ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΚΑΙ ΔΙΚΤΥΩΝ

Διερεύνηση της Εφαρμογής των Γενετικών Αλγορίθμων στο Προσαρμοστικό Φιλτράρισμα της Πληροφορίας

Φοιτητής

Στέφανος Κοντοβάς

Επιβλέποντες Καθηγητές:

Βάβαλης Εμμανουήλ

Νανάς Νικόλαος

1 Μαρτίου 2010

Περίληψη

Η παρούσα διπλωματική εργασία, όπως προδίδει ο τίτλος της, διερευνά την εφαρμογή των εξελικτικών αλγορίθμων στον τομέα του Προσαρμοστικού Φιλτραρίσματος της Πληροφορίας (ΠΦΠ). Αφού παρουσιαστούν οι βασικές αρχές του Φιλτραρίσματος της Πληροφορίας και των Εξελικτικών Αλγορίθμων, περιγράφονται οι ιδιαιτερότητες του ΠΦΠ, που πηγάζουν από τη δυναμική και πολύπλοκη φύση του. Εξετάζονται γνωστές εξελικτικές προσεγγίσεις στο πρόβλημα της προσαρμογής προφίλ και συγκρίνονται τα αποτελέσματα τους με αυτά του δημοφιλούς αλγόριθμου μάθησης Rocchio.

Ευχαριστίες

Κατ' αρχήν θα ήθελα να ευχαριστήσω του καθηγητές μου κ.κ Βάβαλη Εμμανουήλ και Νανά Νικόλαο οι οποίοι πρότειναν το θέμα της διπλωματικής μου εργασίας και προσέφεραν πολύτιμη βοήθεια και καθοδήγηση όλους αυτούς τους έξι μήνες, μέχρι την ολοκλήρωσή της.

Οφείλω, επίσης, ένα «ευχαριστώ» στο Γιάννη, ο οποίος συνέβαλε στην εκτέλεση των πειραμάτων διαχειριζόμενος το cluster του πανεπιστημίου, τον Λευτέρη για τις χρήσιμες συμβουλές του και, ιδιαιτέρως, το Δημήτρη διότι χωρίς τη δική του προτροπή και πρωτοβουλία αυτή η εργασία δεν θα είχε πραγματοποιηθεί.

Τέλος, ευχαριστώ τους γονείς μου και την Κωνσταντίνα για την υπομονή τους και τη στήριξη που μου προσέφεραν.

Περιεχόμενα

. . . στους γονείς μου

Κεφάλαιο 1

Εισαγωγή

Στην εποχή μας, με την πληθώρα ψηφιακών μέσων, την ανάπτυξη της δικτύωσης και κυρίως του Διαδικτύου καθώς και της τεχνολογίας υπολογιστών, οι πληροφορίες που είναι διαθέσιμες στον χρήστη έχουν αυξηθεί δραματικά. Οι πληροφορίες κυριολεκτικά κατακλύζουν τον Ιστό. Σ' αυτό έχουν συμβάλει ο μεγάλος αριθμός ιστότοπων, συμπεριλαμβανομένων και blogs, οι συνδρομές τύπου RSS [1] και τα εκατομμύρια των e-mails. Μεγάλη συμμετοχή στην πλημμυρίδα των πληροφοριών έχουν και τα τελευταία χρόνια τα κοινωνικά δίκτυα όπου κάθε χρήστης τους αποτελεί δέκτη αλλά και πομπό μεγάλου όγκου πληροφοριών. Ο καθε χρήστης έχει πια ενεργή συμμετοχή στο Διαδίκτυο που του επιτρέπει να δημιουργεί και να διαμοιράζει πληθώρα πληροφοριών. Όλα τα παραπάνω δημιουργούν το φαινόμενο της υπερπληροφόρησης. Συσσωρεύεται, δηλαδή, ένας τεράστιος όγκος πληροφοριών με αποδέκτη το χρήστη από τον οποίο ένα μικρό κομμάτι του τον ενδιαφέρει πραγματικά. Γεννάται, επομένως, η ανάγκη για διαχωρισμό της χρήσιμης - ενδιαφέρουσας πληροφορίας από το όλο σύνολο των διαθέσιμων πληροφοριών. Χρειάζεται ένας τρόπος αποδοτικού φιλτραρίσματος πληροφοριών και δεδομένων. Και αυτό δεν αφορά μόνο το χρήστη - δέκτη αλλά και το χρήστη - πομπό πληροφοριών ο οποίος θέλει στοχευμένα να μεταδώσει πληροφορίες σε ομάδες ατόμων με συγκεκριμένα χαρακτηριστικά και ενδιαφέροντα.

[1] Really Simple Syndication: Είναι ένα format ανταλλαγής περιεχομένου βασισμένο σε γλώσσα XML. Ένας νέος τρόπος να ενημερώνεται ο χρήστης του Ίντερνετ για γεγονότα και νέα από άλλους χρήστες ή και κανάλια πληροφορίας.

