντακτικά πρότυπα (Nguyen & Phan 2007 αναφ. Nguyen & Phan 2007 σ.183), όπως και διάφορες τεχνικές κλαδέματος (pruning) όρων που εμφανίζουν μικρή πιθανότητα να είναι φράσεις-κλειδιά (Hasan & Ng, 2014).

Στη συνέχεια, μέσω μεθόδων μάθησης με επίβλεψη ή χωρίς (supervised/unsupervised) επιλέγονται από τη λίστα υποψηφίων οι φράσεις-κλειδιά. Οι παραδοσιακές μέθοδοι με επίβλεψη θέτουν τη διαδικασία επιλογή ως ένα πρόβλημα δυαδικής κατηγοριοποίησης όπου στη μια κατηγορία θα ανήκουν οι υποψήφιες φράσεις που είναι φράσεις-κλειδιά και στην άλλη όσες δεν είναι όπως τα συστήματα KEA, GenEx κ.α.(Frank et al., 1999· Witten et al., 1999· Turney, 1999· Hulth, 2003). Στο Jiang et al. (2009) προτάθηκε η επιλογή των φράσεων-κλειδιών να ιδωθεί σαν πρόβλημα βαθμολόγισης (ranking) όπου οι διάφορες υποψήφιες φράσεις κατατάσσονται σύμφωνα με το βαθμό σχετικότητας τους και επιλέγονται οι υψηλότερα βαθμολογημένες σαν φράσεις-κλειδιά. Με αυτό τον τρόπο λύνει το βασικό μειονέκτημα της δυαδικής κατηγοριοποίησης, η οποία δεν κάνει διάκριση στο βαθμό σχετικότητας των φράσεων παρά τις επιλέγει σαν να ήταν ισάξιες. Για το κατηγοριοποίηση ως φράση-κλειδή ή τον υπολογισμό του

βαθμού σχετικότητας χρησιμοποιούνται διάφορα χαρακτηριστικά (features) είτε εξαγόμενα από το ίδιο το έγγραφο κείμενο είτε από εξωκειμενικές πηγές (Λεξικά, Θησαυροί, Μηχανές Αναζήτησης, Εγκυκλοπαίδειες κ.λπ.). Σύμφωνα και με το Hasan & Ng, (2014) τα σημαντικότερα χαρακτηριστικά που ανήκουν στο ίδιο το κείμενο είναι *στατιστικά* (tf-idf, απόσταση δύο εμφανίσεων, καθορισμός ως φράση (phraseness) σύμφωνα με τον αριθμό εμφανίσεων στα έγγραφα εκπαίδευσης (training set), το μέγεθος της φράσης κ.λπ), *δομικά* (τοποθεσία στο κείμενο, π.χ. στη περίληψη ή στην εισαγωγή), *συντακτικά* και *μορφολογικά*. Τα εξωκειμενικά, επιγραμματικά, είναι ο καθορισμός ως φράση (phraseness) και σημασιολογική συσχέτιση μεταξύ των άλλων φράσεων μέσω εγκυκλοπαιδειών και μηχανών αναζήτησης και άλλες βάσεις δεδομένων (Medelyan et al., 2009· Turney, 2003· Yih et al. 2006 αναφ. Hasan & Ng, 2014· Liu, Z. et al., 2009).

Οι μέθοδοι μάθησης χωρίς επίβλεψη χωρίζονται σε κατηγορίες που σχετίζονται με τον τρόπο μεταχείριση των υποψήφιων φράσεων:

Βάσει Θεματικών

Κάποιες μέθοδοι βρίσκουν τις θεματικές των εγγράφων ως σημασιολογικές συσταδες (clusters) ανάμεσα στις υποψήφιες φράσεις και επιλέγουν για φράση-κλειδί αυτή που έχει μεγαλύτερο βαθμό σχετικότητας από κάθε θεματική (Grineva, 2009 αναφ. Hasan & Ng, 2014· Liu, F. et al., 2009· Liu, Z. et al., 2010).

- *Βάσει Μοντέλων Γλώσσας*

Άλλες χρησιμοποιούν στατιστικά μοντέλα γλωσσών (βλ. Ponte & Croft, 1998) για να επιλέξουν φράσεις-κλειδιά που να ικανοποιούν συγκεκριμένα κριτήρια. Στο Hasan & Ng (2014) διαπιστώνεται πως ενώ στα περισσότερα συστήματα πρώτα εξάγονται οι υποψήφιες φράσεις και στη συνέχεια επιλέγονται οι φράσεις-κλειδιά, στα συστήματα όπου χρησιμοποιούνται μοντέλα γλώσσας "τα δύο βήματα συνδυάζονται". Σαν παράδειγμα, στο Tomokiyo & Hurst (2003) το σύστημα μαθαίνει τέσσερα μοντέλα γλώσσας. Ένα μοντέλο unigram και ένα ngram από ένα *σώμα κειμένων παρασκηνίου (background corpus)* LM_b^1, LM_b^N, – το οποίο αποτελεί ένα μεγάλο σώμα με γενικές γνώσεις (π.χ. το διαδίκτυο) – και αντίστοιχα άλλα δύο από ένα *σώμα κειμένων προσκηνίου (forground corpus)* LM_f^1, LM_f^N, – το οποίο είναι το σώμα που περιλαμβάνει τα έγγραφα από τα οποία θα εξαχθούν φράσεις-κλειδιά. Υπολογίζοντας ένα μέτρο απόκλισης (εδώ KL-divergence) για κάθε φράση ως απώλεια (loss) ανάμεσα στα LM_f^1 και LM_f^N προκύπτει μια τιμή σχετική με το κατά πόσο η επιλεγμένη φράση μπορεί να θεωρηθεί πραγματική φράση (*phraseness*) και κατ' ανάλογο τρόπο, από την απόκλιση ανάμεσα στα LM_b^N και LM_f^N προκύπτει μια τιμή σχετική με την πληροφορία που δίνεται από τη φράση (*informativeness*). Αυτά είναι και τα κριτήρια μέσω των οποίων βαθμολογείται κάθε φράση του κειμένου και, τελικά, επιλέγεται η λίστα με τις φράσεις-κλειδιά.

- *Βάσει Ταυτόχρονης Μάθησης*

Εξαγωγή φράσεων-κλειδιών, Αυτόματη Περίληψη

Ο Zha (2002) προτείνεται η ταυτόχρονη περίληψη και εξαγωγή φράσεων-κλειδιών υπό την υπόθεση ότι οι πιο σημαντικές προτάσεις του κειμένου περιλαμβάνουν και τις πιο σημαντικές φράσεις ή λέξεις του κειμένου (*Αρχή της Αμοιβαίας Ενίσχυσης*). Οι Wan et al. (2007) ορμώμενοι από την παραπάνω υπόθεση και τα αποτελέσματα της προτείνουν δύο παρεμφερείς υποθέσεις εμπνευσμένες από το PageRank. Πρώτον, όταν μια πρόταση είναι σημαντική τότε συνδέεται και με άλλες σημαντικές προτάσεις και, δεύτερον, όταν μια λέξη είναι σημαντική τότε συνδέεται και με άλλες σημαντικές λέξεις. Έτσι τα αποτελέσματα της περίληψης και της εξαγωγής φράσεων-κλειδιών αλληλοενισχύονται.

- *Βάσει Θεωρίας Γράφων*

Από τα παραπάνω συστήματα αυτά που περιγράφονται στα Liu, Z. et al., (2010), Zha (2002) και

Wan et al. (2007) αναπαριστούν τις υποψήφιες λέξεις ή φράσεις ως κόμβους σε δίκτυα έτσι ώστε στη συνέχεια να τις βαθμολογήσουν. Στην επόμενη ενότητα θα περιγραφούν και θα επεκταθούν περισσότερο μαζί με άλλα σχετικά συστήματα που χρησιμοποιούν αυτή τη δομή. Επιγραμματικά, μπορεί να ειπωθεί πως το κυριότερο χαρακτηριστικό που χρησιμοποιείται σε τέτοια συστήματα είναι η σημαντικότητα των κόμβων του γράφου και η συνάρτηση βαθμολόγησης PageRank και οι παραλλαγές της.

ΙΙΙ.3 Γράφοι και Αυτόματη Εξαγωγή Φράσεων-Κλειδιών

Οι παρακάτω εφαρμογές αυτόματης εξαγωγής φράσεων-κλειδιών χρησιμοποιούν, με τον ένα ή τον άλλο τρόπο, αναπαραστάσεις και μέτρα (measures) γράφων.

ΙΙΙ.3.1 (Zha 2002)

Η πρώτη εφαρμογή γράφων σε ένα σύστημα εξαγωγής φράσεων-κλειδιών έγινε έμμεσα, στη εργασία του Zha (2002), όπου αναπαριστούσε σχέσεις ανάμεσα σε προτάσεις και λέξεις και υπό την υπόθεση της Αρχής της Αμοιβαίας Ενίσχυσης κατέληγε σε βαθμολόγηση αμφότερων. Το σύστημα του προ-επεξεργάζεται κάθε έγγραφο του σώματος αφαιρώντας τις διακοπτούσες λέξεις και αποκόπτοντας τις καταλήξεις (stemming) στις υπόλοιπες και παράλληλα αναγνωρίζοντας τις προτάσεις του εγγράφου. Από τα αποτελέσματα της προ-επεξεργασίας παράγει ένα σύνολο όρων $T = \{t_1, t_2, \ldots, t_n\}$ και ένα σύνολο προτάσεων $S = \{s_1, s_2, \ldots, s_m\}$, όπου $s_i \subset T$ για κάθε έγγραφο. Στη συνέχεια κατασκευάζει έναν μη κατευθυνόμενο διμερή γράφο με βάρη στον οποίο όταν $t_i \in s_{iB}$ τότε δημιουργείται μια ακμή ανάμεσα στους κόμβους t_i και s_i. Οι ακμές έχουν βάρος w_{ij}. Από αυτή τη διαδικασία προκύπτει ένας γράφος $G_1(T, S, W)$ όπου ο W αντιστοιχεί στον πίνακα βαρών mxn και η κάθε στήλη αντιπροσωπεύει μια πρόταση. Βασιζόμενος στην υπόθεση την *Αρχή της Αμοιβαίας Ενίσχυσης* υπολογίζει τη σημαντικότητα των όρων, διάνυσμα $u(t_i)$, και των προτάσεων, διάνυσμα $v(s_j)$, βάσει των βαρών των ακμών τους ως εξής :

$$u(t_i) \propto \sum_{(s_j)\sim(t_i)} w_{ij} v(s_j), \qquad v(s_j) \propto \sum_{(t_i)\sim(s_j)} w_{ij} u(t_j) \quad (1)$$

όπου το $\propto$ συμβολίζει την αναλογία και το $(s_j) \sim (t_i)$ την ύπαρξη ακμής οι παραπάνω συναρτήσεις μπορούν να γραφτούν σε σημειογραφία πινάκων :

$$u = \frac{1}{\sigma} W v, \; v = \frac{1}{\sigma} W^T u \quad (2)$$

όπου το $1/\sigma$ είναι η σταθερά της αναλογίας η οποία υπονοείται στο *(1)*.

Στη συνέχεια, για να αποφανθεί θεματικές συστάδες του κειμένου δημιουργεί έναν μη κατευθυνόμενο γράφο με βάρη $G_2(S, WS)$ όπου οι κόμβοι αντιστοιχούν στις προτάσεις του κειμένου και οι ακμές δημιουργούνται μεταξύ δύο προτάσεων αν έχουν κοινές λέξεις. Επιπλέον, τα βάρη αντιστοιχούν στην ομοιότητα μεταξύ δύο προτάσεων ορίζεται ως το εσωτερικό γινόμενο των διανυσμάτων δύο προτάσεων. Η σχέση δύο προτάσεων μεταξύ τους ενδυναμώνεται αν αυτές έχουν κοντινή απόσταση (near-by) και προστίθεται στα βάρη τους μια μεταβλητή *α* που ονομάζει *δύναμης σύνδεσης προτάσεων (sentence link strength)* και μετά από αυτή τη διαδικασία προκύπτει ένας πίνακας βαρών *WS(α)*. Εφαρμόζει φασματική συσταδοποίηση (k-means spectral clustering) (Hartigan & Wong, 1979 αναφ. Zha, 2002) στον πίνακα WS(α) και το αποτέλεσμα, πολύ απλοποιημένα, είναι ένα σύνολο με k συστάδων

προτάσεων Π(α). Για κάθε συστάδα υπολογίζει τη σημαντικότητα των λέξεων και των ακμών *(2)* από τον γράφο G_1 και τις βαθμολογεί σύμφωνα με αυτή. Το αποτέλεσμα είναι ιεραρχικές περιλήψεις βάσει ιεραρχικης συσταδοποίησης που αντικατοπτρίζουν διαφορετικά επίπεδα διακριτότητας (granularity) αναλόγως των όρων παίρνει υπόψιν του (π.χ. παραλείποντας προτάσεις με λέξεις υψηλής σημαντικότητας και έτσι εμφανίζονται περιλήψεις λεπτότερων θεματικών (subtler topics) ως προς την σημαντικότητα τους στο έγγραφο).

Τελικά κάνει εκτίμηση της ακρίβειας συσταδοποίησης που προκύπτει σε ένα σώμα δέκα κειμένων, την οποία θεωρεί προκαταρκτική εργασία για εκτιμήσεις ως προς την εξαγωγή φράσεων-κλειδιών και την περίληψη.

III.3.2 TextRank (Mihalcea & Tarau, 2004)

Όπως φάνηκε παραπάνω, οι σχέσεις μονάδων κειμένου που παράγονται σε διμερείς γράφους είναι σχέσεις συμπερίληψης ανάμεσα σε διαφορετικά επίπεδα του κειμένου, οπότε και κατασκευάζουν ετερογενείς γράφους. Όταν οι επιλεγμένοι όροι, ως κόμβοι, ανήκουν στο ίδιο επίπεδο του κειμένου (ομοιογενής γράφος), οι σχέσεις τους, όποιες κι αν είναι αυτές, υποδηλώνουν μια σύσταση (recommendation) από τον κάθε κόμβο στον γειτονικό του (βλ. υπόθεση PageRank).

Βασιζόμενοι στο παραπάνω οι Mihalcea & Tarau (2004), καθώς αναπτύσσουν το μοντέλο TextRank για τη βαθμολόγηση (ranking) όρων κειμένου, κατασκευάζουν ένα απλό σύστημα εξαγωγής φράσεων-κλειδιών για να εκτιμήσουν το μοντέλο τους. Κατά την προεπεξεργασία, σημειώνεται στις λέξεις των κειμένων το μέρος του λόγου που τους αντιστοιχεί (POS tagging) και περνάνε από συντακτικό φίλτρο που κρατάει μόνο όσες λέξεις αναγνωρίστηκαν ως ουσιαστικά και επίθετα σαν κόμβους στον γράφο. Ως σχέση ανάμεσα στους κόμβους χρησιμοποιείται η συνεμφάνιση τους μέσα σε ένα παράθυρο W μεγέθους που διαπερνά το φιλτραρισμένο κείμενο. Σε αυτόν τον μη κατευθυνόμενο γράφο χωρίς βάρη G(V, E) θα εφαρμοστεί ο TextRank μέχρι σύγκλισης (convergence) με κατώφλι σφάλματος < 0.0001. Τα αποτελέσματα του TextRank θα χρησιμοποιηθούν για τη βαθμολόγηση των λέξεων και από αυτές θα επιλεγούν οι T (= 1/3|V|) με τη μεγαλύτερη βαθμολογία που θα αποτελέσουν τις λέξεις κλειδιά. Αμέσως μετά, κατά τη μετα-επεξεργασία, συντίθενται, από αυτές τις T φράσεις-κλειδιά όποτε βρίσκεται πως δύο ή περισσότερες από αυτές είναι γείτονες λέξεις. Έτσι το αποτέλεσμα είναι μια λίστα με λέξεις και φράσεις-κλειδιά.

Για την εκτίμηση του συστήματος χρησιμοποίηθηκε το σύνολο δοκιμής (test set) 500 περιλήψεων της συλλογής *Inspec* (Hulth, 2003). Για τα αποτελέσματα χρησιμοποίηθηκαν διαφορετικά μεγέθη παραθύρων και εντοπίστηκαν καλύτερα αποτελέσματα με παράθυρα μεγέθους W=2. Το παραπάνω σύστημα συγκρίθηκε με τα αποτελέσματα που δηλώθηκαν στο (Hulth, 2003) και βρέθηκε F-measure 3 μονάδων μεγαλύτερο (Πινακας 3.1) από τη μέθοδο που χρησιμοποιεί n-grams με POS-ετικέτες (tagged n-grams) (Hulth, 2003).

	Υπολογισμένα		Σωστά				
Σύστημα	*Συν.*	*Μέσος*	*Συν.*	*Μέσος*	*Πιστότητα*	*Ανάκλαση*	*F-Measure*
Undirected Textrank (W=2)	6.784	13,7	2.116	4,2	**31,2**	43,1	**36,2**
Hulth Ngrams-tagged	7.815	15,6	1.973	3,9	25,2	**51,7**	33,9
Συν. : Σύνολο φράσεων-κλειδιών, *Μέσος* : μέσος όρων κλειδιών ανά Περίληψη							

Πίνακας 3.1: στη Συλλογή Inspec όπως δηλώνονται στο Mihalcea & Tarau (2004)

Όπως θα φανεί αργότερα, όταν θα περιγραφεί η σύγκριση των διαφόρων μοντέλων

μεταξύ τους στο Hasan & Ng (2010), το σύστημα TextRank όπως περιγράφεται εδώ έδωσε πολύ χαμηλά αποτελέσματα.

III.3.3 SingleRank, ExpandRank, CollabRank (Wan & Xiao 2008a· 2008b)

Μια παραλλαγή του παραπάνω συστήματος αναπτύχθηκε από τους Wan & Xiao (& Yang, 2007· 2008a· 2008b) σαν βάση για τις εκτιμήσεις του συστήματος τους και ονομάστηκε SingleRank ή WordRank (για εξαγωγή unigram). Η κύριες διαφορές του SingleRank ώς προς το TextRank είναι πως α) κατασκευάζεται γράφος με βάρη όπου αντιστοιχούν στον αριθμό συνεμφάνισης δύο κόμβων στο έγγραφο, β) δεν υπάρχει ο περιορισμός να επιλεχθούν λέξεις από τις T σημαντικότερες για τη κατασκευή φράσεων και γ) το μέγεθος του παραθύρου είναι N = 10 (Hasan & Ng, 2010).