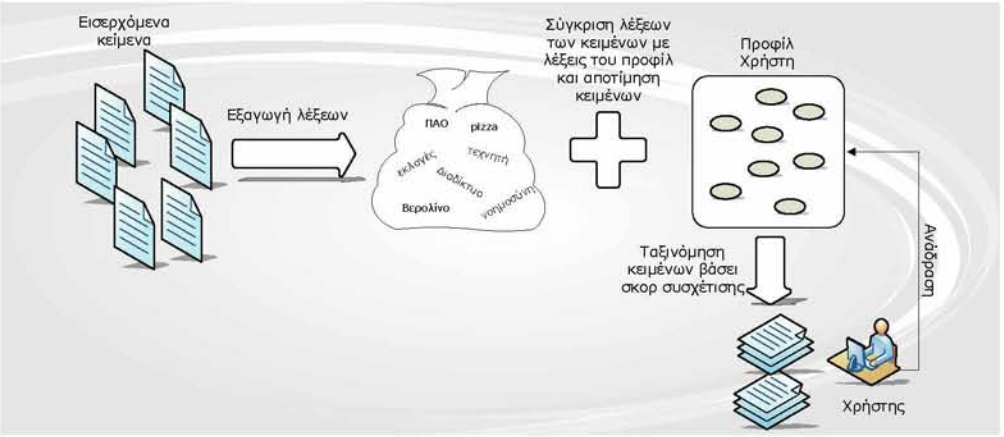
1.1 Φιλτράρισμα της Πληροφορίας (ΦΠ)

Τη λύση σ' αυτό το πρόβλημα προσπαθεί να δώσει ο ερευνητικός τομέας του Φιλτραρίσματος της Πληροφορίας(Information Filteirng). Το ΦΠ βασίζεται στην αναπαράσταση των ενδιαφερόντων του χρήστη υπό τη μορφή του προφίλ (profile). Κάθε πληροφορία συγκρίνεται με το προφίλ του χρήστη. Αν κριθεί «σχετική» τότε παρουσιάζεται στον χρήστη ειδάλλως αγνοείται. Ξεχωρίζουν κυρίως δύο κατηγορίες Φιλτραρίσματος της Πληροφορίας: *collaborative* και *content-based.* Διαφέρουν στον τρόπο με τον οποίο η πληροφορία και το προφίλ του χρήστη αναπαρίστανται και συγκρίνονται.

Collaborative: Η πληροφορία χαρακτηρίζεται από το πώς την έχουν αξιολογήσει τα μέλη μιας κοινότητας. Σκοπός είναι να προταθούν στο χρήστη νέες πληροφορίες που πιθανόν τον ενδιαφέρουν. Αυτό επιτυγχάνεται με τη εύρεση ενός συνόλου χρηστών ("γείτονες") που μοιράζονται τα ίδια ενδιαφέροντα με τον εν λόγω χρήστη. Από τη στιγμή που θα οριστεί αυτό το σύνολο επιλέγονται να παρουσιαστούν στο χρήστη οι πληροφορίες οι οποίες έχουν αξιολογηθεί υψηλά από τα μέλη του παραπάνω συνόλου. Με τον τρόπο αυτό προτείνονται στο χρήστη, βάσει των αξιολογήσεων των "γειτόνων" του, αντικείμενα πληροφορίας τα οποία ο ίδιος δεν είχε προσπελάσει στο παρελθόν ή δε γνώριζε καν την υπαρξή τους. Το Collaborative Filtering δεν προϋποθέτει πρόσβαση σε αυτό καθ' αυτό το περιεχόμενο της πληροφορίας και, συνεπώς, δεν περιορίζεται στην "κειμενική" (textual) πληροφορία. Ένα απλό παράδειγμα εφαρμογής του coollaborative filtering είναι η εισήγηση στο χρήστη, ταινιών, βιβλίων ή μουσικής βάσει των προτιμήσεων των "φίλων" του [6].

Content-based: Η πληροφορία και το προφίλ αναπαρίστανται χρησιμοποιώντας στοιχεία από το περιεχόμενο της πληροφορίας, Λόγω της φύσης του χρησιμοποιείται, κυρίως, σε φιλτράρισμα κειμένων και εγγράφων, από τα οποία μπορούμε να εξάγουμε διακριτά στοιχεία – γνωρίσματα (features). Χρησιμοποιείται ένας κοινός διανυσματικός χώρος (vector space model) όπου τα προφίλ και τα κείμενα, που αναπαρίστανται ως δυαδικά ή σταθμισμένα διανύσματα, προβάλλονται στις ίδιες διαστάσεις. Αυτό

επιτρέπει στη συνέχεια τη χρήση τριγωνομετρικών μεθόδων για την σύγκριση της πληροφορίας με το προφίλ του χρήστη (Σχήμα 1.1).



Σχημα 1.1: Προσαρμοστικό Content-Based Φιλτράρισμα εγγράφων

Πιο συγκεκριμένα, στην περίπτωση της "κειμενικής" πληροφορίας, που θα αποτελέσει αντικείμενο της παρούσης εργασίας, τόσο το προφίλ όσο και το πληροφοριακό αντικείμενο (έγγραφο) αναπαρίστανται ως σταθμισμένα διανύσματα. Οι διαστάσεις του χώρου ορίζονται από τον αριθμό των *μοναδικών* λέξεων που εξάγονται από το περιεχόμενο των κεμένων. Το διάνυσμα του προφίλ αποτελείται από τις λέξεις που χαρακτηρίζουν τα ενδιαφέροντα του χρήστη, συνοδευόμενες από βάρη που δηλώνουν τη σημασία της κάθε λέξης στο προφίλ. Για τη σύγκριση των δύο διανυσματων (προφίλ και κειμένου) οι πιο δημοφιλείς μέθοδοι είναι το *εσωτερικό γινόμενο* :

$$\langle P, D \rangle = \sum_{i=1}^{N} p_i \cdot d_i \tag{1.1}$$

και το *συνημίτονο* της μεταξύ τους γωνίας :

$$\cos(\theta) = \frac{P \cdot D}{\|P\| \|D\|} \tag{1.2}$$

όπου:

$P = \{p_1, p_2, \ldots p_n\}$ είναι το διάνυσμα με τα βάρη των λέξεων του προφίλ,

$D = \{d_1, d_2, \ldots d_n\}$ είναι το διάνυσμα με τα βάρη των λέξεων του κειμένου

Χρησιμοποιώντας κάποια από τις παραπάνω μεθόδους υπολογίζεται μία τιμή (σκορ συσχέτισης) για κάθε κείμενο. Με βάση αυτό το σκορ τα κείμενα ταξινομούνται και παρουσιάζονται, με φθίνουσα σειρά, στο χρήστη. Στη συνέχεια, λαμβάνοντας υπ' όψιν την ανάδραση (feedback) του χρήστη πραγματοποιείται προσαρμογή του προφίλ στα τρέχοντα ενδιαφέροντά του (Σχήμα 1.1). Η ανάδραση μπορεί να είναι είτε *άμεση* π.χ δηλώνει αν τον ενδιαφέρει το κείμενο ή όχι, είτε *έμμεση* π.χ υποδηλώνεται από τη διάρκεια ανάγνωσης του κειμένου.

Στην παρούσα εργασία θα επικεντρωθούμε σε αλγορίθμους content - based Φιλτραρίσματος καθώς είναι γεγονός πως δεν υπάρχει μέχρι σήμερα κάποια καθιερωμένη λύση ούτε επιτυχημένες εφαρμογές για το πρόβλημα αυτό [7].

1.2 Ορισμός Προβλήματος

Αιτία αυτής της αδυναμίας εύρεσης αποδοτικής λύσης, των μέχρι σήμερα εφαρμοσμένων αλγορίθμων, στο πρόβλημα της προσαρμογής προφίλ, είναι ο πολύπλοκος και δυναμικός χαρακτήρας του προβλήματος. Οι περισσότεροι προσαρμοστικοί αλγόριθμοι τα καταφέρνουν καλά όταν πρόκειται για προσαρμογή προφίλ σ' ένα μόνο ενδιαφέρον του χρήστη ή σε περισσότερα αλλά εκ των προτέρων δηλωμένα. Τι γίνεται όμως όταν τα ενδιαφέροντα του χρήστη όχι μόνο είναι περισσότερα του ενός αλλά πληθαίνουν και αλλάζουν με το χρόνο; Π.χ ο χρήστης παύει να ενδιαφέρεται πια για ποδόσφαιρο αλλά ενδιαφέρεται τώρα για μπάσκετ και αργότερα προστίθενται στα ενδιαφέροντα του η φωτογραφία και οι επερχόμενες εκλογές στη χώρα του.

Σε αυτή την περίπτωση το πρόβλημα παύει να αποτελεί απλώς μια αναζήτηση βέλτιστου. Γίνεται δυναμικό και ομοιάζει με πρόβλημα Multimodal Dynamic βελτιστοποίησης [5]. Ο καλύτερος τρόπος να το αντιμετωπίσει κανείς είναι σαν ένα πρόβλημα συνεχούς μάθησης. Ένα πρόβλημα στο οποίο ο αλγόριθμος θα πρέπει να παρακολουθεί συνεχώς τα ενδιαφέροντα του χρήστη και να προσαρμόζεται σε κάθε εναλλαγή, πρόσθεση ή αφαίρεση ενδιαφερόντων του.

Αυτή η δυναμική φύση του προβλήματος ελκύει τη χρήση είτε αλγόριθμων μάθησης είτε αλγορίθμων εμπνευσμένων από τη βιολογία, που αναφέρονται στις επόμενες υποενότητες.

1.2.1 Αλγόριθμοι Μάθησης

Οι αλγόριθμοι μάθησης χρησιμοποιούν μια συνάρτηση προσαρμογής των βαρών. Η συνάρτηση περιέχει συνήθως κάποιους σταθερούς συντελεστές οι οποίοι προκύπτουν είτε από δεδομένα του χρήστη είτε από προηγούμενες πειραματικές εκτελέσεις και καθορίζουν το ρυθμό προσαρμογής του προφίλ. Η συνάρτηση αυτή περιορίζεται στην προσαρμογή προφίλ με συγκεκριμένα χαρακτηριστικά.

Από τους πιο διαδεδομένους αλγόριθμους μάθησης είναι ο αλγόριθμος του Rocchio. Αρχικά εφαρμόστηκε για τη βελτίωση της επερώτησης (query) με βάση τα ήδη υπάρχοντα αποτελέσματα μιας αναζήτησης, αλλά με μερικές μετατροπές εφαρμόζεται και στο Προσαρμοστικό ΦΠ.

Ο αλγόριθμος του Rocchio προϋποθέτει ότι το προφίλ του χρήστη αλλά και το κείμενο που επεξεργάζεται αναπαρίστανται ως σταθμισμένα διανύσματα σε ένα κοινό διανυσματικό χώρο (vector space), διαστάσεων ίσων με τον αριθμό των μοναδικών λέξεων που περιέχονται στα κείμενα. Ο αλγόριθμος προσπαθεί να «μετακινήσει» γραμμικά το διάνυσμα του προφίλ έτσι ώστε να πλησιάσει σε αυτό του εγγράφου που χαρακτηρίζεται ως σχετικό με τα ενδιαφέροντα του χρήστη και να απομακρυνθεί από τα μη - σχετικά.

$$P_{t+1} = \begin{cases} \alpha \cdot P_t + \beta \cdot D & \text{αν D σχετικό,} \\ \alpha \cdot P_t - \gamma \cdot D & \text{αν D μη σχετικό} \end{cases} \tag{1.3}$$

όπου:

P_{t+1}	είναι το νέο προφίλ
P_t	είναι το προηγούμενο προφίλ
D	είναι το διάνυσμα του εξεταζόμενου κειμένου