Όπως αναφέρθηκε παραπάνω, το SingleRank κατασκευάστηκε για την εκτίμηση ενός άλλου συστήματος, του ExpandRank (Wax & Xiao 2008a). Το ExpandRank διευρύνει κάθε έγγραφο υπό εξέταση d_0 με τα κείμενα των θεματικά γειτονικών του εγγράφων, και δημιουργεί γειτονιές εγγράφων. Για το προσδιορισμό της θεματικής εγγύτητας χρησιμοποιείται η συνημιτονοειδής κατά ζεύγη ομοιότητα sim_{doc} των εγγράφων, και τα k έγγραφα με τη μεγαλύτερη ομοιότητα ως προς το d_0 δημιουργούν τη γειτονιά του D με μέγεθος k+1. Όλες οι λέξεις στο διευρυμένο έγγραφο D περνάνε, όπως στο TextRank, από ένα συντακτικό φίλτρο και μένουν μόνο τα ουσιαστικά και τα επίθετα και δημιουργούνται οι κόμβοι. Οι ακμές αντιστοιχούν στις συνεμφανίσεις ανάμεσα σε δύο κόμβους, σε ένα παράθυρο μεγέθους W (2 <= W <= 20) και τα βάρη των ακμών υπολογίζονται με μια συνάρτηση συγγένειας (affinity) ως εξής :

$$aff(v_i, v_j) = \sum_{d_p \in D} \mathrm{sim}_{doc}(d_0, d_p)\, xcount_{d_p}(v_i, v_j) \quad (3)$$

όπου το $count_{dp}(v_i, v_j)$ είναι ο αριθμός συνεμφανίσεων των όρων v_i, v_j στο έγγραφο d_p του D.

Από το παραπάνω προκύπτει ο Καθολικός Γράφος Συγγένειας (Global Affinity Graph) όπου περιγράφεται από τον τετραγωνικό πίνακα M,

$$M_{i,j} = \begin{cases} aff(v_i, v_j), & \text{αν } v_i \text{ συνδέεται με το } v_j,\ i \neq j \\ 0, & \text{διαφορετικά} \end{cases} \quad (4)$$

ο οποίος κανονικοποιείται ώστε οι γραμμές του να αθροίζουν στο 1,

$$\widetilde{M}_{i,j} = \begin{cases} M_{i,j} \ / \ \sum_{j=1}^{|V|} M_{i,j}, & \text{αν } \sum_{j=1}^{|V|} M_{i,j} \neq 0 \\ 0, & \text{διαφορετικά} \end{cases} \quad (5)$$

και η βαθμολογία κάθε λέξης $S(v_i)$ υπολογίζεται με την παραλλαγή του PageRank για γράφους με βάρη

$$S(v_i) = \mu \sum_{j \in V(v_i)} S(v_j) \widetilde{M}_{j,i} + (1-\mu)/|V| \quad (6)$$

μέχρι σύγκλισης (convergence) με κατώφλι σφάλματος < 0.0001.

Μόλις υπολογιστούν τα οι βαθμολογίες για κάθε λέξη, επιλέγονται οι υποψήφιες λέξεις από το d_0 και γειτονικές υποψήφιες λέξεις συντίθενται ως υποψήφιες φράσεις με το περιορισμό να τελειώνουν με ουσιαστικό. Οι υποψήφιες λέξεις θα έχουν τη βαθμολογία που υπολογίστηκε παραπάνω και οι υποψήφιες φράσεις το άθροισμα των βαθμολογιών των λέξεων που περιέχουν. Τέλος, οι φράσεις με την υψηλότερη βαθμολογία επιλέγονται ως φράσεις-κλειδιά.

Στο Wan & Xiao (2008b) ακολουθήθηκε η ίδια διαδικασία με τη διαφορά ότι οι γειτονιές

προσδιορίζονται σύμφωνα με πρότερες συστάδες του σώματος εγγράφων, έτσι κάθε διεύρυνση του έγγραφο d_i ταυτίζεται με την συστάδα (cluster) στην οποία βρίσκεται. Χρησιμοποιούνται διάφορες τεχνικές συσταδοποίησης, όπως K-means, Agglomerative, Gold Standard κ.α. και ανακαλύπτεται ότι η Gold Standard φέρνει τα καλύτερα αποτελέσματα.

Στις εκτιμήσεις τους τα παραπάνω άρθρα χρησιμοποιούν τη συλλογή *DUC-2001* με 309 ειδησεογραφικά άρθρα με χειροκίνητη ανάθεση φράσεων-κλειδιών. Συγκρίθηκαν με το σύστημα SingleRank, που περιγράφηκε παραπάνω, και την βαθμολόγηση με tf-idf score, με τη αντίστοιχη βαθμολογία των φράσεων υπολογισμένη όπως παραπάνω. Για τα αποτελέσματα που δηλώνονται βλ. Πίνακα 3.2. Τα αποτελέσματα υποδεικνύουν ότι η πληροφορία από τα γειτονικά κείμενα συνεισφέρει στα αποτελέσματα της εξαγωγής.

Σύστημα	*Πιστότητα*	*Ανάκλαση*	*F-Measure*
TF-IDF	23,2	28,1	25,4
SingleRank	24,7	30,3	27,2
ExpandRank (k=5)	**28,8**	**35,4**	**31,7**
CollabRank	28,3	34,8	31,2

Table 3.2: στη Συλλογή DUC-2001, όπως δηλώνονται στα Wan & Xiao (2008a· 2008b)

Βασισμένο στο ExpandRank είναι το σύστημα CiteTextRank που αναπτύσσεται Gollapalli & Caragea (2014) για την εξαγωγή φράσεων-κλειδιών από επιστημονικά άρθρα. Το CiteTextRank εμπλουτίζει το βάρος των κόμβων με ένα δίκτυο βιβλιογραφικών αναφορών (Florescu & Caragea, 2017).

III.3.4 Σύγκριση των παραπάνω συστημάτων (Hasan & Ng 2010)

Στη συγκριτική μελέτη των Hasan, Ng (2010) γίνεται εκτίμηση στην αυτόματη εξαγωγή φράσεων-κλειδιών (χωρίς λέξεις-κλειδιά) με τα συστήματα Tf-IDF (στο οποία εξάγονται τα n-grams με το μεγαλύτερο άθροισμα βαθμών σύμφωνα με τις λέξεις από τις οποίες αποτελείται), TextRank, SingleRank, ExpandRank και KeyCluster (Liu, Z. et al. 2009) σε τέσσερις συλλογές, την *INSPEC (500 test set περιλήψεων)*, την *DUC-2001 (309 άρθρα)*, την *NUS (211 επιστημονικά άρθρα)* και την *τροποποιημένη ICSI (161 πρακτικά συνεδριάσεων)* και παρουσιάζουν τα παρακάτω στατιστικά στοιχεία :

	DUC-2001	*Inspec*	*NUS*	*ISCI*
# Εγγράφων	308	500	211	161
# Λέξεις / Έγγρ.	876	134	8291	1611
# Υποψήφιες Λέξεις / Εγγρ.	312	57	3271	453
# Υποψήφιες Φράσεις / Εγγρ.	207	34	2027	296
# Λέξεις / Υποψ. Φράση	1,5	1,7	1,6	1,5
# Έγκυρες φράσεις-κλειδιά	2484	4913	2327	582
# Έγκυρ. Φρ-Κλ. / Έγγρ.	8,1	9,8	11,0	3,6
# Λέξεις / Έγκυρ Φρ-Κλ.	2.1	2,3	2,1	1,3

Table 3.3: Στατιστικά Συλλογών όπως αναφέρονται στο Hasan & Ng (2010)

Γίνονται εκτιμήσεις σε πολλαπλές συλλογές με την υπόθεση πως η διαφοροποίηση στα στατιστικά κάθε συλλογής σχετίζεται με τα αποτελέσματα των εκτιμήσεων. Υπολογίζεται τα F-scores για κάθε ένα σύστημα για κάθε συλλογή αφού έχουν ρυθμιστεί όλες οι παράμετροι στο κάθε σύστημα. (παραλείπεται το KeyCluster από τα παρακάτω)

	DUC-2001		*INSPEC*		*NUS*		*ICSI*	
	Παραμ.	*F-score*	*Παραμ*	*F-score*	*Παραμ*	*F-score*	*Παραμ*	*F-score*
Tf-Idf	N=14	**27,0**	N=14	**36,3**	N=60	**6,6**	N=9	**12,1**
TextRank (W=2)	T = 100%	9,7	T =100%	33,0	T = 5%	3,2	T = 25%	2,7
SingleRank (W=10)	N = 16	25,6	N = 15	35,3	N = 190	3,8	N = 50	4,4
ExpandRank (W=10)	N = 13, κ=5	26,9	N = 15	35,3	N = 177	3,8	N = 51	4,3

Table 3.4: Όπως δηλώνεται στο Hasan & Ng (2010)

		F-score		
		Πρωτότυπα	*Hasan, Ng(2010)*	*Διαφορά*
SingleRank	*Inspec*	36,2	10,0	-26,2
SingleRank	*DUC-01*	27,1	24,9	-2,2
ExpandRank	*DUC-01*	31,7	26,4	-5,3

Table 3.5: Σύγκριση πρωτότυπων αποτελεσμάτων και των αντίστοιχων στο Hasan & Ng (2010)

Στον *πίνακα 3.4* η παράμετρος N είναι ο αριθμός φράσεων-κλειδιών που εξάγονται από το σύστημα. Η παράμετρο W το μέγεθος του παραθύρου που χρησιμοποιείται για να προσδιοριστεί η συνεμφάνιση. Η παράμετρος k στο ExpandRank τον αριθμό των γειτόνων, με την έννοια παρουσιάστηκε παραπάνω. Τα παραπάνω αποτελέσματα συγκρινόμενα με αυτά των πρωτότυπων άρθρων φαίνονται στον πίνακα 3.5.

Οι Hasan & Ng (2010) καταλήγουν πως η μεγάλη διαφορά που εμφανίζεται ανάμεσα στα F-scores των TextRank και SingleRank στα πειράματα τους έγκειται στον περιορισμό του TextRank να κατασκευάζονται φράσεις μόνο από τους T σημαντικότερους κόμβους. Όταν ο ίδιος περιορισμός εφαρμόζεται στο SingleRank τα αποτελέσματα του είναι παρόμοια με του TextRank. Ως προς τις διαφορές ανάμεσα στα πρωτότυπα αποτελέσματα και τα πειράματα τους φαίνεται πως σχετίζεται με τον τρόπο προεπεξεργασίας των κειμένων, αλλά δεν είναι δικαιολογούνται απόλυτα.

Τέλος, το σύστημα που βασίζεται στη βαθμολόγηση με Tf-Idf φαίνεται να έχει καλύτερη επίδοση από τα υπόλοιπα.

III.3.5 Topical PageRank (Liu et al., 2010)

Όμοια με τη συσταδοποίηση που επιτελεί ο Zha (2002), στο Liu et al. (2010) εκπαιδεύεται ένας θεματικός διερμηνέας με τη μέθοδο LDA (Latent Dirichlet Allocation) (Blei et al. 2003) για την αναγνώριση θεματικών λέξεων και εγγράφων. Στη συνέχεια κατασκευάζεται

μια αναπαράσταση του κειμένου με έναν κατευθυνόμενο γράφο με βάρη. Κόμβοι είναι τα ουσιαστικά και τα επίθετα, και ακμές η συνεμφάνιση τους μέσα σε παράθυρο με μέγεθος W (ανάλογο του σώματος κειμένων), παρόμοια με στο (Mihalcea & Tarau 2004). Μέσα στο παράθυρο οι ακμές κατευθύνονται από τη πρώτη λέξη προς όλες τις άλλες και τα βάρη των ακμών είναι ο αριθμός της συνεμφάνισης τους. Οι βαθμολογίες των κόμβων υπολογίζονται με τον Topical PageRank, μία παραλλαγή του PageRank για κατευθυνόμενους γράφους με βάρη που παίρνει υπόψιν του τη πιθανότητα ο κόμβος να ανήκει σε μια συγκεκριμένη θεματική z με την εξίσωση (7). Στην ίδια εξίσωση το $e(v_j, v_i)$ δηλώνει το βάρος της ακμής ij, το $O(v_j)$ τον βαθμό εξόδου του κόμβου v_j με βάρη και το $p(z|v_i)$ την πιθανότητα της θεματική z δοσμένου του v_i, και ερμηνεύεται ως το ποσοστό στο οποίο "η λέξη v_i εστιάζει στη θεματική z". Η τελευταία τιμή βαραίνει την την πιθανότητα τυχαίας φυγής από το κόμβο, η οποία στα κλασικά PageRank είναι ομοιόμορφη, οπότε κάνει τον PageRank προκατειλημμένο (biased). Έγιναν δοκιμές για την επιλογή της πιο αποτελεσματικής τιμής για το παραπάνω βάρος και η $p(z|v_i)$ επιλέχθηκε ως η καλύτερη. Άλλες ήταν η $p(v_i|z)$, "η θεματική να εστιάζει στη λέξη" και $p(v_i|z) * p(z|v_i)$ που είναι το "γινόμενο των τιμών των κεντρικών κόμβων (hub) και των αυθεντιών (authority)" (Cohn & Chang, 2000 αναφ. Liu et al. 2010· Kleinberg 1999)

$$S_z(v_i) = \lambda \sum_{j:v_j \to v_i} \frac{e(v_j, v_i)}{O(v_j)} S_z(v_j) + (1-\lambda)\, p(z|v_i), \quad (7)$$

$$S_z(p) = \sum_{v_i \in p} S_z(v_i), \quad (8)$$

$$S(p) = \sum_{z=1}^{K} S_z(p) \times p(z|d) \quad (9)$$

Οι υποψήφιες φράσεις βαθμολογούνται για κάθε θεματική z σύμφωνα με το άθροισμα των βαθμολογιών των λέξεων τους με την εξ. (8) και τέλος, η καθολική βαθμολογία υπολογίζεται με το άθροισμα των βαθμολογιών των φράσεων για κάθε θεματική επί της κατανομής της θεματικής στο έγγραφο p(z|d) στην εξ. (9). Οι τιμές των p(z|vi) και p(z|d) υπολογίζονται από τον θεματικό διερμηνέα. Οι M υψηλότερα βαθμολογημένες φράσεις εξάγονται από το σύστημα ως φράσεις κλειδιά.

Για την εκτίμηση του συστήματος χρησιμοποιούνται δύο συλλογές, η DUC-2001 όπως έχει επεξεργαστεί από τους Wan & Xiao (2008a· 2008b) με 308 ειδησεογραφικά άρθρα και η Inspec (Hulth, 2003) με 2000 περιλήψεις επιστημονικών άρθρων. Για τη συλλογή Duc-2001 επιλέγεται παράθυρο στο διάστημα [5, 20], ενώ στη συλλογή Inspec στο διάστημα [5, 10] επειδή τα έγγραφα είναι μικρότερα. Τα μέτρα εκτίμησης είναι η Πιστότητα, η Ανάκλαση, το F-measure, η δυαδική προτίμηση (Bpref) και η μέση αμοιβαία κατάταξη (MRR). Το σύστημα συγκρίνεται με το PageRank για κατευθυνόμενους γράφους με βάρη και το Tf-Idf για βάρος των λέξεων σαν βάσεις.

Σύστημα	Πιστότητα	Ανάκλαση	F-Measure	Bpref	MRR
Tf-Idf	23,9	29,5	26,4	17,9	57,6
PageRank	24,2	29,9	26,7	18,4	56,4
Topical PageRank	**28,2**	**34,8**	**31,2**	**21,4**	**63,8**

Table 3.6: Στη συλλογή DUC-2001 (Wan & Xiao 2008a· 2008b*), όπως δηλώνονται στο Liu et al. (2010)*

Σύστημα	Πιστότητα	Ανάκλαση	F-Measure	Bpref	MRR
Tf-Idf	33,3	17,3	22,7	25,5	56,5
PageRank	33,0	17,1	22,5	26,3	57,5
Topical PageRank	**35,4**	**18.3**	**24,2**	**27,4**	**58,3**

Table 3.7: Στη συλλογή INSPEC (2000 abstracts), όπως δηλώνονται στο Liu et al. (2010)

Το Topical PageRank έχει καλύτερη επίδοση από το απλό PageRank που υποδεικνύει ότι οι θεματικές του κειμένου δίνουν σημαντικές πληροφορίες για τον προσδιορισμό των φράσεων που αντιπροσωπεύουν ένα έγγραφο. Επίσης, στη συλλογή DUC-2001 έχει παρόμοια αποτελέσματα με την εργασία των Wan & Xiao (2008a) με διαφορά F-score -0.5. Ως προς τη συλλογή Inspec είναι αδύνατο να γίνουν συγκρίσεις με αποτελέσματα άλλων συστημάτων, αφού το Topical PageRank εκτιμήθηκε και με τις 2000 περιλήψεις του συστήματος, ενώ οι υπόλοιπες μόνο με το σύνολο δοκιμής 500 περιλήψεων.

III.3.2.6 TopicRank (Bougouin et al., 2013)

Ακολουθώντας την ίδια τακτική με το Liu et al. (2010) οι Bougouin et al. (2013) σχεδιάζουν το σύστημα TopicRank που εξάγει φράσεις-κλειδιά βασισμένο στις θεματικές του εκάστοτε εγγράφου. Κατά την προεπεξεργασία σημειώνουν το μέρος του λόγου των λέξεων και απαλείφουν τα χαρακτηριστικά της κλίσης αποκόπτοντας τις καταλήξεις. Για την αναγνώριση των θεματικών επιλέγουν τις μεγαλύτερες σειρές από ουσιαστικά και επίθετα, με περιορισμό το τέλος της σειράς να είναι ουσιαστικό, ως υποψήφιες φράσεις και ανάμεσά τους βρίσκουν μια σχέση ομοιότητας, η οποία βασίζεται στο ποσοστό επικάλυψης ως προς τις λέξεις τους. Ο ελάχιστος βαθμός επικάλυψης ορίζεται εμπειρικά στο 0.25 για να καθοριστεί μια σχέση. Με τον αλγόριθμό ιεραρχικής συσταδοποίησης (Hierarchical Agglomerative Clustering) προσδιορίζονται οι διάφορες συστάδες θεματικών του κειμένου, το αποτέλεσμα είναι μια λίστα με Θ θεματικές. Με τις προσδιορισμένες θεματικές του κειμένου ως κόμβους κατασκευάζεται πλήρης γράφος Θ-τάξης με βάρη G(V, E). Τα βάρη των ακμών μεταξύ δύο θεματικών είναι ο βαθμός στον οποίο οι φράσεις τους εμφανίζονται σε κοντινή απόσταση στο κείμενο και υπολογίζεται με την εξ. (10),

$$w_{i,j} = \sum_{c_i \in v_i} \sum_{c_j \in v_j} dist(c_i, c_j), \qquad (10)$$

$$dist(c_i, c_j) = \sum_{p_i \in pos(c_i)} \sum_{p_j \in pos(c_j)} \frac{1}{|p_i - p_j|} \qquad (11)$$

όπου ci είναι οι φράσεις που ανήκουν σε σε μια θεματική v_i, η εξ. (11) υπολογίζει την απόσταση ανάμεσα σε δύο φράσεις, pos είναι η απόσταση από την αρχή του κειμένου. Για την εξαγωγή φράσεων-κλειδιών, υπολογίζεται η βαθμολογία κάθε θεματική με τον PageRank για γράφους με βάρη και από τις k θεματικές με την υψηλότερη βαθμολογία εξάγονται φράσεις-κλειδιά.