Οι συντελεστές α, β, γ ορίζουν τη συνεισφορά του τρέχοντος προφίλ και του σχετικού(ή μη σχετικού κειμένου αντίστοιχα) στα βάρη του νέου προφίλ. Ουσιαστικά καθορίζουν την ταχύτητα σύγκλισης των βαρών. Πιο συγκεκριμένα, ο συντελεστής α αποτελεί ένα δείκτη απόσβεσης μνήμης. Καθορίζει, δηλαδή, πόσο γρήγορα «ξεχνά» το προφίλ την προηγούμενη του κατάσταση. Μπορεί να θεωρηθεί σαν δείκτης εξασθένησης (decay) ο οποίος μειώνει αναλογικά το βάρος μιας λέξης με το χρόνο, ασυμπτωτικά με το 0 (μηδέν). Μεγαλύτερη τιμή του α σημαίνει και μικρότερη εξασθένηση και αντιστρόφως. Οι συντελεστές β, γ προσδιορίζουν το ποσοστό του βάρους των λέξεων του σχετικού κειμένου που προστίθεται/αφαιρείται αντίστοιχα στις/από τις αντίστοιχες του προφίλ. Να σημειωθεί πως ο αλγόριθμος του Rocchio δεν παρέχει κάποιο μηχανισμό αφαίρεσης λέξεων από το προφίλ επειδή προϋποθέτει διανυσματικό χώρο με προκαθορισμένες διαστάσεις. Όπως επίσης λόγω της απουσίας κανονικοποίησης (normalization) εάν ο συντελεστής β είναι μεγαλύτερος από τον α τότε τα βάρη των λέξεων του προφίλ θα αυξάνονται συνεχώς.

Η Ενισχυτική Μάθηση (Reinforcement Learning) αποτελεί μία παραλλαγή των αλγόριθμων μάθησης. Η διαφοροποίηση της αφορά στους συντελεστές μάθησης, οι οποίοι δεν είναι πια σταθεροί. Οι τιμές των συντελεστών μειώνονται καθώς αυξάνεται ο αριθμός των σχετικών εγγράφων που έχουν επεξεργαστεί, ούτως ώστε ο αλγόριθμος που, μετά από κάποια βήματα, έχει "μάθει" το θέμα ενδιαφέροντος να μπορεί να διατηρηθεί σε αυτή την κατάσταση.

Και σ' αυτή την περίπτωση όμως, παρ 'όλο που υπάρχει δυνατότητα μεταβολής των συντελεστών μάθησης, όταν ο αλγόριθμος έχει "μάθει" ένα συγκεκριμένο θέμα, τείνει να διατηρηθεί σ' αυτή την κατάσταση και είναι δύσκολου να προσαρμοστεί σε ένα νέο θέμα ενδιαφέροντος [10].

Γενικότερα οι Αλγόριθμοι μάθησης δεν είναι αποτελεσματικοί όταν παρουσιάζονται εναλλαγές στα θέματα ενδιαφέροντος [14]. Μία πιθανή εξήγηση είναι ότι αυτοί οι αλγόριθμοι επινοήθηκαν για να επιλύσουν προβλήματα βελτιστοποίησης που αφορούν ένα μόνο βέλτιστο σημείο – θέμα και δεν έχουν κάποιο μηχανισμό εύρεσης νέας περιοχής ενδιαφέροντος ή απαλοιφής κάποιας προηγούμενης μειωμένου ενδιαφέροντος [7].

1.2.2 Αλγόριθμοι εμπνευσμένοι από τη Βιολογία

Τα τελευταία χρόνια υπάρχει ένα αυξανόμενο ενδιαφέρον στην εφαρμογή εμπνευσμένων από τη βιολογία (biologically inspired) προσεγγίσεων στον τομέα του προσαρμοστικού φιλτραρίσματος. Οι προσεγγίσεις αυτές αποτελούν προσομοιώσεις της συμπεριφοράς των φυσικών βιολογικών συστημάτων με σκοπό την επίλυση πολύπλοκων υπολογιστικών προβλημάτων.

Ο τομέας των biologically inspired αλγορίθμων περιλαμβάνει:

- **Νευρωνικά Δίκτυα**: Εμπνευσμένα από τη λειτουργία του εγκεφάλου (νευρώνες, συνάψεις), χρησιμοποιούνται για την αναγνώριση μοτίβων.
- **Αποικίες Μυρμηγκιών**: Βασίζονται στη συμπεριφορά των "κοινωνικών" εντόμων και εφαρμόζονται στην αναζήτηση αποδοτικών και σύντομων διαδρομών σε γράφους.
- **Τεχνητό Ανοσοποιητικό Σύστημα**: Είναι προσαρμοστικά συστήματα τα οποία εκμεταλλεύονται τα χαρακτηριστικά του ανοσοποιητικού συστήματος, τη μάθηση και τη μνήμη, για να λύσουν ένα υπολογιστικό πρόβλημα. Βρίσκουν εφαρμογές στον τομέα της Τεχνητής Νοημοσύνης και πιο συγκεκριμένα στη Μηχανική Μάθηση αλλά και στο ΠΦΠ.
- **Εξελικτικοί Αλγόριθμοι**: Εμπνευσμένοι από τη θεωρία της εξέλιξης των ειδών, χρησιμοποιούνται σε προβλήματα βελτιστοποίησης.

Στην εργασία αυτή θα επικεντρωθούμε στου Γεντικούς Αλγόρθμους (ΓΑ) που υ-πάγονται στην κατηγορία των Εξελικτικών Αλγορίθμων (ΕΑ). Η επόμενη ενότητα περιγράφει τις βασικές αρχές οι οποίες διέπουν τους ΕΑ.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Φυσική Επιλογή

Η *Φυσική Επιλογή* είναι η διαδικασία εξέλιξης των ειδών μέσω της οποίας οι οργανισμοί που είναι καλύτερα προσαρμοσμένοι στο περιβάλλον αφήνουν περισσότερους απογόνους από εκείνους που είναι λιγότερο προσαρμοσμένοι. Η θεωρία της φυσικής επιλογής διατυπώθηκε επίσημα το 1858, από τους Κάρολο Δαρβίνο και τον Alfred Russel Wallace, που πραγματοποιούσαν εκείνη την περίοδο ανεξάρτητες μεταξύ τους έρευνες. Στηρίζεται στην παρατήρηση πως ορισμένες διαφορές μεταξύ των ατόμων σε έναν πληθυσμό είναι κληρονομήσιμες.

Αν λοιπόν ένα κληρονομήσιμο γνώρισμα προσφέρει προσαρμοστικό πλεονέκτημα στο φορέα του, ο οργανισμός αυτός (είτε γιατί επιβιώνει περισσότερο, είτε γιατί επιλέγεται περισσότερο από τα άτομα του άλλου φύλου, σε σύγκριση με όσους δεν το φέρουν) αφήνει περισσότερους απογόνους με αποτέλεσμα να το μεταβιβάζει με αυξημένη συχνότητα στα άτομα της επόμενης γενιάς. Με τον τρόπο αυτό συσσωρεύονται από γενιά σε γενιά τα ευνοικά για την επιβίωση γνωρίσματα, κάτι που μπορεί να οδηγήσει βαθμιαία στη δημιουργία ενός νέου είδους. Με τη διαδικασία αυτή, που είναι απολύτως συμβατή με τα επιστημονικά δεδομένα έχει επιτευχθεί η εξέλιξη των

σύνθετων οργανισμών από απλούστερους. Η θεωρία της Φυσικής επιλογής δεν απέδειξε απλώς την ύπαρξη της εξέλιξης, αλλά υπέδειξε και ένα πειστικό μηχανισμό για το πώς έχει συμβεί η διαδικασία της εξέλιξης[1].

2.2 Νεο-Δαρβινισμός

Στο *νεο-Δαρβινισμό* η εξέλιξη περιγράφεται με τους όρους **ποικιλότητα**, **κληρονομικότητα** και **φυσική επιλογή** [3]. Η ποικιλότητα αναφέρεται στις διαφορές μεταξύ των οντοτήτων και υπονοεί ότι η διαδικασία της εξέλιξης εφαρμόζεται σε έναν πληθυσμό από οργανισμούς. Η κληρονομικότητα είναι το στοιχείο που διαφοροποιεί τον νεο-Δαρβινισμό από τον "κλασσικό" Δαρβινισμό[2]. Λίγο αργότερα από το Δαρβίνο(1864), ο Gregor Mendel, ένας Αυστριακός μοναχός και βοτανολόγος, παρατήρησε ότι τα χαρακτηριστικά των ειδών δεν συγχωνεύονται αλλά κληρονομούνται από γενιά σε γενιά. Τη θεωρία του Mendel ήρθε να επιβεβαιώσει πολύ αργότερα(1953) η ανάπτυξη της μοριακής βιολογίας και η ανακάλυψη της ύπαρξης των γονιδίων στο DNA.

Η θεωρία του νεο-Δαρβινισμού υιοθετήθηκε, επίσης, για την επεξήγηση της πολιτισμικής εξέλιξης. Ο Richard Dawkins εισήγαγε(1976) την έννοια του *meme*[3] σαν ένα τμήμα πληροφορίας που αναπαραγάγει τον εαυτό του καθώς οι άνθρωποι ανταλλάσσουν ιδέες [1]. Η ουσιαστική διαφορά μεταξύ βιολογικής και πολιτισμικής εξέλιξης είναι ότι ενώ τα γονίδια μεταφέρονται αναλλοίωτα, τα memes, αντιθέτως, διασκευάζονται από τα γνωστικά χαρακτηριστικά του ατόμου που τα μεταδίδει. Σε αντίθεση με τη θεωρία της βιολογικής εξέλιξης, η θεωρία του μιμητισμού υποστηρίζει ότι η πολιτισμική εξέλιξη μπορεί αξιοποιεί επίκτητες βελτιώσεις των memes με στόχο την επιτάχυνση της εξελικτικής διαδικασίας.

Στην επόμενη ενότητα αναφέρεται ο τρόπος με τον οποίο όλες οι παραπάνω θεωρίες μεταφράζονται σε αλγόριθμους για την επίλυση υπολογιστικών προβλημάτων και ιδιαιτέρως για προβλήματα σχετιζόμενα με το Φιλτράρισμα της Πληροφορίας.

[1] http://el.wikipedia.org/wiki/Φυσική_Επιλογή

[2] Ο νεο-Δαρβινισμός αρνείται τον επίκτητο χαρακτήρα στην κληρονομικότητα

[3] Η ετυμολογία της λέξης προέρχεται από την ελληνική λέξη *μιμητισμός*

Κεφάλαιο 3

Γενετικοί και Μιμητικοί Αλγόριθμοι

Στόχος των γενετικών αλγορίθμων (ΓΑ) είναι οι χρήση των θεμελιωδών αρχών της εξέλιξης των ειδών για την εύρεση λύσεων σε υπολογιστικά προβλήματα. Ανήκουν στο κλάδο της επιστήμης υπολογιστών και αποτελούν τεχνικές στοχαστικής αναζήτησης σε προβλήματα βελτιστοποίησης.