Για την εκτίμηση του συστήματος χρησιμοποιούνται τέσσερις διαφορετικές συλλογές. Η Inspec (500 περιλήψεις, σύνολο δοκιμής), SemEval-2010 (100 επιστημονικά άρθρα, σύνολο δοκιμής), WikiNews-fr (100 ειδησεογραφικά άρθρα στα Γαλλικά) και Deft-2012-fr (93 επιστημονικά άρθρα στα γαλλικά, σύνολο δοκιμής). Το σύστημα συγκρίνεται με το Tf-Idf, το TextRank και το SingleRank όπως περιγράφονται παραπάνω.

Συλλογές	Έγγραφα		Φράσεις-Κλειδιά	
	Αριθμός Εγγράφων	Λέξεις / Έγγρ.	Συνολικά	Φράσεις-Κλειδιά/Εγγρ.
Inspec	500	136,3	4913	9,8
SemEval	100	5179,6	1466	14,7
WikiNews	100	309,6	964	9,6
Deft	93	6844,0	485	5,2

Table 3.8: Στατιστικά Συλλογών όπως αναφέρονται στο Bougouin et al. (2013)

Σύστημα	Inspec			SemEval-2010			WikiNews-fr			Deft-2012-fr		
	Π	Α	F	Π	Α	F	Π	Α	F	Π	Α	F
Tf-Idf	32,7	38,6	33,4	13,2	8,9	10,5	33,9	35,9	34,3	10,3	19,1	13,2
TextRank	14,2	12,5	12,7	7,9	4,5	5,6	9,3	8,3	8,6	4,9	7,1	5,7
SingleRank	**34,8**	**40,3**	**35,2**	4,6	3,2	3,7	19,4	20,7	19,7	4,5	9,0	5,9
TopicRank	27,6	31,5	27,9	**14,9**	**10,3**	**12,1***	**35,0**	**37,5**	**35,6***	**11,7**	**21,7**	**15,1***
* : υψηλή σημαντικότητα $p < 0.001$ με Student t-test												

Table 3.9: Όπως δηλώνονται στο Bougouin et al. (2013)

Όπως και στο Topical PageRank βαθμολόγηση βάσει θεματικών δίνει καλύτερα αποτελέσματα. Συγκριτικά, ως προς τα άλλα συστήματα, παρατηρείται πτώση της επίδοσης στη συλλογή Inspec που φαίνεται να σχετίζεται με το μικρό μέγεθος των περιλήψεων (κατά μέσο όρο 135 λέξεις/έγγραφο, βλ. Πίνακα 3.3). Η ίδια πτώση δεν ήταν σημαντική στο Topical PageRank, οπότε μπορεί κανείς να υποθέσει πως η μέθοδος εξαγωγής θεματικών καθορίζει την ευελιξία σε διαφορετικά μεγέθη συλλογών. Σε γενικές γραμμές, το TopicRank έχει καλύτερη απόδοση από το Tf-Idf αν εξαιρεθούν τα αποτελέσματα στο Inspec.

III.3.7 (Wan et al., 2007)

Συνδυάζοντας τις τεχνικές στο Zha (2002) και του TextRank, το Wan et al. (2007) κατασκευάζει έναν μη κατευθυνόμενο γράφο με σχέσεις πρότασης – πρότασης, λέξης – λέξης, και πρότασης – λέξης σαν αναπαράσταση των κειμένων στο σύστημα του. Το σύστημα επιτελεί ταυτόχρονη εργασία εξαγωγής φράσεων-κλειδιών και περίληψης και όπως σύστημα στο Zha (2002) (MutualRank στο Wan et al. 2007), βασίζεται στην Αρχή της Αμοιβαίας Ενίσχυσης.

Οι σχέσεις ανάμεσα σε προτάσεις υπολογίζονται με τη συνημιτονοειδή ομοιότητα τους ως προς τις κοινές τους λέξεις. Αν δύο προτάσεις εμφανίσουν ομοιότητα μεγαλύτερη του μηδενός σχηματίζεται μια ακμή ανάμεσα τους. Ο τετραγωνικός πίνακας γειτνίασης U που προκύπτει κανονικοποιείται σε U^{norm} ώστε οι γραμμές του να αθροίζουν στο 1, όπως στη συνάρτηση (5). Παρόμοια, σύμφωνα με ένα μέτρο σημασιολογικής ομοιότητας (είτε βασισμένο σε εξωκειμενική γνώση π.χ. WordNet (WN), είτε βασισμένο σε στατιστικές του σώματος κειμένων, όπως Mutual Information (MI) με παράθυρο (βλ. Mihalcea et al., 2006) προστίθενται ακμές ανάμεσα στις λέξεις και ο τετραγωνικός πίνακας γειτνίασης V που προκύπτει

κανονικοποιείται σε V^{norm}. Τέλος, οι σχέσεις λέξεων-προτάσεων είναι σχέσεις συμπερίληψης των λέξεων στις προτάσεις και τα βάρη τους υπολογίζονται μέσω μιας συνάρτησης συγγένειας (affinity) ως εξής :

$$aff(s_i, t_j) = \frac{tf_{t_j} \cdot isf_{t_j}}{\sum_{t \in s_i} tf_t \cdot isf_t} \quad (12)$$

όπου isf_t είναι η αντίστροφη συχνότητα στη πρόταση του όρου t_j κατ' αναλογία με το idf. Από αυτή τη διαδικασία προκύπτει ένας πίνακας mxn W και ο ανάστροφός του W^T κανονικοποιούνται ώστε το άθροισμα των γραμμών τους να είναι ίσο με 1. (W^{norm}, W^{Tnorm})

Βάσει της Αρχής της Αμοιβαίας Ενίσχυσης οι βαθμολογίες των λέξεων και των προτάσεων δίνονται από τις παρακάτω εξισώσεις,

$$u(s_i) = \alpha \sum_{j=1}^{m} U_{j,i}^{norm} u(s_j) + \beta \sum_{j=1}^{n} W_{j,i}^{Tnorm} v(t_j) \quad (13)$$

$$v(t_j) = \alpha \sum_{j=1}^{n} V_{i,j}^{norm} v(t_j) + \beta \sum_{j=1}^{m} W_{i,j}^{norm} u(s_j) \quad (14)$$

Σύμφωνα με τις βαθμολογίες που έχει λάβει κάθε λέξη γίνεται εκτίμηση με 34 έγγραφα, από τις πρώτες πέντε συστάδες εγγράφων της συλλογή DUC-2002, από τα οποία εξάγονται χειροκίνητα έως και δέκα λέξεις κλειδιά για το καθένα, με μέσο όρο 6.8 / έγγραφο. Σύγκριναν τη μέθοδο τους με τα Zha (2002) και Mihalcea & Tarau (2004) στην εξαγωγή λέξεων-κλειδιών (unigrams) (MutualRank, WordRank αντίστοιχα). Κάθε μέθοδος έχει ρυθμιστεί ώστε να εξάγει 10 λέξεις-κλειδιά.

Σύστημα	*Πιστότητα*	*Ανάκληση*	*F-measure*
Wan et al. (2007) (MI) W = 5	**42,5**	49,1	**45,6**
Wan et al. (2007) (WN)	41,3	**50,4**	45,4
WordRank W = 10	37,9	40,7	39,3
WordRank W = 5	36,8	42,2	39,3
WordRank W = 2	37,2	41,2	39,2
MutualRank	35,5	39,7	37,5
MI: υπολογισμός ακμών με Mutual Information, WN: μέσω του WordNet			

Table 3.10: στη Συλλογή DUC-2002 όπως δηλόντονται στο Wan et al. (2007)

Τα παραπάνω αποτελέσματα δεν μπορούν να συγκριθούν με τα παραπάνω αφού η συλλογή από την οποία εκτιμήθηκαν είναι διαφορετική. Μια σημαντική παρατήρηση που έγινε στο Wan & Xiao (2008a) για το ExpandRank είναι ότι το μέγεθος παραθύρου που επιλέγεται δεν επηρεάζει αισθητά την επίδοση στο διάστημα [4, 20], ενώ όταν το επιλεγμένο μέγεθος είναι W = 2 η επίδοση πέφτει περίπου 2 μονάδες κάτω στο F-measure, συγκριτικά . Εδώ φαίνεται πως το παράθυρο δεν αλλάζει αισθητά την επίδοση του συστήματος ως προς το F-measure και αυτό ίσως έχει σχέση με την φύση του συστήματος όπου εξάγει unigrams, οπότε δεν χρειάζεται να συνθέσει φράσεις μετά το πέρας της επεξεργασίας (post-processing). Πέρα από αυτά, το

σύστημα τους φαίνεται να φέρνει τα καλύτερα αποτελέσματα από τα άλλα δύο ανεξάρτητα της μεθόδου σημασιολογικής ομοιότητας μεταξύ των λέξεων (MI - WN : Π +1,2% Α -1,3% F +0,2%).

III.3.2.8 PositionRank (Florescu & Caragea, 2017)

Τα καλύτερα αποτελέσματα, έως το 2018, από συστήματα βασισμένα σε δομές γράφων φαίνεται να έρχονται από το σύστημα PositionRank των Florescu & Caragea, (2017). Στο εν λόγω σύστημα κατασκευάζεται ένας γράφος με βάρη κατά το πρότυπο του SingleRank όπου κόμβοι είναι τα ουσιαστικά και τα επίθετα του κειμένου, ακμές η συνεμφάνιση τους σε ένα παράθυρο μεγέθους W και το βάρος τον ακμών ο αριθμός συνεμφανίσεων μέσα στο κείμενο. Ο πίνακας γειτνίασης με βάρη κανονικοποιείται όμοια με την εξ. (5). Η βαθμολογία του κάθε κόμβου υπολογίζεται με έναν προκατειλημμένο PageRank όπως στην εξ. (7) με βάρος προκατάληψης το |V| διαστάσεων *διάνυσμα* p^{δ}, ίσο με,

$$p^{\delta}=[\frac{p_1}{p_1+p_2+...+p_{|(V)|}},\ \frac{p_2}{p_1+p_2+...+p_{|(V)|}},\ ...,\ \frac{p_{|(V)|}}{p_1+p_2+...+p_{|(V)|}}] \quad (15)$$

όπου για κάθε λέξη i το p_i είναι το άθροισμα,

$$p_i=\sum_{a\in pos(i)}\frac{1}{a} \quad (16)$$

με pos(i) οι θέσεις στις οποίες εμφανίζεται η λέξη v_i στο κείμενο.

Μία παρόμοια μέθοδος είναι να υπολογίζεται το βάρος προκατάληψης μόνο με την πρώτη εμφάνιση της λέξης στο κείμενο και οπότε,

$$fp_i=\frac{1}{pos(i)[1]} \quad (17)$$

Οι υποψήφιες προτάσεις συντίθενται από τις λέξεις με τις μεγαλύτερες βαθμολογίες,που είναι γειτονικές, όπως στο SingleRank και η βαθμολογία των φράσεων είναι το άθροισμα της βαθμολογίας των λέξεων τους όπως στο ExpandRank (Wan & Xiao, 2008a). Οι k φράσεις με την υψηλότερη βαθμολογία επιλέγονται σαν φράσεις-κλειδιά.

Η εκτίμηση του PositionRank έγινε με τρεις διαφορετικές συλλογές. Οι πρώτες δύο είναι οι, συνταγμένες από τους Gollapalli & Caragea, (2014), συλλογές από το CiteSeerX επιστημονικών άρθρων του συνεδρίου της ACM "Knowledge Discovery and Data Mining" (KDD) και του World Wide Web Conference (WWW). Η τρίτη συντάχθηκε από τους Nguyen & Kan (2007) από διάφορα επιστημονικά άρθρα. Εξάγονται λέξεις κλειδιά από τους τίτλους και τις περιλήψεις των επιστημονικών άρθρων των παραπάνω συλλογών. (Πίνακας 3.11)

Συλλογές	#Εγγράφα	#Φράσεις-Κλειδιά	#Φράσεις-Κλειδιά/Εγγρ.
KDD	834	3093	3,70
WWW	1350	6405	4,74
Nguyen & Kan 2007	211	882	4,18

Table 3.11: Στατιστικά Συλλογών όπως αναφέρονται στο Florescu & Caragea, (2017)

Το σύστημα συγκρίνεται με τα περισσότερα από τα παραπάνω περιγραφόμενα συστήματα και τα αποτελέσματα προβάλλονται στους πίνακες [12], [13], [14] για κάθε συλλογή.

Από τα αποτελέσματα των Florescu & Caragea, (2017) φαίνεται πως το σύστημα PositionRank έχει τις καλύτερες επιδόσεις συγκριτικά με τα άλλα συστήματα που περιγράφηκαν σε αυτή την ενότητα. Το χαμηλό F-score δικαιολογείται από την επιλογή των τίτλων και των περιλήψεων των άρθρων για εξαγωγή φράσεων-κλειδιών, αφού είναι δύσκολο να περιέχονται όλες οι σχετικές φράσεις-κλειδιά εκεί.

Σύστημα	k=2			k=4			k=6			k=8		
	Π	Α	F	Π	Α	F	Π	Α	F	Π	Α	F
PositionRank	11,1	5,6	**7,3**	10,8	11,1	**10,6**	9,8	15,3	**11,6**	9,2	18,9	**12,1**
PositionRank-fp	10,3	5,3	6,8	10,2	10,4	10,0	9,1	13,8	10,9	8,6	17,2	11,3
Tf-Idf	10,5	5,2	6,8	9,6	9,7	9,4	9,2	13,8	10,7	8,7	17,4	11,3
TextRank	8,1	4,0	5,3	8,3	8,5	8,1	8,1	12,3	9,4	7,6	15,3	9,8
SingleRank	9,1	4,6	6,0	9,3	9,4	9,0	8,7	13,1	10,1	8,1	16,4	10,6
ExpandRank	10,3	5,5	6,9	10,4	10,7	10,1	9,2	14,5	10,9	8,4	17,5	11,0
Topical PageRank	9,3	4,8	6,2	9,1	9,3	8,9	8,8	13,4	10,3	8,0	16,2	10,4

Table 3.12: στη Συλλογή KDD όπως δηλόντονται στο Wan et al.

Σύστημα	k=2			k=4			k=6			k=8		
	Π	Α	F	Π	Α	F	Π	Α	F	Π	Α	F
PositionRank	11,3	5,3	**7,0**	11,3	10,5	**10,5**	10,8	14,9	**12,1**	9,9	18,1	**12,3**
PositionRank-fp	9,6	4,5	6,0	10,3	9,6	9,6	10,1	13,7	11,2	9,4	17,2	11,7
Tf-Idf	9,5	4,5	5,9	10,0	9,3	9,3	9,6	13,3	10,7	9,1	16,8	11,4
TextRank	7,7	3,7	4,8	8,6	7,9	8,0	8,1	12,3	9,8	8,2	15,2	10,2
SingleRank	9,1	4,2	5,6	9,6	8,9	8,9	9,3	13,0	10,5	8,8	16,3	11,0
ExpandRank	10,4	5,3	6,7	10,4	10,6	10,1	9,5	14,7	11,2	8,6	17,7	11,2
Topical PageRank	8,8	4,2	5,5	9,6	8,9	8,9	9,5	13,2	10,7	9,0	16,5	11,2

Table 3.13: στη Συλλογή WWW όπως δηλόντονται στο Wan et al.

Σύστημα	k=2			k=4			k=6			k=8		
	Π	Α	F	Π	Α	F	Π	Α	F	Π	Α	F
PositionRank	10,5	5,8	**7,3**	10,6	11,4	**10,7**	11,0	17,2	**13,0**	10,2	21,1	**13,5**
PositionRank-fp	10,0	5,4	6,8	10,4	11,1	10,5	11,2	17,4	13,2	10,1	21,2	13,3
Tf-Idf	7,3	4,0	5,0	9,5	10,3	9,6	9,1	14,4	10,9	8,9	18,9	11,8
TextRank	6,3	3,6	4,5	7,4	7,4	7,2	7,8	11,9	9,1	7,2	14,8	9,4
SingleRank	9,0	5,2	6,4	9,5	9,9	9,4	9,2	14,5	11,0	8,9	18,3	11,6
ExpandRank	9,5	5,3	6,6	9,5	10,2	9,5	9,1	14,4	10,8	8,7	18,3	11,4
Topical PageRank	8,7	4,9	6,1	9,1	9,5	9,0	8,8	13,8	10,5	8,8	18,0	11,5

Table 3.14: στη Συλλογή Nguyen & Kan 2007 όπως δηλόντονται στο Wan et al.

ΚΕΦΑΛΑΙΟ IV. Αυτόματη Περίληψη

IV.1 Εισαγωγή

Η περίληψη μπορεί να οριστεί ως εξής, είναι η διαδικασία κατά την οποία ένα κείμενο παράγεται βασισμένο σε ένα ή περισσότερα άλλα κείμενα και διατηρεί και να μεταδίδει τις πιο σημαντικές πληροφορίες τους, ενώ ταυτόχρονα το μέγεθος του παραμένει μικρό. Στην αυτόματη περίληψη, από υπολογιστικά συστήματα, ο σκοπός είναι να παραχθεί μια περίληψη που να προσεγγίζει τις περιλήψεις που γίνονται από επαγγελματίες. (Hovy & Lin, 1998)

Υπάρχουν δύο βασικοί τρόπο να παραχθεί μια περίληψη. Ο αφαιρετικός (abstractive) τρόπος περίληψης δεν περιορίζεται στις λέξεις, φράσεις, προτάσεις ή παραγράφους που εμφανίζονται στο ίδιο το κείμενο, άλλα επιτελεί ταυτόχρονα και παραγωγή φυσικής γλώσσας. χωρίζεται σε περιορισμένη αφαιρετική μετάφραση, όταν έχει ένα περιορισμένο λεξιλόγιο με το οποίο μπορεί να "εκφραστεί" και μη περιορισμένη και μη περιορισμένη αφαιρετική μετάφραση, όταν έχει όλο το εύρος λεξιλογίου της γλώσσας του αρχικού κειμένου. Αντίθετα ο εξορυκτικός (extractive) τρόπος περίληψης αναγνωρίζει τα πιο σημαντικά κομμάτια ενός κειμένου και τα συνθέτει σε ένα νέο κείμενο, την περίληψη (Radev et al., 2002).

Ακόμα, αν η περίληψη γίνεται από ένα μόνο έγγραφο τότε ονομάζεται *περίληψη μονού εγγράφου* και αν γίνεται από πολλά έγγραφα, περίληψη πολλαπλών εγγράφων. Στη πρώτη περίπτωση συνηθίζεται να λέγεται απλά περίληψη και μόνο αν η πηγή της είναι πολλαπλά έγγραφα υπάρχει προσδιορισμός.