Ένας ΓΑ χρησιμοποιεί έναν *πληθυσμό* τυχαίων λύσεων. Κάθε οντότητα στον πληθυσμό καλείται *χρωμόσωμα* και αναπαριστάται, συνήθως, ως ένα σύνολο συμβόλων πεπερασμένου αλφάβητου. Ο πληθυσμός εξελίσσεται μέσω επαναληπτικών βημάτων που αποκαλούνται *γενιές*. Σε κάθε γενιά καθένα από τα χρωμοσώματα αποτιμάται, χρησιμοποιώντας μια *συνάρτηση καταλληλότητας*, η οποία υπολογίζει πόσο καλή είναι η λύση που αναπαριστά το χρωμόσωμα. Τα χρωμοσώματα επιλέγονται για αναπαραγωγή σύμφωνα με την καταλληλότητα τους. Δύο χρωμοσώματα ζευγαρώνουν χρησιμοποιώντας τη μέθοδο της *διασταύρωσης* (crossover), η οποία συνδυάζει τμήματα και από τους δύο γονείς με σκοπό να παράγει καλύτερους απογόνους που θα αντικαταστήσουν τα πιο αδύναμα μέλη του πληθυσμού. Σε πολλές περιπτώσεις για να διευρυνθεί το πεδίο αναζήτησης και να εξερευνηθεί κάποιο άλλο πεδίο λύσεων τα

χρωμοσώματα υπόκεινται σε μικρές τροποποιήσεις μέσω του μηχανισμού των *μεταλλάξεων* [6].

Εμπνευσμένος και από τις αρχές του Δαρβίνου για την εξέλιξη των ειδών αλλά και από την έννοια του meme που διατυπώθηκε από τον Dawkins, ο όρος *Μιμητικός Αλγόριθμος* (Μ.Α) επινοήθηκε για πρώτη φορά από τον Pablo Moscato [4] το 1989 για να περιγράψει ένα Γ.Α που ενσωμάτωνε τοπική αναζήτηση. Σε αντίθεση με τους Γ.Α όπου το χρωμόσωμα διατηρείται αναλλοίωτο στη διάρκεια του κάθε κύκλου ζωής του, οι Μ.Α παρέχουν στο χρωμόσωμα τη δυνατότητα να "μάθει" με στόχο να επιτύχουν βελτιωμένες λύσεις. Ουσιαστικά, ο Μ.Α αποτελεί ένα υβριδικό Γ.Α σε συνδυασμό με μία ξεχωριστή συνάρτηση μάθησης που του επιτρέπει να πραγματοποιεί τοπικές βελτιώσεις στα χρωμοσώματα, οι οποίες κληρονομούνται στους απογόνους.

Κεφάλαιο 4

Εξελικτικό Φιλτράρισμα Πληροφορίας

Η δυναμική φύση του ΠΦΠ προσέλκυσε την εφαρμογή εξελικτικών μεθόδων συνθέτοντας το *Εξελικτικό Φιλτράρισμα Πληροφορίας*[1] (ΕΦΠ). Οι ΓΑ λόγω των ιδιοτήτων τους αποτελούν σημαντικό εργαλείο για την υλοποίηση του ΕΦΠ.

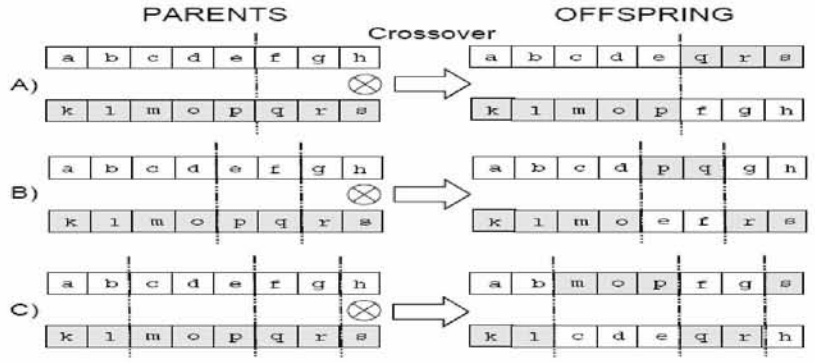
Το ΕΦΠ ανήκει κυρίως στη κατηγορία content-based φιλτραρίσματος (βλ. Ενότητα 1.1) και χρησιμοποιεί ΓΑ ή ΜΑ για την προσαρμογή του προφίλ σε αλλαγές των ενδιαφερόντων του χρήστη. Η φιλοσοφία του ΕΦΠ περιλαμβάνει ένα πληθυσμό από προφίλ τα οποία αναπαριστούν τα ενδιαφέροντα του χρήστη και εξελίσσονται βασιζόμενα στην ανάδραση του χρήστη. Ακολουθώντας την ανάδραση του χρήστη ο πληθυσμός των προφίλ μετακινείται σε περιοχές που ενδιαφέρουν τον χρήστη και απομακρύνονται από τις μη - ενδιαφερόμενες. Τα προφίλ που είναι πιο κοντά σε αυτές τις περιοχές θεωρούνται καλύτερα και τους αποδίδονται μεγαλύτερες τιμές από τη *συνάρτηση καταλληλότητας* (Σ.κ). Σαν συνάρτηση καταλληλότητας μπορεί να χρησιμοποιηθεί ο μέσος όρος ή το άθροισμα των σκορ συσχέτισης για τα σχετικά κείμενα που έχουν αποτιμηθεί ως εκείνη τη στιγμή [2, 11] καθώς και η αναλογία των επιτυχών αποτιμήσεων σχετικών κειμένων [12, 13].