Άλλο κριτήριο για το χαρακτηρισμό μιας περίληψης είναι το κατά πόσο παράγεται από τις πληροφορίες όλου του εγγράφου, όπου και ονομάζεται Γενική Περίληψη (Generic Summarization), ή βάσει ενός ερωτήματος (Query-based Summarization ή Personalized Summarization) που προσδιορίζει τη πληροφορία ως προς την οποία θα παραχθεί. (Gambhir & Gupta, 2017)

Εν παρόδω, αξίζει να σημειωθεί ότι η εκτίμηση ενός συστήματος περίληψης είναι πολύ δύσκολη, αφού υπάρχει ασυμφωνία στο ποια στοιχεία της πηγής αποτελούν άξια να βρεθούν στη περίληψη. Ενδεικτικά, έχει παρατηρηθεί πως δύο διαφορετικοί επαγγελματίες θα επιλέξουν διαφορετικά περιεχόμενα για τη περίληψη τους. Έτσι δεν μπορεί να βρεθεί σταθερή βάση για σωστή εκτίμηση με κάποια πρότυπη περίληψη. Πολλές μελέτες έχουν διεξαχθεί πάνω σε αυτό το θέμα. (Radev et al., 2003 · Teufel & Van Halteren, 2004). Τα συνέδρια SUMMAC (1998), DUC (2001-2007) και TAC (2008-2011) διεξάγουν διαγωνισμούς και εκτιμήσεις για τα διάφορα συστήματα προσφέροντας συλλογές για δοκιμές.

Τις βάσεις για ένα σύστημα αυτόματης περίληψης τις έθεσε ο Luhn (1958). Σχεδόν όλα τα μετέπειτα συστήματα βασίζονται στις γενικές ιδέες που υλοποίησε το σύστημα που ανέπτυξε στην IBM για την εξορυκτικής περίληψη τεχνικών κειμένων και που περιγράφεται στο παραπάνω άρθρο. Αρχικά, τα κείμενα περνάνε από μια *διαδικασία προεπεξεργασίας* ώστε να αποκοπούν οι καταλήξεις τους (stemming) και διαγράφονται οι διακόπτουσες λέξεις (stopwords removal). Στη συνέχεια "ο παράγοντας σημαντικότητας μιας πρότασης βρίσκεται από την ανάλυση των λέξεων της." (Luhn, 1958 σ.160) Όσο πιο σημαντικές είναι οι λέξεις που περιέχονται στη πρόταση τόσο πιο σημαντική είναι η ίδια η πρόταση. Για να προσδιοριστεί η σημαντικότητα μιας λέξης λαμβάνεται υπόψιν η συχνότητα με την οποία εμφανίζεται μέσα στο κείμενο και οι σχετικές θέσεις στις οποίες εμφανίζεται μέσα στη πρόταση. Οι προτάσεις με το μεγαλύτερο βαθμό σημαντικότητας επιλέγονται για την παραγωγή της περίληψης. Αυτό είναι επίσης ένα από τα πρώτα συστήματα που προσεγγίζουν το θέμα στατιστικά.

Την ίδια χρονιά, και επίσης στην IBM ο Baxendale (1958) δείχνει πως η τοποθεσία της κάθε πρότασης μέσα στις παραγράφους δίνει πληροφορίες για την σημαντικότητα της. Στα πειράματα που έκανε σε 200 παραγράφους κατέληξε πως η θεματική πρόταση εμφανιζόταν κατά

85% στην αρχή της παραγράφου και κατά 7% στο τέλος της. Ενστάσεις στα παραπάνω αποτελέσματα υπέβαλλαν οι Braddock (1974) και Singer & Donlan (1985) όπου υποστηρίχθηκε πως μόνο στο 13% των παραγράφων εμφανίζεται η θεματική πρόταση στην αρχή της και, αντίστοιχα, ότι η θεματική μπορεί να εμφανιστεί οπουδήποτε μέσα στη παράγραφο ή και καθόλου. Έτσι, μερικά εργασίες επέλεξαν τεχνικές που βασίζονται στην εξαγωγής σημαντικών παραγράφων αντί προτάσεων. (Salton et al., 1994· Salton et al., 1996· Mitra et al., 1997)

Πολύ σημαντική για τα μετέπειτα συστήματα ήταν η εργασία του Edmundson (1969) για περίληψη επιστημονικών εργασιών. Εκεί, χρησιμοποίησε έναν γραμμικό συνδυασμό διαφόρων ευρετικά χαρακτηριστικών για τον υπολογισμό του βάρους της κάθε πρότασης. Τα καινούρια χαρακτηριστικά που παρουσίασε ήταν α) ο αριθμός των λέξεων της πρότασης που εμφανίζονται στο τίτλο και την περίληψη (abstract), β) ο αριθμός των λέξεων στη πρόταση που υπάρχει σε μια προκατασκευασμένη λίστα λέξεων, τη λίστα συνθηματικών λέξεων, που θεωρούνται ενδεικτικές μιας πλούσιας σε πληροφορία πρότασης,

Στη βάση αυτών των εργασιών, πιο σύγχρονα συστήματα διατήρησαν και ανέπτυξαν αυτές τις υποθέσεις : α) η σημαντικότητα μιας πρότασης από τις λέξεις που περιέχει, β) η σημαντικότητα μιας πρότασης από τη θέση της στο κείμενο, γ) η σημαντικότητα μιας πρότασης από τις θέσεις των λέξεων της στο κείμενο.

Έτσι, οι Kupiec et al. (1995) και οι Aone et al. (1999) προσέγγισαν την περίληψη ως ένα πρόβλημα κατηγοριοποίησης όπου οι προτάσεις χωρίζονταν σε άξιες εξόρυξης και ανάξιες, βάσει διαφόρων χαρακτηριστικών όπως το μέγεθος της πρότασης, λέξεις που εμφανίζονται με κεφαλαία και άλλα χαρακτηριστικά που πρότεινε ο Edmundson (1969). (Das & Martins, 2007). Συνεχίζοντας, οι Conroy & O'leary (2001) χρησιμοποίησαν Κρυφές Μαρκοβιανές Αλυσίδες (Hidden Markov Chains βλ Rabiner (1989)) όπου βασίζεται στα παραπάνω χαρακτηριστικό και ένα, επιπλέον, πρωτότυπο όπου υπολογίζει τη πιθανότητα των όρων της πρότασης δοσμένου των όρων του κειμένου. Το σύστημα των Conroy & O'leary εμφανίζει καλύτερες επιδόσεις από τα προηγούμενα. (Nenkova & McKeown, 2011)

Για διάφορα ευρετικά χαρακτηριστικά που μπορούν να χρησιμοποιηθούν βλέπε την εργασία των Fattah & Ren (2009, σσ 128-130).

Όλες οι παραπάνω τεχνικές θεωρούσαν τις προτάσεις ανεξάρτητες τη μία από την άλλη. Αντίθετα, οι Carbonell & Goldstein (1998) προτείνουν, για τη περίληψη πολλαπλών εγγράφων, τη χρησιμοποίηση της Μέγιστης Περιθωριακής Συνάφειας (Maximal Marginal Relevance, MMR), η οποία υπολογίζει την ανομοιότητα μεταξύ προτάσεων ώστε να απαλειφθεί ο πλεονασμός από προτάσεις που θα εμφάνιζαν υψηλό βαθμό αν και περιέχουν κοινές πληροφορίες. Έτσι ο βαθμός κάθε πρότασης υπολογίζεται σύμφωνα με την ομοιότητα της με της άλλες. Η MMR είναι χρήσιμη και στις βασισμένες σε ερώτημα περιλήψεις για να ανακαλύψει προτάσεις με καινούριες πληροφορίες που βασίζονται στο ερώτημα. Προς αυτή τη κατεύθυνση, το σύστημα MEAD για περίληψη πολλαπλών εγγράφων, όπως αναπτύσσεται στο Radev et al. (2000), συσταδοποιεί τα έγγραφα της συλλογής σε θεματικές και κάθε θεματική χαρακτηρίζεται από ένα ψευτο-έγγραφο σε μορφή διανύσματος, το centroid, στο οποίο εγγράφονται όλες οι λέξεις που έχουν συχνότητα μεγαλύτερη από ένα προκαθορισμένο κατώφλι. Κάθε πρόταση βαθμολογείται, έπειτα, βάσει των centroids και στη συνέχεια επανα-βαθμολογείται για την απαλοιφή πλεονασμός, όπως παραπάνω.

Στην παρατήρηση της μη ανεξαρτησίας των προτάσεων βασίζονται και οι τεχνικές με γράφους. Στα συστήματα που αναπτύσσονται από αυτές τις τεχνικές, πολλά από τα παραπάνω ευρετικά χαρακτηριστικά παραλείπονται και όπως φάνηκε στα συστήματα εξαγωγής φράσεων-κλειδιών με διαβάθμιση PageRank, μπορούν να ενσωματώθούν με τη μορφή προκατάληψης στον αλγορίθμου διαβάθμισης. (Wan, 2010)

IV.2 Γράφοι και Αυτόματη Περίληψη

Οι γράφοι πριν εφαρμοστούν στην αναπαράσταση σχέσεων προτάσεων είχαν χρησιμοποιηθεί για να περιγράψουν σχέσεις ανάμεσα σε κείμενα (π.χ. παραγωγή σχέσεων υπερσυνδέσμου ανάμεσα σε έγγραφα). Αυτό ήταν και το έναυσμα για τη μελέτη της σχέσεως ομοιότητας ανάμεσα σε προτάσεις για την διαπίστωση της πιο σημαντικής πρότασης στα κείμενα με την υπόθεση πως προτάσεις που συνδέονται νοηματικά με πολλές άλλες προτάσεις του κειμένου ή των κειμένων ενδέχεται να εμπεριέχουν σημαντικές πληροφορίες και άρα είναι εύλογο να συμπεριληφθούν σε μια εξορυκτική περίληψη κειμένων.

IV.2.1 (Skorokhod'ko, 1972)

Ο πρώτος που φαίνεται να διατύπωσε αυτή την υπόθεση ήταν ο Skorokhod'ko (1972). Στο Mani (2001) περιγράφεται το μοντέλο το οποίο χρησιμοποιεί η παραπάνω εργασία για τον προσδιορισμό της σχετικότητας των προτάσεων.

Οι κόμβοι είναι προτάσεις και οι ακμές μεταξύ τους σχέσεις σημασιολογική ομοιότητας. Η σημασιολογική ομοιότητα διαπιστώνεται όταν οι προτάσεις περιέχουν είτε κοινές λέξεις, είτε λέξεις με σχέσεις υπωνυμία ή υπερνυμίας μεταξύ τους ή είναι, έκαστες, θεματικά ισχυρές στο κείμενο (βάσει των συχνοτήτων τους στα κείμενα και τις συλλογές). Ο βαθμός σημασιολογικής συσχέτισης ανάμεσα στις προτάσεις είναι βασισμένος στον αριθμό αυτών των σχέσεων των λέξεων τους. Δύο είναι τα κριτήρια για την σημαντικότητα των προτάσεων, το κριτήριο Συνδεσιμότητας (Connectivity) και το κριτήριο Αναγκαιότητας (Indispensability), όπως τα ονόμασε ο Mani (2001). Το κριτήριο Συνδεσιμότητας λέει πως η σημαντικότητα μιας πρότασης είναι ανάλογη με τον αριθμό των προτάσεων που συνδέονται μαζί της με σημασιολογική ομοιότητα. Συγχρόνως, το κριτήριο Αναγκαιότητας λέει πως η σημαντικότητα μιας πρότασης είναι ανάλογη με το βαθμό αλλαγής που θα προκληθεί με την αφαίρεση της. Αυτά τα δύο μπορούν να συνδυαστούν όπως στην εξίσωση (1) όπου Fi είναι η σημαντικότητα της πρότασης i, N το μέτρο το σχετικό με την συνδεσιμότητα, το M ο αριθμός κόμβων στον γράφο ή προτάσεων στο κείμενο και το Mi ο αριθμός κόμβων στην μέγιστη συνδεδεμένη συνιστώσα (connected component) του γράφου μετά την αφαίρεση του κόμβου i.

$$F_i = N_i(M - M_i) \quad (1)$$

Αυτή η μέθοδος δεν επηρέασε ιδιαίτερα το πεδίο της αυτόματης περίληψης και αναφέρεται σπάνια (Jones, 1999· Mani, 2001· Moens, 2002). Περιλαμβάνει όμως τις βασικές ιδέες των μελλοντικών μοντέλων.

IV.2.2 Τρείς αλγόριθμοι περίληψης (Salton et al., 1997)

Όπως αναφέρθηκε παραπάνω οι Salton et al. (1994) εξάγουν μεγαλύτερες μονάδες κειμένου από την πρόταση. Η ίδια πρακτική συνεχίζεται και στην εργασία Salton et al. (1997) όπου συνδέονται οι παράγραφοι σύμφωνα με την ομοιότητα τους. Οι ακμές εγγράφονται αν η ομοιότητα ξεπεράσει ένα ορισμένο κατώφλι. Η ομοιότητα ανάμεσα σε δύο παραγράφους υπολογίζεται με το εσωτερικό γινόμενο των διανυσμάτων τους και κανονικοποιείται ώστε να πέφτει στο διάστημα [0, 1].

Αφού κατασκευασθεί ο γράφος, η περίληψη κατεσκευάζεται με τρεις διαφορετικούς αλγορίθμους. Ο πρώτος ονομάζεται *θαμνώδης διαδρομή (bushy path)*, σε αυτόν επιλέγονται οι n παράγραφοι με το μεγαλύτερο βαθμό στον γράφο και η περίληψη αποτελείται από αυτές τις

παραγράφους με τη σειρά που εμφανίζονται στο κείμενο. Αυτός ο αλγόριθμος όμως δεν συγκρατεί, αναγκαστικά, την συνεκτικότητα (coherence) του κειμένου αφού οι επιλεγμένες παράγραφοι μπορεί να μην έχουν ιδιαίτερη ροή. Για να λυθεί αυτό το θέμα προτείνεται το *μονοπάτι κατά βάθος (depth-first path)*, ξεκινώντας από μια παράγραφο με υψηλό βαθμό εισάγεται στη περίληψη η πιο σχετική παράγραφος με την επιλεγμένη, στη συνέχεια εισάγεται η πιο σχετική παράγραφος με τη δεύτερη και ούτω καθεξής για n παραγράφους. Αυτός ο αλγόριθμος αν δημιουργεί συνεκτικές περιλήψεις μπορεί να χάσει διάφορες θεματικές του κειμένου και εξαρτάται απόλυτα από την αρχική επιλογή. Ένας συνδυασμός των δύο παραπάνω, ο αλγόριθμος *κατακερματισμένης θαμνώδους διαδρομής (segmented bushy path)* θεωρητικά λύνει το πρόβλημα. Δημιουργείται ένα κομμάτι της περίληψης με τον αλγόριθμο μονοπατιού κατά βάθος, αν από αυτό φαίνεται να περισσεύουν παράγραφοι με υψηλό βαθμό τότε επιλέγεται ο υψηλότερος από αυτούς και με τον αλγόριθμο μονοπατιού κατα βάθος δημιουργείται ένα δεύτερο κομμάτι περίληψης κ.ο.κ. ώσπου να έχουν μπει στη περίληψη όλες οι θεματικές του κειμένου.

Η εκτίμηση των τριών αυτών τεχνικών έγινε σε μια συλλογή 50 εγγράφων. Μετρήθηκε η ανάκλαση ανάμεσα στις παραγράφους που εξάγονται από τους αλγορίθμων και σε δύο χειροποίητες περιλήψεις. Στη στήλη optimistic είναι τα αποτελέσματα σύμφωνα με την περίληψη που είναι πιο κοντά στις αυτόματες, στην pessimistic με τη περίληψη που είναι πιο μακριά από τις αυτόματες, intersection σύμφωνα με τις παραγράφους που θεωρήθηκαν σημαντικές και από τις δύο περιλήψεις και union μια αυτόματη εξαγόμενη παράγραφος θεωρείται ορθή αν βρέθηκε τουλάχιστον σε μία από τις δύο. Ο αλγόριθμος Random επιλέγει τυχαίες παραγράφους από το κείμενο.

Αλγόριθμοι	Η επικάλυψη ανάμεσα στις δύο περιλήψεις είναι 46%				
	Optimistic	Pessimistic	Intersection	Union	Avg btwn opt & pes
Bushy Path	45,60	30,74	47,33	55,16	**38,17**
Depth-First Path	43,98	27,76	42,33	52,48	35,84
Segmented Bushy	45,48	26,37	38,17	52,95	35,92
Random	39,16	22,07	38,47	44,24	30,62

Πίνακας 4.1 όπως δηλόντονται στο Salton et al. 1997

Τα αποτελέσματα δείχνουν πως ο αλγόριθμος θαμνώδης διαδρομής φέρνει τα καλύτερα αποτελέσματα, 55% των παραγράφων που επιλέχθηκαν συμφωνούν με τις χειροποίητες περιλήψεις των δύο ειδικών. Σύμφωνα με τα οπτιμιστικά αποτελέσματα η περίληψη bushy path συμφωνεί κατά 46% με την εν λόγω περίληψη, το οποίο είναι και το ποσοστό συμφωνίας ανάμεσα στις δύο χειροποίητες περιλήψεις. Κατά κύριο λόγο, οι παράγραφοι που επιλέχθηκαν θεωρούνται σχετικές με το κείμενο κατά 38%.

IV.2.3 (Zha, 2002· Wan et al., 2007)

Ο τρόπος δημιουργίας περιλήψεων που αναπτύσσεται στο Zha (2002) περιγράφεται στην αυτόματη εξαγωγή φράσεων-κλειδιών. Όπως αναφέρθηκε, βασει στην αρχή της αμοιβαίας ενίσχυσης προτάσεις που περιέχουν λέξεις-κλειδιά ενισχύονται από το σύστημα. Την ίδια ιδέα ανέπτυξε η εργασία των Wan et al. (2007) που επίσης περιγράφηκε παραπάνω. Εδώ θα δηλωθούν τα στοιχεία εκτίμησης του συστήματος τους.

Η εκτίμηση για την αυτόματη εξορυκτική περίληψη για ένα μόνο κείμενο έγινε με τη

συλλογή DUC-2002 με περιλήψεις περίπου 100 λέξεων, το σώμα εγγράφων παρασκηνίου για τον προσδιορισμό της κοινής πληροφορίας (mutual information) των λέξεων ήταν τα DUC-2001 και DUC-2005. Για την εκτίμηση χρησιμοποιήθηκε το εργαλείο ROUGE με μέτρα εκτίμησης τα Rouge-1gram, 2gram και το Rouge-w (weigthed longest common subsequence). Παρακάτω φαίνονται δύο εκτιμήσεις μια χωρίς επαναβαθμολόγιση MMR και μια με επαναβαθμολόγιση που προσφέρει το ROUGE[4].

Σύστημα	Rouge-1	Rouge-2	Rouge-W
Wan et al. (2007) (WN)	**47,10**	**20,42**	**16,34**
Wan et al. (2007) (MI) W=2	46,71	20,20	16,26
Wan et al. (2007) (MI) W=5	46,80	20,26	16,31
Wan et al. (2007) (MI) W=10	46,82	20,30	16,29
SentenceRank (Mihalcea & Tarau, 2004)	45,59	19,20	15,79
MutualRank (Zha, 2002)	43,74	17,99	15,33

Πίνακας 4.2: στη Συλλογή DUC-2002 όπως δηλόντονται στο Wan et al. (2007) χωρίς MMR

Σύστημα	Rouge-1	Rouge-2	Rouge-W
Wan et al. (2007) (WN)	**47,33**	20,25	16,35
Wan et al. (2007) (MI) W=2	47,28	**20,28**	**16,37**
Wan et al. (2007) (MI) W=5	47,28	20,24	16,34
Wan et al. (2007) (MI) W=10	47,22	20,23	16,31
SentenceRank (Mihalcea & Tarau, 2004)	46,26	19,46	16,02
MutualRank (Zha, 2002)	43,80	17,25	15,22

Πίνακας 4.3: στη Συλλογή DUC-2002 όπως δηλόντονται στο Wan et al. (2007) με MMR

Απο τα αποτελέσματα φαίνεται πως τα προτεινόμενα συστήματα έχουν καλύτερες επιδόσεις στη περίληψη απο το σύστημα του Zha (2002) και των Mihalcea & Tarau (2004) (εδώ SentenceRank) που θα περιγραφούν αμέσως μετά. Επίσης τα καλύτερα αποτελέσματα επιτευχθηκαν με τη βαθμολόγηση που πήρε υποψιν της δεδομένα γνώσης απο το WordNet. Στις περιλήψεις που έγιναν χωρίς επαναβαθμολόγηση φαίνεται ότι το μεγαλύτερο παράθυρο φέρνει καλύτερα αποτελέσματα, κάτι που δεν παρατηρείται όταν έχει εφαρμοστεί η MMR, η οποία απαλείφει οποιοδήποτε πλεονασμό που μπορεί να υπάρξει σε μικρά παράθυρα.

4 https://github.com/gregdurrett/berkeley-doc-summarizer/tree/master/rouge/ROUGE

IV.2.4 TextRank (Mihalcea & Tarau, 2004· 2005)

Το σύστημα που αναπτύσσουν οι Mihalcea & Tarau (2004) είναι πολύ απλό και έγινε για να εκτιμηθεί η αξία του τρόπου βαθμολόγησης όρων με PageRank. Οι προτάσεις έχουν το ρόλο κόμβου στον γράφο που κατασκευάζεται και οι σχέσεις μεταξύ τους είναι σχέσης ομοιότητας που υπολογίζεται σύμφωνα με την επικάλυψη λέξεων των δύο προτάσεων βλ. Εξ (2). Αν δύο προτάσεις δεν περιέχουν καμία κοινή λέξη τότε αυτή η ομοιότητα θα είναι μηδενική. Με τον υπολογισμό του PageRank για μη κατευθυνόμενους γράφους με βάρη βαθμολογείται η κάθε πρόταση.

$$\mathrm{Sim}(S_i, S_j) = \frac{\left|w_k \vee w_k \in S_i \wedge w_k \in PS_j\right|}{\log\left(\left|S_i\right|\right) + \log\left(\left|S_j\right|\right)} \quad (2)$$

Η εκτίμηση του συστήματος γίνεται στο DUC-2002 με το εργαλείο ROUGE στην δημιουργία περίληψης 100 λέξεων, με μέτρα εκτίμησης το Rouge-1gram και τα αποτελέσματα που δηλώνουν στην εργασία είναι στα πέντε καλύτερα του συνεδρίου (DUC-2002). Εξετάζονται τρείς διαφορετικές προεπεξεργασίες. α) Το κείμενο ως έχει (raw), β) με αποκοπή καταλήξεων (stemmed) γ) με αποκοπή καταλήξεων και αφαίρεση διακοπτουσών λέξεων (stemmed, stopwords). Στο Mihalcea & Tarau (2005) έγινε το ίδιο πείραμα με προεπεξεργασία αποκοπής καταλήξεων, δοκιμάζοντας κατευθυνόμενους γράφους όπου η κατεύθυνση είναι η εμφάνιση των προτάσεων κατα τη ροή του κειμένου και η αντίστροφη. Η βαθμολογίση των κόμβων γίνεται με HITS και TextRank. Φαίνεται πως με κατευθυνόμενους γράφους η επίδοση και του TextRank και του HITS είναι συγκρίσιμη με το καλύτερο σύστημα του DUC-2002.

Σύστημα	raw	stemmed	stemmed, stopwords
S27	48,14	50,11	44,05
S31	47,15	49,14	41,60
TextRank	*47,08*	*49,04*	*42,29*
S28	47,03	48,90	43,46
S21	46,83	48,69	42,22

Πίνακας 4.4

Σύστημα	Undirected	Forward Dir.	Backward Dir.
HITS (-stemmed)	49,12	45,84	**50,23**
TextRank (-stemmed)	49,04	42,02	**50,08**

Πίνακας 4.5

Η σύγκριση μπορεί να γίνει και με τα αποτελέσματα στο Wan et al. (2007), το πρώτο που παρατηρήται είναι η διαφορά ανάμεσα στα αποτελέσματα του TextRank στα δύο πειράματα. Το Wan et al. (2007) δεν είναι λεπτομεριακό ως προς το πως υπολόγισε το TextRank όμως εμφανίζει πτώση 2 μονάδων απο τα αποτελέσματα που παρουσιάζουν οι Mihalcea & Tarau (2004) αν υποθέτεθεί πως υπολογίστηκε χωρίς προεπεξεργασία στο κείμενο. Πολύ είναι οι παράγωντες που μπορούν να έχουν συνδράμει στη μείωση, αλλα λείπουν πληροφορίες για να προσδιορισθόυν.

Τα πειράματα γίναν και σε μια συλλογή στη πορτογαλλική για να διερευνηθεί η υπόθεση

ότι το TextRank και το HITS λειτουργούν ανεξάρτητα απο τη γλώσσα. Τα αποτελέσματα ήταν παρόμοια με την αγγλική. Αυτη η υπόθεση μπορεί να υποστηριχθεί τουλάχιστον για τις λατινογενείς γλώσσες.

Πέρα απο το σύστημα αυτόματης εξορυκτικής περίληψης για ένα μόνο έγγραφο που ανέπτυξαν στα Mihalcea & Tarau (2004· 2005), στην εργασία του 2005 γίνονται πειράματα και με περιλήψεις απο πολλαπλά έγγραφα. Η μέθοδος τους είναι να γίνει περίληψη για κάθε έγγραφο στη συλλογή ή σε μια συστάδα εγγράφων με το παραπάνω σύστημα και στη συνέχεια να γίνει μια "περίληψη περιλήψεων" (ή μεταπερίληψη) με το ίδιο σύστημα και με τις βαθμολογίσεις PageRank ή HITS, όπως παραπάνω. Για την αποφυγή επαναλήψεων των ίδιων θεματικών και την παράθεση των ίδιων πληροφοριών εισάγουν στο σύστημα ένα ανώτερο όριο στην ομοιότητα (ρυθμισμένο στο 0,5 στα πειράματα τους) πάνω απο το οποίο δεν συμερρίζονται οι σχέσεις και δεν εγγράφονται ακμές.

Η εκτίμηση γίνεται πάλι στη συλλογή DUC-2002 για περίληψη πολλαπλών εγγράφων, με το εργαλείο Rouge και με μέτρο εκτίμησης το Rouge-1gram. Τα αποτελέσματα φαίνονται παρακάτω, συγκρινόμενα με τα καλύτερα πέντε του DUC του 2002.

Τα καλύτερα 5 συστήματα για περίληψη πολλαπλών εγγράφων στο DUC-2002					Βάση
S26	S19	S29	S25	S20	
35,78	34,47	32,64	30,56	30,47	29,32

Πίνακας 4.6

Αλγόριθμος για περίληψη των μονών εγγράφων	Αλγόριθμος μετα-περίληψης		
	TextRank (-und)	TextRank (-dback)	HITS (-dback)
TextRank (-und)	**35,52**	34,99	34,65
TextRank (-dback)	35,02	34,48	34,39
HITS (-dback)	**35,72**	35,20	34,73

Πίνακας 4.7

Τα αποτελέσματα δείξαν πως όπως και στη περίληψη ενός μόνο εγγράφου ο HITS με αντίστροφη κατεύθυνση απο τη ροή του κειμένου δίνει καλύτερα αποτελέσματα αν και κατα τη περίληψη των περιλήψεων η κατεύθυνση (τουλάχιστον η αντίστροφη κατεύθυνση) μειώνει την επίδοση. Η διαφορές δεν είναι πολύ μεγάλες για να θεωρηθούν σημαντικές. Ώς προς τα καλύτερα αποτελέσματα του DUC-2002, φαίνεται πως όλες οι μέθοδοι που περιγράφονται εδώ μπορούν να συγκριθούν με την καλύτερη (Rouge-1 = 34,37).

IV.2.5 LexRank (Erkan & Radev, 2004· Otterbacher et al. 2009)

Συγχρόνως, και ανεξάρτητα απο τους Mihalcea & Tarau, οι Erkan & Radev (2004) στην εργασία του ανέπτυξαν μια μέθοδο υπολογισμού σημαντικότητας των προτάσεων για περίληψη πολλαπλών εγγράφων, βασισμένο στο PageRank, ως εναλλακτική στις μεθόδους με ψευδοέγγραφο (centroid). Για τη διαπίστωση της ομοιότητας ανάμεσα στις προτάσεις υπολογίζουν μια παραλλαγή της συνημιτονοειδούς ομοιότητας που ενσωματώνει τον υπολογισμό της αντίστροφου της συχνότητας των κειμένων idf (βλ. επόμενη ενότητα).

$$cosine-idf(x,y)=\frac{\sum_{w\in x,y} tf_{wx}\, tf_{wy}\,(idf_w)^2}{\sqrt{\sum_{x_i\in x}(tf_{x_i x}\, idf_{x_i})^2}\times\sqrt{\sum_{y_i\in y}(tf_{y_i y}\, idf_{y_i})^2}} \quad (3)$$

Οι κόμβοι του γράφου είναι οι προτάσεις των κειμένων και οι ακμές δηλώνουν την συσχέτιση τους με βάρη τον βαθμό ομοιότητας αναμεταξύ τους. Για τη διαπίστωση του βαθμού σημαντικότητας της κάθε πρότασης υπάρχουν δύο μεθόδοι. Ο πρώτος ονομάζεται LexRank και διατηρεί τις ακμές που έχουν βάρος (δηλαδή τις προτάσεις που βαθμό ομοιότητα) πάνω απο ένα ορισμένο κατώφλι, στη συνέχεια υπολογίζεται ο PageRank για μη κατευθυνόμενους γράφους χωρίς βάρη. Ο δεύτερος, Continuous LexRank, υπολογίζει τον PageRank για μη κατευθυνόμενους γράφου με βάρη.

Η εκτιμήσεις για αυτές τις μεθόδους έγινε στη συλλογή DUC-2003 (Task-2), DUC-2004 (Task-2, Task-4a, Task-4b) με το εργαλείο ROUGE με μέτρο εκτίμησης το Rouge-1gram (εμφανίζονται η μικρότερη, η μεγαλύτερη και η μέση τιμή) ενταγμένες στο σύστημα MEAD (Radev et al., 2004) που έχει επιλογή MMR επαναβαθμολόγησης για απαλειφή πλεονασμού, η οποία και χρησιμοποιήθηκε. Οι εξαγώμενες περιλήψης είχαν μέγεθος 665-byte για κάθε συστάδα εγγράφων του σώματος. Για το LexRank το κατώτατο κατώφλι είναι 0,1. Έγιναν συγκρίσης με το βαθμολόγηση βάσει ψευδοεγγράφου (Centroid) και βάσει του βαθμού των κόμβων (Degree). Παρακάτω εμφανίζονται τα αποτελέσματα απο τις συλλογές DUC-2003 (Task-2) και DUC-2004 (Task-2) όπως και τα καλύτερα αποτελέσματα των συνεδρίων σε αυτές τις συλλογές.

Μέθοδος	DUC-2003 Task-2			Μέθοδος	DUC-2004 Task-2		
	Min	Max	Average		Min	Max	Average
Centroid	35,23	37,13	36,24	Centroid	35,80	37,67	36,70
Degree	35,66	36,50	35,95	Degree	35,90	38,30	37,07
LexRank	36,10	37,26	36,66	LexRank	36,46	38,08	37,36
Cont. LexRank	35,94	37,00	36,46	Cont. LexRank	36,17	38,26	37,58

Πίνακας 4.8

DUC-2003 Task-2		DUC-2002 Task-2	
Κωδικός	Avg. Rouge-1 Score	Κωδικός	Avg. Rouge-1 Score
12	37,98	65	38,22
13	36,76	104	37,44
16	36,60	35	37,43
6	36,07	19	37,39
26	35,82	124	37,06

Πίνακας 4.9

Ο LexRank και ο Continuous LexRank αποδίδουν συγκριτικά με τις βάσεις του και μάλιστα είναι βρίσκεται στους πέντε καλύτερους σύμφωνα με τα αποτελέσματα του DUC-2002. Σύγκριση με τις παραπάνω μεθόδους δεν είναι δυνατό να γίνει γιατί δεν υπάρχουν αποτελέσματα απο κοινή συλλογή όμως διαγωνίζονται εξίσου καλά με τα σύγχρονα συστήματα όπως και τα άλλα συστήματα που περιγράφηκαν πιο πάνω.

ΚΕΦΑΛΑΙΟ V. Ανάκτηση Πληροφοριών

V.1 Εισαγωγή

Μια καλή βάση για την διασαφήνιση της έννοιας της ανάκτησης πληροφοριών είναι η περιγραφή της λειτουργίας μιας βιβλιοθήκης. Σε αυτό το παράδειγμα, η πληροφορία προς ανάκτηση είναι τίτλοι βιβλίων· γίνεται ερώτηση στη γραμματεία της βιβλιοθήκης για βιβλία σχετικά με ένα θέμα – ο τρόπος που η γραμματεία συγκεντρώνει τη πληροφορία δεν έχει σημασία προς το παρόν, αφού πέφτει σε ένα άλλο είδος εφαρμογής, της ταξινόμησης εγγράφων – όμως το γεγονός είναι ότι μία λίστα τίτλων βιβλίων παραδίδεται σε αυτόν που ρωτάει, τα οποία έχουν σχέση με το θέμα που αναζητεί. Αυτό είναι ακριβώς το πεδίο και ο σκοπός της ανάκτησης πληροφοριών. Τα *μέσα* προς την ανάκτηση είναι υπολογιστικές διαδικασίες που σχετίζονται με τη πληροφορία που υποβάλλεται ως ερώτημα στο σύστημα και την όλη συλλογή πληροφοριών στην οποία αναζητείται "απάντηση".

V.1.1 Ανάλυση Κειμένων

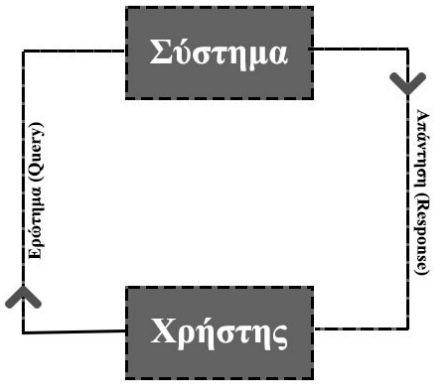
Τα κείμενα ως έχουν είναι ακατέργαστα πράγματα, κατ'αρχην αδόμητα και μόνο ως προς κάτι που θα του προσδώσει δομικά χαρακτηριστικά αποκτάει δομή, όπως στη περίπτωση του ανθρώπου-αναγνώστη ή ενός συστήματος ανάλυσης κειμένων. Ένα σύστημα ανάλυσης κειμένων σχεδιάζεται με τέτοιο τρόπο ώστε να κάνει "κατανοητά" το κείμενα, δηλαδή να τα κάνει μηχανικά αναγνώσιμα (machine-readable), ως προς την επιτέλεση μιας εφαρμογής και γιαυτό τον λόγο διαφοροποιείται από εφαρμογή σε εφαρμογή. Σε γενικές γραμμές, όμως, κάθε ανάλυση δημιουργεί είτε μια συλλογή όρων για κάθε κείμενο είτε μια συλλογή κειμένων για κάθε όρο. Η παραπάνω δομή ονομάζεται *ευρετήριο (index)* και η διαδικασία δημιουργίας της *κατασκευή ευρετηρίου (index construction)* ή *ευρετηρίαση (indexing)*. Αναλόγως του αν η ευρετηρίαση γίνεται από κάποιον ειδικό ή αυτοματοποιημένα, ονομάζεται *χειροκίνητη (manual)* ή *αυτόματη (automatic)* αντίστοιχα. Συνήθως, οι *αυτόματοι ευρετηριοποιητές (indexers)* προσπαθούν να προσεγγίσουν τα αποτελέσματα των ειδικών.
Η δομή του ευρετηρίου είναι παρόμοια με του λεξικού (dictionary), με κλειδιά και τιμές. Το κλασσικό ευρετήριο έχει για κλειδιά κείμενα και για τιμές του όρους του κειμένου. Σε μερικές εφαρμογές, και ειδικά στις εφαρμογές ανάκτησης, αυτός ο τύπος ευρετηρίου δεν είναι αποτελεσματικός, αφού, για παράδειγμα, μια αναζήτηση πρέπει να προσπελάσει όλα τα κείμενα του ευρετηρίου προς εύρεση αυτών που περιλαμβάνουν τον ζητούμενο όρο. Σε αντίθεση, ένα ευρετήριο του οποίου τα κλειδιά είναι οι όροι και οι τιμές τα κείμενα που τους περιλαμβάνουν δεν παρουσιάζει αυτό το πρόβλημα. Αυτό ονομάζεται *ανεστραμμένο ευρετήριο (inverted index)*. Πάνω στα ευρετήρια θα βασιστούν όλες οι εφαρμογές που θα περιγραφούν και η διαδικασία κατασκευής ευρετηρίου θα υποδηλώνεται σε κάθε μια από αυτές.
Η δομή ενός συστήματος ανάκτησης δεδομένων

V.1.2 Ο Χρήστης και το Σύστημα

Το πιο γενικό επίπεδο του μοντέλου ανάκτησης πληροφοριών μπορεί να προσδιορισθεί από τον γενικότερο τρόπο λειτουργίας της ίδιας της ανάκτησης. Η ανάκτηση πληροφοριών στην πραγματικότητα *είναι* ανάκτηση *εγγράφων*, που περιέχουν, με τον ένα ή τον άλλο τρόπο, πληροφορίες που "αφορούν" τον χρήστη (ανάκτηση εγγράφων). (Rijsbergen, 1979) Έτσι, τα πρωταρχικά στοιχεία του μοντέλου είναι ο χρήστης και το σύστημα.