[1] Evolutionary Information Filtering

Στην περίπτωση των ΜΑ γίνεται χρήση συναρτήσεων μάθησης για την βελτίωση των προφίλ που αποτίμησαν υψηλότερα κάποιο σχετικό κείμενο.

Ένα ποσοστό από τα "καταλληλότερα" προφίλ επιλέγονται για αναπαραγωγή. Το ποσοστό αυτό μπορεί να είναι είτε σταθερό είτε μεταβαλλόμενο με το πέρασμα των γενεών. Για πιο πιστή αναπαράσταση της διαδικασίας της φυσικής επιλογής μπορεί να χρησιμοποιηθεί ο μηχανισμός της ρουλέτας (roulette wheel selection.

Τα ζεύγη των προφίλ που έχουν επιλεχθεί αναπαράγονται χρησιμοποιώντας τη μέθοδο της διασταύρωσης. Ανάλογα με τον αριθμό των σημείων που επιλέγονται για την κατάτμηση του διανύσματος του προφίλ χωρίζονται κυρίως σε τρεις κατηγορίες: α) ενός, β) δύο και γ) τριών σημείων (Σχήμα 4.1). Περισσότερα σημεία διασταυρώσεων καθιστούν πιο τυχαίο το αποτέλεσμα αλλά μπορούν να καταστρέψουν σημαντικούς συνδυασμούς λέξεων.



Σχημα 4.1: Τεχνικές Διασταύρωσης: *a*) ενός, *b*) δύο και *c*) τριών σημείων

Οι απόγονοι είναι δυνατόν να υποστούν μεταλλάξεις κατά τις οποίες μία τυχαία λέξη μπορεί να αντικατασταθεί από μία άλλη ενός διαφορετικού προφίλ. Μία άλλη επιλογή είναι η μεταβολή του βάρους μίας λέξης με τυχαίο ή προκαθορισμένο τρόπο.

Οι απόγονοι αντικαθιστούν τα χειρότερα, βάσει της ΣΚ, προφίλ διατηρώντας πάντα τον πληθυσμό σταθερό.

Κεφάλαιο 5

Θεωρητικά ζητήματα

Οι ΓΑ και οι ΜΑ βρίσκουν επιτυχημένες εφαρμογές σε προβλήματα βελτιστοποίησης ειδικά όταν πρόκειται για την εύρεση ενός μόνο βέλτιστου σημείου. Το πρόβλημα του ΠΦΠ δεν ανήκει σε αυτή την κατηγορία. Ανήκει περισσότερο στην κατηγορία Multimodal Dynamic Optimization (MDO) όπου αναζητείται η εύρεση πολλών βέλτιστων σημείων τα οποία δεν είναι σταθερά και μεταβάλλονται με το χρόνο. Οι ΓΑ είναι δύσκολο να αποδώσουν σ' ένα MDO πρόβλημα καθώς έχουν την τάση να συγκλίνουν σε ένα μόνο βέλτιστο σημείο και παρουσιάζουν έλλειψη ποικιλομορφίας (diversity). Για την αντιμετώπιση αυτής της έλλειψης εφαρμόζονται ποικίλες τεχνικές μερικές από τις οποίες είναι:

- Η αύξηση του ρυθμού μετάλλαξης.
- Τυχαία τοποθέτηση νέων χρωμοσωμάτων στον πληθυσμό.
- Ενσωμάτωση μνήμης στους Γ.Α όπου τα καλύτερα χρωμοσώματα από προηγούμενες γενιές αποθηκεύονται και επαναχρησιμοποιούνται όταν υπάρξουν αλλαγές στο περιβάλλον.
- Χρησιμοποίηση υποσυνόλων του πληθυσμού για την (ανα)κάλυψη περισσότερων περιοχών λύσεων.

Οι παραπάνω λύσεις, όμως, είναι επιπρόσθετες τεχνικές και δεν αποτελούν βιολογικές διαδικασίες των ΓΑ. Η έλλειψη ποικιλομορφίας είναι ένα εγγενές πρόβλημα των ΓΑ.

Προκύπτει από τη διαδικασία της επιλογής των χρωμοσωμάτων για αναπαραγωγή, με βάση τα σχετικά τους βάρη, και το σταθερό μεγέθος του πληθυσμού. Σε κάθε κύκλο επιλέγονται για διασταύρωση τα χρωμοσώματα που βρίσκονται πιο κοντα στη λύση και οι απόγονοι που δημιουργούνται σε κάθε γενιά τείνουν να περιορίζονται γύρω από τη λύση. Επιπλέον τα λιγότερα κατάλληλα χρωμοσώματα, που θα μπορούσαν, ενδεχομένως, σε επόμενες γενιές να εξερευνήσουν νέες λύσεις, αντικαθίστανται έτσι ώστε να διατηρείται σταθερό το μέγεθος του πληθυσμού και αναπόφευκτα μειώνεται η ποικιλομορφία.