Σε αυτή τη μορφή του μοντέλου, το σύστημα και ο χρήστης είναι σαν "μαύρα κουτιά" και χαρακτηρίζονται μόνο από τη συνδιαλλαγή τους. Ο χρήστης ρωτάει και το σύστημα δέχεται την ερώτηση και απαντάει στον χρήστη που δέχεται την απάντηση.(Σχήμα 5.2) Εφόσον ο χρήστης ξεκινάει τη συνδιαλλαγή και αιτεί πληροφορίες από το σύστημα, πρέπει να διευκρινισθεί πρώτα από που προέρχεται και τη μορφή έχει αυτό το αίτημα. Οπότε, αναλύοντας περαιτέρω το στοιχείο "χρήστης" φαίνεται πως, ως επί το πλείστον, το απαρτίζουν τα εξής χαρακτηριστικά στοιχεία· έχει ένα *γνωστικό υπόβαθρο (background knowledge),* βρίσκεται μέσα ένα *πλαίσιο (context)* και έχει *ανάγκη πληροφοριών (information need)* (Nie et al., 1998 σσ.19-21), βάσει των οποίων θα σχηματίσει το αίτημα του με τη μορφή *ερωτήματος (query)*. Όλα τα παραπάνω στοιχεία είναι άγνωστα στο σύστημα και το μόνο σημείο διεπαφής είναι το ερώτημα. Οπότε, βάσει του ερωτήματος, κατ'αναλογία, σκοπός του συστήματος, τώρα, είναι να επιστρέψει μια *απάντηση (response)* που θα ικανοποιεί την ανάγκη του χρήστη, ο οποίος βρίσκεται μέσα στο δικό του πλαίσιο, κατέχοντας το δικό του γνωστικό υπόβαθρο. Εδώ εμφανίζεται η ανάγκη εισαγωγής της έννοιας της *συνάφειας (relevance)*· Αναδιατυπώνοντας, η απάντηση του συστήματος πρέπει να είναι συναφής με το ερώτημα του χρήστη και κατ'επέκταση με την ανάγκη του που "μεταφράζεται" σε αυτό κ.λπ.

Ένα πρώτο πρόβλημα που εμφανίζεται είναι η συχνή *αναντιστοιχία* της ανάγκης πληροφοριών του χρήστη και της μετατροπής της σε ερώτημα. Οπότε, η υπόθεση ότι ο χρήστης ξέρει πως να θέτει τα ερωτήματα είναι αβέβαιη και προϋποθέτει πολλούς παράγοντες που, κατά κύριο λόγο, έχουν σχέση με το γνωστικό του υπόβαθρο. Θεωρητικά ο βαθμός *αποτελεσματικότητας (effectiveness)* του συστήματος είναι ανάλογος με τη δυνατότητα ταύτισης της κρίσης της συνάφειας (relevance judgment) του χρήστη και της κρίσης της συνάφειας του συστήματος, δηλαδή, της ικανοποίησης των αναγκών του χρήστη από το σύστημα. Αυτή είναι η πρώτη μορφή του μοντέλου και περιλαμβάνει το σύστημα και τον χρήστη. Εν συνεχεία, η σχέση του συστήματος με τον χρήστη θα παραμείνει υπόρρητη μέσα στις υποθέσεις κάθε πραγματικού συστήματος.



Σχήμα 5.1: Ο χρήστης και το σύστημα ανάκτησης

V.1.3 Το σύστημα

Ένα άλλο επίπεδο του μοντέλου εστιάζει στο *σύστημα* της ανάκτησης πληροφοριών και αποτελεί τον πυρήνα του ανώτερου επιπέδου εφόσον είναι το μόνο από τα δύο στοιχεία του που μπορεί να τροποποιηθεί τεχνικά και να δομηθεί αρχιτεκτονικά – ο χρήστης μπορεί να θεωρηθεί απρόβλεπτος και δεν είναι στόχος αυτής της εργασίας να τον περιγράψει. Όπως επισημάνθηκε παραπάνω, το σύστημα ανάκτησης δέχεται ένα ερώτημα στη είσοδο του και πρέπει να ανακτήσει σχετικά έγγραφα τα οποία στέλνει στην έξοδο του.

Το σύστημα διαθέτει μια συλλογή εγγράφων από την οποία γίνεται η ανάκτηση και η οποία ονομάζεται *σώμα κειμένων ή corpus*. Τα έγγραφα είναι έγγραφα κειμένων, όπως αναφέρθηκε παραπάνω. Από εδώ και πέρα έγγραφα και κείμενα θα χρησιμοποιούνται με το ίδιο σχεδόν νόημα. Τα έγγραφα ευρετηροποιούνται αφού πρώτα έχουν περάσει από κάποια διαδικασία προεπεξεργασίας. Στην προεπεξεργασία μπορούν επιτελούνται διάφορες εργασίες αναλόγως το εκάστοτε σύστημα, π.χ. μπορούν να αποκοπούν οι καταλήξεις των λέξεων, να αφαιρεθούν οι διακοπτούσες λέξεις κ.λπ. Συνήθως τα συστήματα ανάκτησης πληροφοριών διατηρούν ένα ανεστραμμένο ευρετήριο γιατί είναι πιο αποδοτικό στην εύρεση και οι λέξεις ενός ευρετηρίου θα ονομάζονται λεξικά.

Αφού το σύστημα λάβει το ερώτημα το επεξεργάζεται παρόμοια για να μετασχηματίσει τους όρους του σε αντιστοιχία με τους όρους των ευρετηρίων. Στη συνέχεια, επιτελείται μια σύγκριση ανάμεσα στους όρους του ερωτήματος και τους όρους κάθε εγγράφου. Η γενική ιδέα είναι πως αν οι όροι του ερωτήματος εμφανίζονται σε κάποιο έγγραφο αυτό πρέπει να είναι σχετικό με το ερώτημα. Αυτή ακριβώς είναι η στιγμή που υπολογίζεται η συνάφεια του εγγράφου προς το ερώτημα. Ο υπολογισμός γίνεται μέσω ενός αλγορίθμου που θα βαθμολογεί το κάθε έγγραφο του σώματος.

Οι πρώτοι αλγόριθμοι υπολογισμού συνάφειας ήταν δυαδικοί και βασίζοντας απλώς στην εμφάνιση ή μη του όρου στο έγγραφο, επιπλέον το ερώτημα ήταν ανάγκη να είναι γραμμένο με λογικούς τελεστές (Boolean operators). Τα έγγραφα που πληρούν τις προϋποθέσεις του ερωτήματος (π.χ. να εμφανίζονται οι λέξεις “quantum” και “computer” χωρίς την εμφάνιση της λέξης “magic”) επιστρέφονται στον χρήστη από το σύστημα (Lancaster & Fayen, 1973).

Αναγνωρίζοντας τα μειονεκτήματα του δυαδικού μοντέλου, το διανυσματικό μοντέλο επιτρέπει στον αλγόριθμο του να μην υπάρχει απόλυτο ταίριασμα και βαθμολογεί την εγγύτητα του ερωτήματος και του εγγράφου. Κάθε έγγραφο αναπαριστάται από ένα διάνυσμα που περιλαμβάνει όλους του λεξικού με τιμές τα βάρη που διαθέτει ο κάθε όρος σε αυτό. Μια συλλογή κειμένων μπορεί, τότε, να αναπαρασταθεί ως ένας πίνακας όρων-εγγράφων συντεθειμένο από τα παραπάνω διανύσματα. Όταν δίνεται ένα ερώτημα η συνάφεια υπολογίζεται μέσω μιας συνάρτησης ομοιότητας ανάμεσα στο ερώτημα και τα έγγραφα που παίρνει υπόψιν της αυτά τα βάρη. Τα βάρη τον όρων στο κείμενο έχουν διάφορους τρόπους να υπολογισθούν. Συστατική στις περισσότερες μεθόδους στάθμισης των όρων είναι η συχνότητα tf_{ik} με την οποία εμφανίζεται ένας όρος i σε ένα κείμενο k. Η συχνότητα όμως, ακόμα κι αν έχουν αφαιρεθεί κατά την προεπεξεργασία οι διακοπτούσες λέξεις, επιτρέπει σε όρους χωρίς ιδιαίτερη σημαντικότητα να αποκτήσουν μεγάλο βάρος μονάχα επειδή εμφανίστηκαν πολλές φορές στο έγγραφο. Τις περισσότερες φορές, για να αποφευχθεί αυτό το πρόβλημα η συχνότητα κάθε όρου μπορεί να πολλαπλασιαστεί με τον αντίστροφο της συχνότητας των κειμένων (inverse document frequency, idf) που περιέχουν τον όρο (tf-idf), βλ. Εξ (2), (3). Έτσι, λέξεις που εμπεριέχονται σε πολλά κείμενα και άρα δεν έχουν διακριτική αξία για τα έγγραφα τιμωρούνται. Από την άλλη υπάρχουν επίσης διάφοροι τρόποι υπολογισμού της εγγύτητας ερωτήματος-εγγράφου. Μια διαδεδομένη συνάρτηση ομοιώτητας είναι η συνημιτονοειδής ομοιότητα (cosine similarity) μεταξύ διανυσμάτων. Επίσης, κάποιες φορές εφαρμόζονται βάρη και στους όρους του ερωτήματος. Όταν βαθμολογηθούν όλα τα έγγραφα σύμφωνα με την

συνάφεια τους προς το ερώτημα επιλέγεται ένας αριθμός ή ένα ποσοστό από αυτά και επιστρέφεται στο χρήστη.(σύστημα SMART: Salton & McGill, 1983).

Ακολούθησαν κι άλλα μοντέλα όπου κατά κύριο λόγο επεμβαίνουν και αναπτύσουν τον τρόπο που σταθμίζονται οι όροι και υπολογίζεται η ομοιότητα ανάμεσα στο ερώτημα και τα έγγραφα. [π.χ. BM25 (Robertson et al., 1995), Dirichlet prior (Zhai & Lafferty, 2001), Piv (Singhal, 2001)]

Η εκτίμηση του κάθε συστήματος γίνεται με δοσμένες συλλογές, ερωτήματα και μία λίστα με τα αντίστοιχα συναφή έγγραφα ταξινομημένα κατά το βαθμό συνάφειας τους για κάθε ερώτημα.

Τα συστήματα που βασίζονται σε γράφους συνήθως υπολογίζουν τη στάθμιση των όρων με τεχνικές γράφων και τη συνάρτηση ομοιότητας με τα διανύσματα που προκύπτουν όπως και στα άλλα συστήματα.

$$Cosine(D_i, Q_j) = \frac{\sum_{k=1}^{t} (d_{ik} \cdot q_{jk})}{\sqrt{\sum_{k=1}^{t} d_{ik}^{2}} \cdot \sqrt{\sum_{k=1}^{t} q_{jk}^{2}}} \qquad (1)$$

$$idf_t = \frac{N}{df_t}, \qquad (2)$$

$$w(t,d) = tf_{td} * idf_t, \qquad (3a)$$

ή

$$w(t,d) = \log(tf_{td}) * \log(idf_t) \qquad (3b)$$

V.2 Γράφοι και Ανάκτηση Κειμένων

V.2.1 TextLink, PosLink, TextRank, PosRank (Blanco & Lioma, 2012).

Η στάθμιση των όρων του κειμένου στο πλαίσιο ενός γράφου όρων αντιστοιχεί στον υπολογισμό του βαθμού σημαντικότητας του κάθε κόμβου. Όπως φάνηκε παραπάνω υπάρχουν διάφοροι τρόποι υπολογισμού της σημαντικότητας των κόμβων. Οι Blanco & Lioma (2012) μελετούν τους περισσότερους απο τους τρόπους βαθμολόγησης των κόμβων στην εργασία ανάκτησης πληροφοριών. Κατά την προεπεξεργασία δεν αφαιρούν τις διακοπτούσες λέξεις, δεν αποκόπτουν τις καταλήξεις και ούτε επιλέγουν για όρους συγκεκριμένα μέρη του λόγου ενώ σημειώνεται το μέρος του λόγου που αντιστοιχέι σε κάθε λέξη.

Αρχικά κατασκευάζουν δύο γράφους, έναν κατευθυνόμενο κι έναν μη κατευθυνόμενο όπου οι κόμβοι αντιστοιχουν σε όρους και οι ακμές σε συνεμφανίση των όρων μέσα σε ένα παράθυρο μεγέθους W = 10. Για τον κατευθυνόμενο γράφο, επιπλέον, εγγράφεται κατεύθυνση που βασίζεται στα μέρη του λόγου της κάθε λέξης. Η κατεύθυνση είναι βασισμένη στη Θεωρία Βαθμού (Rank Theory) του Jespersen (1929) στην οποία ορισμένα μέρη του λόγου τροποοιούν κάποια άλλα. Εφόσον οι κατευθύνσεις στο γράφο μπορούν να ερμηνευτούν ως σχέσεις εξάρτησης, όταν μια λέξη τροποποιεί μια άλλη, τότε εγγράφεται στην ακμή μια κατεύθυνση απο τη πρώτη στη δεύτερη. α)Τα *ουσιαστικά* μπορούν να τροποποιήσουν άλλα ουσιαστικά, β) τα *ρήματα* τροποποιούν ουσιαστικά, επίθετα και άλλα ρήματα, γ) τα *επίθετα* ουσιαστικά, ρήματα και άλλα επίθετα και δ) τα *επιρρήματα* τροποποιούν όλα τα παραπάνω. Στον μη κατευθυνόμενο

γράφο οι κόμβοι σταθμίζονται με δύο διαφορετικούς τρόπους, μέσω του βαθμού τους (όπου ονομάζεται βαθμολογία TextLink) και μέσω του PageRank (TextRank) τους για μη κατευθυνόμενους γράφους. **(**εξίσωσεις (I.1), (I.6)**)** Στον κατευθυνόμενο γράφο οι κόμβοι σταθμίζονται με τους αντίστοιχους υπολογισμούς με κατευθύνσεις, με τις ονομασίες PosLink και PosRank **(**εξισώσεις (I.2), (I.7)**)**. Στη συνέχεια, απο τα αποτελέσματα των παραπάνω υπολογισμών (*tw),* υπολογίζεται η βαθμολογία κάθε όρου και αντιστοιχεί στον λογάριθμο του βάρους του κόμβου που του αντιστοιχεί επι του λογαρίθμου του αντιστρόφου της συχνότητας των κειμένων της συλλογής που περιέχουν τον όρο, όπως στα κλασικά μοντέλα, (raw) βλ. εξ (3b). Στη βαθμολογία μπορούν να ενσωματοθούν και μέτρα του γράφου ως τοπολογικές ιδιότητες του γραφου. Για την ενσωμάτωση χρησιμοποιείται η τεχνική της ενσωμάτωσης κορεσμούς (*satu integration* βλ. Craswell et al., 2005) όπως στην εξίσωση (4), όπου P_d είναι το επιλεγμένο μέτρο του γράφου και ψ, κ είναι παράμετροι. Τα μέτρα γράφου που μελετώνται είναι ο μέσος βαθμός $\bar{d}(G)$, το μέσο μέγεθος διαδρομής *l(G)* και ο καθολικός συντελεστής συσταδοποίησης *c(G)*, όπως επίσης κι ένα μέτρο, αντίστοιχο του μεγέθους εγγράφου (document length), το αθροίσμα των βαρών όλων των κόμβων (sum). Τα l(G) και c(G) υπολογίζονται προσεγγιστηκά με τις εξισώσεις (5), (6), σύμφωνα με το Albert & Barabási (2002) και τις παρατηρήσεις στο κεφάλαιο που περιγράφει το γλωσσικό δίκτυο. Κάθε μία απο τις τοπολογικές ιδιότητες του γράφου ερμηνεύεται σαν μια ιδιότητα του εν λόγω κειμένου και αναλόγως ενσωματώνονται στη βαθμολόγιση. Ο μέσος βαθμός στον γράφο εμφανίζει τη διαπλοκή που έχουν οι κόμβοι ενός γράφου, για ένα κείμενο αυτό δηλώνει το αντίστροφως ανάλογο της συνοχής του, έτσι $P_d = 1/\bar{d}(G)$. Όπως αναφέρθηκε παραπάνω (βλ. Εισαγωγή), το μέσο μέγεθος διαδρομής του γράφου είναι ο μέσος των ελάχιστων βημάτων που πρέπει να διανυθούν μεταξύ όλων των κόμβων του, σε ένα κείμενο όσο μικρότερο είναι το μέγεθος διαδρομής τόσο πιο καλά συνδεδεμένη είναι η ανάπτυξη του λόγου του, άρα $P_d = 1/l(G)$. Όσο για το καθολοκό συντελεστή του κειμένου, όσο μεγαλύτερος είναι τόσο πιο πολλές θεματικές αναπτύσσονται σε αυτό, με $P_d = c(G)$. Και τέλος, το άθροισμα των βαρών των κόμβων θα ενσωματωθεί αντιστρόφος εν είδει κανονικοποίησης που τιμωρεί μεγάλα έγγραφα, $P_d = 1/\text{sum}$.

$$w_g(t,d) = \log(tw) * \log(idf) + \psi \frac{P_d}{\kappa + P_d} \quad (4)$$

$$l(G) \approx \ln \frac{(|(V(G))|)}{\ln(\bar{d}(G))} \quad (5)$$

$$c(G) \approx \frac{\bar{d}(G)}{|(V(G))|} \quad (6)$$

Για την εκτίμηση των παραπάνω μοντέλων χρησιμοποιήθηκαν τρείς διαφορετικές συλλογές του συνεδρίου ανάκτησης (TREC), το Disk4&5 (εξαιρουμένων των αρχείων του Κονγκρέσου) με ειδησεογραφικά άρθρα, το WT2G με σελίδες διαδυκτίου και το BLOG06 που αποτελείται απο ροές μπλόγκ (feeds). **Τα στατιστικά των συλλογών είναι στον πίνακα .**

Τα αποτελέσματα συγκρίνονται με βάση αυτά του BM25, tf-idf και τα μέτρα εκτίμησης που χρησιμοποιούνται είναι η Μέση Αντιπροσωπευτική Πιστότητα (Mean Average Precision, MAP), η Πιστότητα στα 10 (P@10) και η η δυαδική προτίμηση (Bpref).

Συλλογή	# Έγγραφα	# Όροι
WT2G	247.491	1.002.586
Disk4&5	528.155	840.536
BLOG06	3.215.171	4.968.020

Πίνακας 5.1: Στατιστικά συλλογών, όπως αναφέρονται στο Blanco & Lioma (2012)

MAP	Raw	+ $\bar{d}(G)$	+l(G)	+c(G)	+sum	BM25	TFIDF
WT2G							
TextRank	30,33	30,64	30,57	30,83	30,53	29,98	22,68
TextLink	27,77*	30,23	29,76	29,62	28,50		
PosRank	29,50	30,40	29,78	29,71	29,42		
PosLink	28,02	**31,27**	29,17	29,28	28,94		
D4&5							
TextRank	22,43	23,04	22,43	**23,29***	23,07	22,98	19,35
TextLink	20,30	22,33	21,47*	21,97*	21,80*		
PosRank	21,91	22,56	22,05	22,89	22,55		
PosLink	20,20	22,39	21,24*	21,72*	21,78*		
BLOG06							
TextRank	35,03	35,01	36,17	35,83	35,31	36,62	29,63
TextLink	36,57	36,97	**39,47**	***39,06****	38,19		
PosRank	38,74	***38,97****	39,44*	***39,18****	***39,03****		
PosLink	36,74	***39,03****	37,78	***38,33****	38,33		

Πίνακας 5.2: Mean Average Precision. Όπως δηλώνονται στο Blanco & Liomal (2012). Με έντονους χαρακτήρες τα καλύτερα αποτελέσματα ανα συλλογή. (*): Στατιστική σημαντικότητα στο p<0.05

P@10	Raw	+ $\bar{d}(G)$	+l(G)	+c(G)	+Sum	BM25	TFIDF
WT2G							
TextRank	48,20	39,60	50,00	50,00	49,80	49,50	41,20
TextLink	42,60	48,40	46,60	47,00	44,00		
PosRank	48,20	50,20	50,20	**50,80**	49,60		
PosLink	43,20	50,40	58,40	45,40	45,00		
D4&5							
TextRank	40,60	42,41	41,12	42,29*	41,65	43,29	38,55
TextLink	35,18*	40,76	38,96*	39,40	37,71*		
PosRank	40,00	41,00	40,20	41,24	**50,00***		
PosLink	35,22	40,20	38,31*	38,47*	37,55*		
BLOG06							
TextRank	63,80	64,20*	66,80	65,60	64,20	66,80	60,00
TextLink	63,20	65,00*	66,80*	65,00*	63,40		
PosRank	**71,60**	71,00	71,40	***70,60****	***69,60****		
PosLink	63,60	69,40	64,00	64,00	64,00		

Πίνακας 5.3: P@10. Όπως δηλώνονται στο Blanco & Liomal (2012). Με έντονους χαρακτήρες τα καλύτερα αποτελέσματα ανα συλλογή. (*): Στατιστική σημαντικότητα στο p<0.05

BPREF	Raw	$+\bar{d}(G)$	+l(G)	+c(G)	+Sum	BM25	TFIDF
WT2G							
TextRank	29,63	30,09	30,55	30,43	29,81	29,48	23,38
TextLink	27,57*	30,00	29,94	29,80	28,82		
PosRank	29,26	29,61	29,55	29,21	29,20		
PosLink	27,95	**30,81**	29,32	29,29	28,84		
D4&5							
TextRank	23,31	23,75	23,29	24,04*	23,76	24,05	20,94
TextLink	21,47	23,03*	22,46*	22,77*	22,75*		
PosRank	23,19	23,96	23,45*	**24,12**	23,85		
PosLink	21,47	23,15	22,34*	22,61*	22,74*		
BLOG06							
TextRank	40,32	40,28	41,04	40,71	40,41	41,61	36,38
TextLink	42,98	43,00	***44,69****	***44,04****	43,54		
PosRank	44,77	**45,11***	45,51	***44,93****	***44,92****		
PosLink	43,11	***44,89****	43,91	43,89	44,04		

Πίνακας 5.4: Bpref. Όπως δηλώνονται στο Blanco & Liomal (2012). Με έντονους χαρακτήρες τα καλύτερα αποτελέσματα ανα συλλογή. (*): Στατιστική σημαντικότητα στο p<0.05

Παρατηρείται πως οι απλές βαθμολογίσεις χωρίς τις τοπολογικές ιδιότητες (Raw) έχουν εξίσου καλά αποτελέσματα με την B25 (και στη BLOG06 καλύτερα) και πολύ καλύτερα απο την tf-idf, το οποίο υποδεικνύει πως ο βαθμός των κόμβων είναι πιο πληροφοριακό μέτρο απο τον αριθμό εμφανίσεων των όρων στο κείμενο. Επίσης, το BM25 έχει κανονικοποίηση κατα μέγεθος κειμένου, κάτι που οι απλές βαθμολογίσεις (Raw) δεν περιλαμβάνλουν και απο τα στοιχεία του +Sum φαίνεται πως δεν βελτιώνει ιδιαίτερα τα αποτελέσματα, πέρα απο τη πιστότητα στα 10 πρώτα έγγραφα στη συλλογή Disk4&5 με PosRank που έχει σχεδόν 10 μονάδες αύξηση απο την απλή. Οι ενισχυμένες βαθμολογίσεις με τοπολογικές ιδιότητες έχουν σταθερά καλύτερη επίδοση απο τις απλές και εξίσου καλά με την B25 και μεταξύ τους δεν φαίνεται να έχουν σημαντικές διαφορές στα τρία μέτρα εκτίμησης. Τέλος, η διαφορά ανάμεσα στη χρήση κατευθυνόμενων και μη κατευθυνόμενων γράφων φαίνεται να μην έχει ιδιαίτερη διαφορά στα αποτελέσματα.

V.2.2 Graph-of-Words (Rousseau & Varzigiannis, 2013)

Οι Rousseau & Varzigiannis (2013) παρουσιάζουν άλλη μια μέθοδο βαθμολόγησης πάνω στο βαθμό κόμβου, τον graph-of-word. Κατασκευάζεται ένας κατευθυνόμενος γράφος με βάρη, όπου οι κόμβοι αντιπροσωπεύουν τους επεξεργασμένους όρους του κειμένου και οι ακμές την συνεμφάνιση τους σε ένα παράθυρο μεγέθους W=4. Η κατεύθυνση των ακμών ακολουθεί τη φυσική ροή του κειμένου. Κατα τη προεπεξεργασία αφαιρούν τις διακοπτούσες λέξεις. Η βαθμολόγηση των όρων γίνεται πολλαπλασιάζοντας το βαθμό των κόμβων τους με μια συνάρτηση κανονικοποίησης κατά μέγεθος κειμένου, όπως περιγράφεται στο Κεφάλαιο Ι**.** Ο υπολογισμός συνάφειας ερωτήματος-εγγράφου είναι ίδιος με παραπάνω.

Στην εργασία κάναν εκτιμήσεις με τέσσερις διαφορετικές συλλογές απο το TREC, απο τα οποία θα αναφερθούν τα τρία, η συλλογή Disk4&5 που είναι κοινή στις δύο εργασίες και οι συλλογές Disk1&2, WT10G. Τα αποτελέσματα συγκρίνονται με διάφορους άλλους τρόπου

βαθμολόγισης, εδώ επιλέγονται οι BM25 με ρυθμισμένες παραμέτρους και χωρίς (b=0,75) και το Tf-Idf με κανονικοποίηση κατα το μέγεθος των κειμένων. Επίσης εμφανίζονται οι τύποι των παραπάνω και με κανονικοποίηση με κάτω φραγή (βλ. Κεφ. Ι**)**. Τα μέτρα εκτίμησης είναι είναι η Μέση Αντιπροσωπευτική Πιστότητα (Mean Average Precision, MAP) και η Πιστότητα στα 10 (P@10).

Σύστημα	Disk4&5		Disk1&2		WT10G	
	MAP	P@10	MAP	P@10	MAP	P@10
BM25 (untuned, b=0,75)	23,68	41,61	16,60	37,00	18,70	24,79
BM25 + low_bnd	24,66	41,45	15,58	32,07	20,26	25,21
Tf-Idf	21,32	40,64	18,32	41,07	14,30	22,71
Tf-Idf + low_bnd	23,68	41,57	18,25	38,13	16,43	24,38
BM25 (tuned)	25,02	**43,82**	18,93	40,80	21,04	**32,10**
BM25 + low_bnd (tuned)	**25,47**	43,49	16,03	33,40	**21,69**	27,71
TW	21,90	41,33	15,76	40,40	19,46	24,79
TW-Idf	24,03	41,20	**19,73**	**41,48**	21,25	29,17

Πίνακας 5.5: Όπως δηλώνονται στο Rousseau & Varzigiannis(2013). Με έντονους χαρακτήρες τα καλύτερα αποτελέσματα ανα συλλογή.

V.2.3 ConRank (Tu et al., 2016)

Διαφορετική μέθοδο κατασκευής γράφου ακολουθούν οι Tu et al. (2016) για να υπολογίσουν τα βάρη των ακμών σύμφωνα με τη σημασιολογική συσχέτιση των εφαπτόμενων κόμβων τους. Κατα την προεπεξεργασία αφαιρούνται οι διακοπτούσες λέξεις και αφαιρούνται οι καταλήξεις. Στη προκειμένη, ο γράφος είναι ένας κατευθυνόμενος πλήρης γράφος με βάρη. Κόμβοι είναι οι λέξεις και ακμές εγγράφονται ανάμεσα ανάμεσα σε όλους τους κόμβους προς όλες τις κατευθύνσεις. Τα βάρη της ακμής αντιπροσωπεύουν τη πιθανότητα μετάφρασης του ένα όρου στον άλλο και υπολογίζεται με τις εξισώσεις (7) με (15). Στην εξίσωση (9) το dist(ti, tj) είναι η μέση απόσταση μεταξύ των όρων ti και tj σε όλη τη συλλογή. Η εξίσωση (10) υπολογίζει την κοινή πληροφορία (mutual information) ανάμεσα στους όρους. Στις εξισώσεις (11) με (15) το c(Xi=1) είναι ο αριθμός κειμένων που περιέχει τον όρο ti και τα υπόλοιπα (*c(Xi=1, Xj=1), c(Xi=1, Xj=0)* κ.λπ.) ανάλογα. Το βάρος της ακμής ανάμεσα σε δύο κόμβους δηλώνει την σημασιολογική τους συσχέτιση. Αφού υπολογιστούν τα βάρη ο βαθμός του κάθε κόμβου υπολογίζεται με μια παραλλαγή του PageRank για μη κατευθυνόμενους γράφους με βάρη, που αποτελεί ένα στοιχείο του ConRank, βλ εξ. (16). Ο βαθμός συνάφειας ερωτήματος-εγγράφου υπολογίζεται με την εξίσωση 17, όπου συνδυάζονται δύο διαφορετικοί υπολογισμοί, α) τη βαθμολόγηση PageRank για όλους τους όρους του ερωτήματος επι ένα idf [*Con(q, d)*], β) έναν απο τους κλασικούς τρόπους βαθμολόγησης συνάφειας [*R(q, d)*], βλ εξ. (17), (18).

Σύστημα	WT2G		WT10G	
	MAP	P@10	MAP	P@10
BM25	31,28	49,57	21,07	36,26
ComRank-BM25	33,49 (+7,07%)	53,20 (+7,32%)	21,84 (+3,65%)	37,47 (+3,34%)
LMDir	30,57	50,63	20,94	31,08
ComRank-LMDir	32,11 (+5.04%)	54,09 (+6,83%)	21,28 (+1,62%)	31,54 (+1,48%)
MATF	32,41	54,81	22,26	32,83
ComRank-MATF	33,92 (+4,66%)	55,94 (+2,06%)	23,01 (+3,37%)	33,27 (+1,34%)

Πίνακας 5.6

Η σημασιολογική συσχέτιση μεταξύ όρων μπορεί να υπολογιστεί με διάφορους τρόπους. Αξίζει να σημειωθεί πως τα τελευταία χρόνια αναπτύχθηκαν μοντέλα νευρωνικών δικτύων, βασιζόμενα στη μοντελοποίηση γλώσσας, που δημιουργούνε σύμφωνα με το πλαίσιο (συνήθως ένα παράθυρο) κάθε όρου, σε ένα πολύ μεγάλο σώμα κειμένων, διανύσματα πολλών διαστάσεων που αναπαριστούν αυτούς τους όρους, το αποτέλεσμα ονομάζεται word2vec. Στο Mikolov et al. (2013), όπου παρουσιάστηκε για πρώτη φορά με τα μοντέλα continuous bag-of-words και continuous skip-gram, το πιο διαδεδομένο παράδειγμα που περιγράφει τις δυνατότητες τέτοιων αναπαραστάσεων είναι το αλγεβρικού τύπου “King” – “Man” + “Woman” = “Queen”. Έτσι εάν ένα σύστημα έχει στη διάθεση του διανύσματα για κάθε όρο που εμφανίζεται στο σώμα κειμένων η συσχέτιση μεταξύ τους θα είναι δυνατή με κάποια συνάρτηση ομοιώτητας, όπως π.χ. της απλής συνημιτονοειδούς ομοιότητας.

$$trans(t_i,t_j)=\frac{\mathrm{sim}(t_i,t_j)}{\sum_{t' \text{ in } d}\mathrm{sim}(t_i,t')}, \tag{7}$$

$$\mathrm{sim}(t_i,t_j)=MI(ti,tj)*df(t_i,t_j), \tag{8}$$

$$df(t_i,t_j)=\mathrm{e}^{-a*(dist(t_i,t_j)-1)}, \tag{9}$$

$$MI(t_i,t_j)=\sum_{X_i=0,\ 1}\sum_{X_j=0,\ 1}P(X_i,X_j)\log\frac{P(Xi,Xj)}{P(X_i)P(X_j)} \tag{10}$$

$$P(X_i=1)=\frac{c(X_i=1)}{N} \tag{11}$$

$$P(X_i=0)=1-P(X_i=1) \tag{12}$$

$$P(X_i=1,X_j=1)=\frac{c(X_i=1,X_j=1)}{N} \tag{13}$$

$$P(x_i=1,X_j=0)=\frac{c(X_i=1)-c(X_i=1,X_j=1)}{N} \tag{14}$$

$$P(X_i=0,X_j=0)=1-c(X_i=1,X_j=1)-c(X_i=1,X_j=0)-c(X_i=0,X_j=1) \tag{15}$$

$$S_T(t_i)=\lambda\sum_{j\in d}S_T(t_j)trans(t_i,t_j)+(1-\lambda) \tag{16}$$

$$ConRank=k*\frac{Con(q,d)}{1+Con(q,d)}+(1-k)\frac{R(q,d)}{1+R(q,d)} \tag{17}$$

$$Con(q,d)=\sum_{t_j\in q}S(t_j)*IDF(t_j) \tag{18}$$

ΒΙΒΛΙΟΓΡΑΦΙΑ

Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, *74*(1), 47.

Aone, C., Okurowski, M. E., Gorlinsky, J., and Larsen, B. (1999). A trainable summarizer with knowledge acquired from robust NLP techniques. *Advances in automatic text summarization*, *71*.
Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Baralis, E., Cagliero, L., Mahoto, N., & Fiori, A. (2013). GRAPHSUM: Discovering correlations among multiple terms for graph-based summarization. *Information Sciences*, *249*, 96-109.

Baxendale, P. B. (1958). Machine-made index for technical literature—an experiment. *IBM Journal of Research and Development*, *2*(4), 354-361.

Bharti, S. K., & Babu, K. S. (2017). Automatic Keyword Extraction for Text Summarization: A Survey. *arXiv preprint arXiv:1704.03242*.

Blanco, R., & Lioma, C. (2012). Graph-based term weighting for information retrieval. *Information retrieval*, *15*(1), 54-92.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Bollobás, B. (1998). Random graphs. In *Modern graph theory* (pp. 215-252). Springer, New York, NY.

Bougouin, A., Boudin, F., & Daille, B. (2013, October). Topicrank: Graph-based topic ranking for keyphrase extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)* (pp. 543-551).

Braddock, R. (1974). The frequency and placement of topic sentences in expository prose. *Research in the Teaching of English*, *8*(3), 287-302.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107-117.

Carbonell, J., & Goldstein, J. (1998, August). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 335-336). ACM.

Chow, S. L. (1998). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and brain sciences*, *21*(2), 169-194.

Chung, F. (2008). A whirlwind tour of random graphs. *Encyclopedia on Complex Systems, Springer.*

Cohn, D., & Chang, H. (2000, June). Learning to probabilistically identify authoritative documents. In *ICML* (pp. 167-174).

Conroy, J. M., & O'leary, D. P. (2001, September). Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 406-407). ACM.

Craswell, N., Robertson, S., Zaragoza, H., & Taylor, M. (2005, August). Relevance weighting for query independent evidence. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 416-423). ACM.

Das, D., & Martins, A. F. (2007). A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, *4*, 192-195.

Dorogovtsev, S. N., & Mendes, J. F. (2002). Evolution of networks. *Advances in physics*, *51*(4), 1079-1187.

Dorogovtsev, S. N., & Mendes, J. F. F. (2001). Language as an evolving word web. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*(1485), 2603-2606.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, *16*(2), 264-285.

Erdös, P., & Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae (Debrecen)*, *6*, 290-297.

Fang, H., Tao, T., & Zhai, C. (2004, July). A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 49-56). ACM.

Fattah, M. A., & Ren, F. (2009). GA, MR, FFNN, PNN and GMM based models for automatic text summarization. *Computer Speech & Language*, *23*(1), 126-144.

Florescu, C., & Caragea, C. (2017). PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1105-1115).

Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-specific keyphrase extraction. In *16th International Joint Conference on Artificial Intelligence (IJCAI 99)* (Vol. 2, pp. 668-673). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, *47*(1), 1-66.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, *30*(4), 1141-1144.

Gollapalli, S. D., & Caragea, C. (2014, July). Extracting Keyphrases from Research Papers Using Citation Networks. In *AAAI* (pp. 1629-1635).

Grineva, M., Grinev, M., & Lizorkin, D. (2009, April). Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web* (pp. 661-670). ACM.

Guan, J. (2016). A study of the use of keyword and keyphrase extraction techniques for answering biomedical questions.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.

Hasan, K. S., & Ng, V. (2010, August). Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 365-373). Association for Computational Linguistics.

Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1262-1273).

Hovy, E., & Lin, C. Y. (1998, October). Automated text summarization and the SUMMARIST system. In *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998* (pp. 197-214). Association for Computational Linguistics.

Hulth, A. (2003, July). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 216-223). Association for Computational Linguistics.

Hulth, A., & Megyesi, B. B. (2006, July). A study on automatically extracted keywords in text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 537-544). Association for Computational Linguistics.

i Cancho, R. F., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*(1482), 2261-2265.

i Cancho, R. F., Solé, R. V., & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E*, *69*(5), 051915.

Jiang, X., Hu, Y., & Li, H. (2009, July). A ranking approach to keyphrase extraction. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 756-757). ACM.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11-21.

Jones, K. S. (1999). Automatic summarizing: factors and directions. *Advances in automatic text summarization*, 1-12.

Kasture, N. R., Yargal, N., Singh, N. N., Kulkarni, N., & Mathur, V. (2014). A survey on methods of abstractive text summarization. *Int. J. Res. Merg. Sci. Technol*, *1*(6), 53-57.

Kleinberg, J. M. (1997). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, *46*(5), 604-632.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (1999, July). The web as a graph: Measurements, models, and methods. In *International Computing and Combinatorics Conference* (pp. 1-17). Springer, Berlin, Heidelberg.

Kulig, A., Drożdż, S., Kwapień, J., & Oświęcimka, P. (2015). Modeling the average shortest-path length in growth of word-adjacency networks. *Physical Review E*, *91*(3), 032810.

Kupiec, J., Pedersen, J., & Chen, F. (1995, July). A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 68-73). ACM.

Lancaster, F. W., & Fayen, E. G. (1973). Information Retrieval On-Line, Melville Publ. *Co., Los Angeles, Calif*, 15.

Liu, F., Pennell, D., Liu, F., & Liu, Y. (2009, May). Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics* (pp. 620-628). Association for Computational Linguistics.

Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010, October). Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 366-376). Association for Computational Linguistics.

Liu, Z., Li, P., Zheng, Y., & Sun, M. (2009, August). Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 257-266). Association for Computational Linguistics.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159-165.

Lv, Y., & Zhai, C. (2011, October). Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 7-16). ACM.

Mani, I. (2001). *Automatic summarization* (Vol. 3). John Benjamins Publishing.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, No. 1, p. 496). Cambridge: Cambridge university press. Web: https://nlp.stanford.edu/IR-book/

Medelyan, O., Frank, E., & Witten, I. H. (2009, August). Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3* (pp. 1318-1327). Association for Computational Linguistics.

Mehler, A., Lücking, A., Banisch, S., Blanchard, P., & Job, B. (Eds.). (2016). *Towards a*

theoretical framework for analyzing complex linguistic networks. Springer Berlin Heidelberg.

Mel′čuk, I. A. (1988). *Dependency syntax: theory and practice*. SUNY press.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI* (Vol. 6, pp. 775-780).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mitra, M., Singhal, A., & Buckleytt, C. (1997). Automatic text summarization by paragraph extraction. *Intelligent Scalable Text Summarization*.

Moens, M. F. (2002). *Automatic indexing and abstracting of document texts*. Kluwer Academic Publishers.

Nenkova, A., & McKeown, K. (2011). Automatic summarization. *Foundations and Trends® in Information Retrieval*, *5*(2–3), 103-233.

Nguyen, C. Q., & Phan, T. T. (2007). A Pattern-based Approach to Vietnamese Key Phrase Extraction, In *Addendum Contributions of the 5th International IEEE Conference on Computer Sciences- RIVF'07*: (pp. 41-46).

Nguyen, C. Q., & Phan, T. T. (2009, August). An ontology-based approach for key phrase extraction. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 181-184). Association for Computational Linguistics.

Nguyen, T. D., & Kan, M. Y. (2007, December). Keyphrase extraction in scientific publications. In *International Conference on Asian Digital Libraries* (pp. 317-326). Springer, Berlin, Heidelberg.

Nguyen, T. D., & Luong, M. T. (2010, July). WINGNUS: Keyphrase extraction utilizing document logical structure. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 166-169). Association for Computational Linguistics.

Nie, J. Y., Isabelle, P., & Foster, G. (1998). Using a probabilistic translation model for cross-language information retrieval. In *Sixth workshop on Very Large Corpora*.

Otterbacher, J., Erkan, G., & Radev, D. R. (2009). Biased LexRank: Passage retrieval using random walks with question-based priors. *Information Processing & Management*, *45*(1), 42-54.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Parveen, D., & Strube, M. (2015, July). Integrating Importance, Non-Redundancy and Coherence in Graph-Based Extractive Summarization. In *IJCAI* (pp. 1298-1304).

Ponte, J. M., & Croft, W. B. (1998, August). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 275-281). ACM.

Prikhod'Ko, S. M., & Skorokhod'ko, E. F. (1982). Automatic abstracting from analysis of links between phrases. *Nauchno-Tekhnicheskaya Informatsiya, Seriya*, *2*(16), 1.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

Radev, D. R., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., ... & Otterbacher, J. (2004, May). MEAD-A Platform for Multidocument Multilingual Text Summarization. In *LREC*.

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. *Computational linguistics*, *28*(4), 399-408.

Radev, D. R., Jing, H., & Budzikowska, M. (2000, April). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4* (pp. 21-30). Association for Computational Linguistics.

Radev, D. R., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., ... & Drabek, E. (2003, July). Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 375-382). Association for Computational Linguistics.

Rijsbergen, C. J. V. (1979). Information retrieval. dept. of computer science, university of glasgow. *URL: citeseer. ist. psu. edu/vanrijsbergen79information. html*, *14*.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. *Nist Special Publication Sp*, *109*, 109.

Rousseau, F., & Vazirgiannis, M. (2013, October). Graph-of-word and TW-IDF: new approach to ad hoc IR. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 59-68). ACM.

Salton, G., & McGill, M. (1983). Introduction to Modern information retrieval.

Salton, G., Allan, J., Buckley, C., & Singhal, A. (1994). Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, *264*(5164), 1421-1426.

Salton, G., Singhal, A., Buckley, C., & Mitra, M. (1996, March). Automatic text decomposition using text segments and text themes. In *Proceedings of the the seventh ACM conference on Hypertext* (pp. 53-65). ACM.

Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and

summarization. *Information Processing & Management*, *33*(2), 193-207.

Singer, H., & Donlan, D. (1985). *Reading and learning from text*. Lawrence Erlbaum Associates, Inc., Publishers, 365 Broadway, Hillsdale, NJ 07642.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, *24*(4), 35-43.

Sinha, S., Pan, R. K., Yadav, N., Vahia, M., & Mahadevan, I. (2009, August). Network analysis reveals structure indicative of syntax in the corpus of undeciphered Indus civilization inscriptions. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing* (pp. 5-13). Association for Computational Linguistics.

Skorokhod'ko, E. F. (1972). Adaptive method of automatic abstracting and indexing. *Information Processing 71*, 1179-1182.

Teufel, S., & Van Halteren, H. (2004). Evaluating information content by factoid analysis: human annotation and stability. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Tomokiyo, T., & Hurst, M. (2003, July). A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18* (pp. 33-40). Association for Computational Linguistics.

Tu, X., Huang, J. X., Luo, J., & He, T. (2016, July). Exploiting semantic coherence features for information retrieval. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 837-840). ACM.

Turney, P. D. (1997). Extraction of keyphrases from text: evaluation of four algorithms, National Research Council. *Institute for Information Technology, Technical Report ERB-1051*.

Turney, P. D. (1999). Learning to extract keyphrases from text. *arXiv preprint cs/0212013*.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information retrieval*, *2*(4), 303-336.

Turney, P. D. (2003). Coherent keyphrase extraction via web mining. *arXiv preprint cs/0308033*.

Wan, X. (2010, August). Towards a unified approach to simultaneous single-document and multi-document summarizations. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 1137-1145). Association for Computational Linguistics.

Wan, X., & Xiao, J. (2008a). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *AAAI* (Vol. 8, pp. 855-860).

Wan, X., & Xiao, J. (2008b). CollabRank: towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1* (pp. 969-976). Association for Computational Linguistics.

Wan, X., Yang, J., & Xiao, J. (2007). Towards an iterative reinforcement approach for

simultaneous document summarization and keyword extraction. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 552-559).

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, *393*(6684), 440.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999, August). KEA: Practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries* (pp. 254-255). ACM.

Yih, W. T., Goodman, J., & Carvalho, V. R. (2006, May). Finding advertising keywords on web pages. In *Proceedings of the 15th international conference on World Wide Web* (pp. 213-222). ACM.

Zha, H. (2002, August). Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 113-120). ACM.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*. 2001.

ΓΛΩΣΣΑΡΙ

(un)supervised	με/χωρίς επίβλεψη
adjacency	γειτνίασης
affinity	συγγένεια
attachment	προσάρτηση
authority	αυθεντία
average	μέσος
bushy	θαμνώδης
candidate	υποψήφιος
circuit	κύκλωμα
cluster	συστάδα
clustering	συσταδοποίηση
co-occurence	συνεμφάνιση
coherence	συνεκτικότητα
cohesion	συνοχή
collocation	συμπαράθεση
column	στήλη
component	συστατικό στοιχείου
construction	κατασκευή
corpus	σώμα κειμένων/εγγράφων
correlation	συσχέτιση
cycle	κύκλος
degenerate	εκφυλισμένος
degree	βαθμός
dimension	διάσταση
distribution	κατανομή
domination	κυριαρχία
effectiveness	αποτελεσματικότητας
eigenspace	ιδιοχώρος
eigenvalue	ιδιοτιμής
eigenvector	ιδιοδιάνυσμα
endvertex	τελικός κόμβος
feature	χαρακτηριστικό
global	καθολικός
graph	γράφος
graph order	τάξη
graph size	μέγεθος
hub	κεντρικός κόμβος
idf	αντίστροφος της συχνότητας των κειμένων
incident	εφαπτόμενη
incoporate	ενσωματώνω
index (inverted)	ευρετήριο (ανεστραμμένο)
isolated	απομονωμένος, μονήρης
keyphrase	φράση-κλειδί
Keyword	λέξη-κλειδί
length	μέγεθος
local	τοπικός
manual	χειροκίνητο
mean	μέσος
neighbor	γείτονας
normalization	κανονικοποίηση
Part-of-Speech	μέρος του λόγου
path	μονοπάτι
precision	πιστότητα
preprocessing	προεπεξεργασία
property	ιδιότητα
query	ερώτημα
rank	βαθμολογίας
ranking	βαθμολόγιση
recall	ανάκλαση
relation	σχέση, συσχέτιση
relevance	συνάφεια
response	απάντηση
row	σειρά
scalar	βαθμωτός
segmented	κατακερματισμένο
semantic	σημασιολογικό
sequence	σειρά, ακολουθία
similarity	ομοιώτητα
small-world	μικρός κόσμος
stemming	αποκοπή καταλήξεων

stopwords	**διακοπτούσες λέξεις**	**triangle**	**τρίγωνο**
tag	**ετικέτα**	**vector**	**διάνυσμα**
tf	**συχνότητα όρου**	**walk**	**περίπατος**
tokeniaztion	**στοιχειοποίηση**	**weighted**	**με βάρος, σταθμισμένος**
topic	**θεματική**	**neighborhood**	**γειτονιά**
trail	**μονοπάτι**	**vertex set**	**σύνολο κόμβων**

ΠΑΡΑΡΤΗΜΑ Α: Μέτρα Εκτίμησης

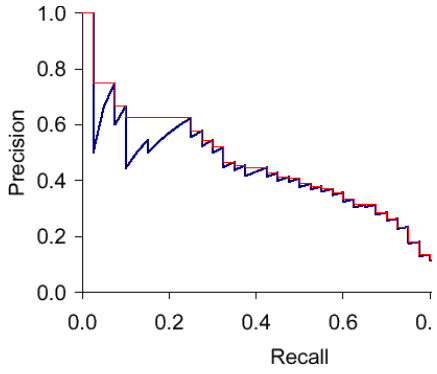
Η εκτίμηση είναι η διαδικασία κατά την οποία μετριέται η επίδοση ενός συστήματος βάσει των αποτελεσμάτων του. Κατά την εκτίμηση γίνεται σύγκριση ανάμεσα σε δύο ειδών αποτελέσματα, τα πραγματικά και τα υπολογισμένα από το σύστημα. Τα υπολογισμένα αποτελέσματα μπορούν να αναχθούν σε μια διάκριση ανάμεσα σε θετικά και αρνητικά. Τα θετικά είναι όσα στοιχεία δήλωσε το σύστημα ως έγκυρα ως προς την εφαρμογή και τα αρνητικά όσα αρνήθηκε ως άκυρα. Έτσι,τα πραγματικά αποτελέσματα (πραγματικά θετικά κι αρνητικά) συγκρίνονται με τα υπολογισμένα και προκύπτουν τέσσερεις κατηγορίες. Όταν το σύστημα εγκρίνει ένα πραγματικά θετικό αποτέλεσμα τότε αυτό σημειώνεται ως ορθά θετικό (true positive, TP) και όταν το ακυρώνει σημειώνεται ως λανθασμένα αρνητικό (false negative, FN). Αντιστοίχως, όταν ακυρώνει ένα πραγματικά αρνητικό τότε σημειώνεται ως ορθά αρνητικό (true negative, TN) και όταν το εγκρίνει ως λανθασμένα θετικό (false positive, FP). Στη δυαδική εκτίμηση αυτού του τύπου οι κατηγορίες αυτές είναι πολύ σημαντικές , αφού τα πιο κλασσικά μέτρα εκτίμησης βασίζονται σε αυτές. Αυτά είναι η Ανάκλαση (Recall) που δηλώνει το ποσοστό των ορθών υπολογισμένων θετικών σε σχέση με τα πραγματικά θετικά (1), η Πιστότητα (Precision) που δηλώνει το ποσοστό των ορθών υπολογισμένων θετικών σε σχέση με όλα τα υπολογισμένα θετικά (2) και η Ακρίβεια (Accuracy) που δηλώνει το ποσοστό των ορθών κρίσεων σε σχέση με όλα τα αποτελέσματα (3). Ένα μέτρο που συνδυάζει την πιστότητα και την ανάκλαση είναι το F-Measure, ο αρμονικός τους μέσος (4).

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$F - Measure = \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$



Σε εφαρμογές όπου τα αποτελέσματα είναι διατεταγμένα προς κάποια βαθμολόγηση η εκτίμηση πρέπει να λάμβανει υποψιν και τη διάταξη τους. Τα κλασσικά μέτρα μπορούν να χρησιμοποιηθούν και σε βαθμολογημένα αποτελέσματα. Υπολογίζονται κάθε φορά τα k υψηλότερα αποτελέσματα και αυξάνεται το k κατά ένα ανά επανάληψη. Όσο το k αυξάνεται τόσο θα αυξάνεται και η ανάκλαση. Έτσι, προκύπτουν ζευγαρωτές τιμές πιστότητας και ανάκλασης για κάθε k και μπορούν να αναπαρασταθούν με τη καμπύλη πιστότητας-ανάκλασης (precision-recall curve). Η καμπύλη έχει πολλά πριονωτά σημεία αφού για κάθε αύξηση του k αν το καινούριο αποτέλεσμα είναι λανθασμένα θετικό η πιστότητα θα μειωθεί ενώ η ανάκλαση θα παραμείνει ίδια, αν είναι ορθά θετικό τότε και τα δύο αυξάνονται. Ένας τρόπος να απαλειφθούν αυτοί οι πριονισμοί είναι ο υπολογίζεται της πιστότητα εκ παρεμβολής (interpolated precision) όπου στη καμπύλη περιλείπονται οι αυξομειώσεις και η πιστότητα να παίρνει την μέγιστη τιμή μέχρι την επόμενη σίγουρη μείωση, έτσι ώστε όταν η πιστότητα μειώνεται δεν μπορεί να

αυξηθεί πάλι (η κόκκινη γραμμή στο σχήμα Α.1). Η αντιπροσωπευτική πιστότητα (average precision) είναι ένα σημαντικό μέτρο και μπορεί να υπολογιστεί ως η αντιπροσωπευτική πιστότητα εκ παρεμβολής 11-σημείων, δηλαδή ο υπολογισμός του μέσου όρου πιστότητας για 11 τιμές της ανάκλασης (0.0 μέχρι 1.0 με βήμα 0.1 τη φορά). Όταν γίνεται εκτίμηση σε πολλές εφαρμογές του συστήματος συγχρόνως γίνεται να υπολογιστεί η μέση αντιπροσωπευτική πιστότητα (Mean Average Precision, MAP), δηλαδή ο μέσος όρος της αντιπροσωπευτικής πιστότητας όλων των εφαρμογών. Ένα άλλο, διαδεδομένο, μέτρο εκτίμησης είναι η πιστότητα στα k (P@k) στην οποία υπολογίζεται η πιστότητα των k υψηλότερων αποτελεσμάτων. Τέλος, η η δυαδική προτίμηση (Binary Preference Bpref)[5] υπολογίζει τη συχνότητα με την οποία τα πραγματικά θετικά διατάζονται χαμηλότερα από τα πραγματικά αρνητικά στα αποτελέσματα.

Σημείωση: Στην ανάκτηση πληροφοριών τα θετικά αντιπροσωπεύουν τα συναφή έγγραφα και τα αρνητικά τα μη συναφή. Στην αυτόματη περίληψη τα θετικά αντιπροσωπεύουν τις μονάδες (ngrams, προτάσεις, παράγραφοι ή ότι άλλο) που εμφανίζονται στις περιλήψεις των ειδικών και τα αρνητικά όσες δεν εμφανίζονται. Τέλος, στην εξαγωγή φράσεων-κλειδιών τα θετικά είναι οι οι συμφωνίες με τις φράσεις-κλειδιά των ειδικών και τα αρνητικά οι ασυμφωνίες.

Για τα μέτρα εκτιμήσεως του εργαλείου Rouge βλ. Lin (2004)[6].

5 Buckley, C., & Voorhees, E. M. (2004, July). Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 25-32). ACM.

6 Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

ΠΑΡΑΡΤΗΜΑ Β: Μοντέλα Βαθμολόγησης Όρων (Πίνακας)

Οι παρακάτω υπολογισμοί δίνονται επειδή αυτά τα μοντέλα αναφέρονται στο κύριο σώμα της εργασίας και δεν αναπτύχθηκαν καθόλου. Δίνεται συγκεκριμένα η βαθμολόγηση όρου εγγράφου αφού έχει χρησιμότητα ανεξαρτήτου εφαρμογής.

Μοντέλο	Βαθμολόγηση όρου Εγγράφου
Tf-idf (Jones, 1972)	$c(t,D)\times\frac{N}{df(t)}$
Okapi BM25 (Robertson et al., 1995)	$\frac{(k_1+1)c(t,D)}{k1(1-b+b\times\lvert D\rvert/avdl)+c(t,D)}\times\log(\frac{N+1}{df(t)})$
Dirichlet Prior (Zhai & Lafferty, 2001)	$\log(\frac{\mu}{\lvert D\rvert+\mu}+\frac{c(t,D)}{(\lvert D\rvert+\mu)p(t\lvert C)})$
Pivoted Normalization (Singhal, 2001)	$\left\{\begin{array}{l}\frac{1+\log(1+\log(c(t,D)))}{1-s+s\times\lvert D\rvert/avdl}\times\log\frac{N+1}{df(t)},\ \ \alpha v\ \ c(t,D)>0\\ 0,\alpha\lambda\lambda\iota\acute{\omega}\varsigma\end{array}\right\}$

Πίνακας Β.1: Όπως εμφανίζονται στο Lv, Zhai (2011)

c(t, D)	Συχνότητα εμφάνισης όρου t στο έγγραφο D
N	Αριθμός εγγράφων στη συλλογή
df(t)	Αριθμός εγγράφων που περιλαμβάνουν τον όρο t
|D|	Μέγεθος εγγράφου D
avdl	Μέσο μέγεθος εγγράφων στη συλλογή
c(t, C)	Συχνότητα εμφάνισης όρου t στη συλλογή C
p(t|C)	Πιθανότητα του όρου t δοσμένης της συλλογής C (Zhai & Lafferty, 2001)

Πίνακας Β.2: Η σημειογραφία δανεισμένη από το Lv, Zhai (2011)

♠ Τέλος ♠