Λόγω αυτής της αδυναμίας των ΓΑ είναι αναμενόμενο να υστερούν στο πρόβλημα του ΠΦΠ καθώς θα αντιμετώπιζαν δυσκολίες να ακολουθήσουν τις εναλλαγές των θεμάτων ενδιαφέροντος. Ανακαλύψαμε όμως, μέσω πειραμάτων, πιο βασικά προβλήματα που αντιμετωπίζουν οι ΓΑ στο τομέα του ΠΦΠ. Προβλήματα που πηγάζουν από την αναπαράσταση των χρωμοσωμάτων σ' ένα διανυσματικό χώρο πολλών διαστάσεων. Για να μπορέσει ο αλγόριθμος να καλύψει μεγάλο εύρος θεμάτων απαιτείται από τα χρωμοσώματα ένας πολύ μεγάλος αριθμός λέξεων στο βεβαρημένο τους διάνυσμα. Καθώς όμως αυξάνονται οι διαστάσεις του διανύσματος μεγαλώνει και ο χώρος αναπαράστασης. Έτσι οι μετρικές που βασίζονται στην απόσταση δύο διανυσμάτων και χρησιμοποιούνται ως συναρτήσεις καταλληλότητας χάνουν μέρος της χρηστικότητας τους. Τα διανύσματα απέχουν πια τόσο πολύ το ένα από το άλλο τους με αποτέλεσμα η μέτρηση της μεταξύ τους απόστασης να μην αποτελεί ποιοτική σύγκριση (δεν προσδίδει κάποια επιπλέον ιδιότητα). Τέλος, καθώς αυξάνεται ο διανυσματικός χώρος αυξάνονται εκθετικά οι πιθανοί συνδυασμοί των λέξεων στο χρωμόσωμα. Επομένως, μιας κα οι λέξεις έχουν μεταβλητά βάρη σε κάθε χρωμόσωμα είναι πάρα πολύ δύσκολο (σχεδόν απίθανο) να δημιουργηθεί/παραχθεί, μέσω των διαδικασιών της διασταύρωσης και της μετάλλαξης, ο σωστός συνδυασμός λέξεων - βαρών που θα αντιπροσωπεύει τα ενδιαφέροντα του χρήστη.

Στην επόμενη ενότητα παραθέτουμε αποτελέσματα πειραμάτων που ενισχύουν τη συγκεκριμένη υπόθεση.

Κεφάλαιο 6

Πειράματα

Για την υλοποίηση του αλγορίθμου χρησημοποιήσαμε τη συλλογή εγγράφων reuters 21578, αποτελούμενη από 21,578 κείμενα ειδήσεων, δημοσιευμένα από το γνωστό πρακτορείο ειδήσεων, που αφορούσαν την επικαιρότητα το έτος 1987. Τα κείμενα έχουν χωριστεί σε κατηγορίες ανάλογα με το θέμα που διαπραγματεύονται. Από όλα τα θέματα χρησιμοποιούμε για τις ανάγκες των πειραμάτων μόνο αυτά τα οποία έχουν στη κατηγορία τους πάνω από 100 σχετικά κείμενα στη συλλογή – συνολικά 23 θέματα (Πίνακας 6.1). Να σημειωθεί πως τα κείμενα είναι στην αγγλική γλώσσα, χωρίς αυτό να σημαίνει ότι υπάρχουν περιορισμοί στη χρησιμοποίηση οποιασδήποτε άλλης γλώσσας[1].

6.1 Επεξεργασία Εγγράφων

Τα έγγραφα για να μπορούν να αξιοποιηθούν αποτελεσματικότερα υπόκεινται σε μια απαραίτητη προεργασία. Αφαιρούνται από το κείμενο του εγγράφου, με τη μέθοδο *stopword* όλες οι κοινές και συχνά χρησιμοποιούμενες λέξεις (π.χ and, an, is, with, for, about κ.α) καθώς και αυτές που έχουν λιγότερους από 2 χαρακτήρες. Έπειτα, από τις λέξεις που έχουν απομείνει αφαιρούνται οι περιττές καταλήξεις προσπαθώντας

[1]Για την υλοποίηση του αλγορίθμου χρησιμοποιήθηκε η γλώσσα προγραμματισμού Java.

Θέμα	Μέγεθος	Θέμα	Μέγεθος
earn	3987	money-supply	190
acq	2448	sugar	184
money-fx	801	gnp	163
crude	634	coffee	145
grain	628	veg-oil	137
trade	552	gold	135
interest	513	nat-gas	130
wheat	306	soybean	120
ship	305	bop	116
corn	254	livestock	114
dlr	217	cpi	112
oilseed	192		

ΠΙΝΑΚΑΣ 6.1: Τα θέματα που μελετώνται στα πειράματα και το αντίστοιχο μέγεθός τους

να απομονώσουμε τη " ρίζα " τους (μέθοδος *stemming*) σύμφωνα με τον αλγόριθμο του Porter [9].

Μετά την επεξεργασία των κειμένων κα την εξαγωγή όλων των *μοναδικών* λέξεων από τα κέιμενα της συλλογής δημιουργείται ένας πίνακας με στήλες όσες οι παραπάνω λέξεις και γραμμές όσα και τα κείμενα της συλλογής. Έτσι σε κάθε γραμμή υπολογίζεται η τιμή του Term Frequency Inverse Document Frequency (TFIDF) της κάθε λέξης από το συγκεκριμένο κείμενο. Το TFIDF της λέξης t ενός κειμένου περιγράφεται από τον τύπο[2]:

$$TFIDF_t = tf_t \times \ln \frac{|D|}{df_t} \tag{6.1}$$

όπου:

tf_t ο αριθμός των εμφανίσεων της λέξης t στο κείμενο(κανονικοποιημένος)

df_t ο αριθμός των κειμένων που περιέχουν τη λέξη t

$|D|$ ο συνολικός αριθμός των κειμένων

Με βάση το μέσο όρο των τιμών TFIDF της κάθε λέξης μπορούμε να επιλέξουμε τις πιο σημαντικές λέξεις από τα κείμενα της συλλογής.

[2]http://en.wikipedia.org/wiki/TFIDF

6.2 Πειραματικό πλαίσιο

Τα προφίλ - χρωμοσώματα αναπαρίστανται ως διανύσματα αριθμών διπλής ακρίβειας (double). Οι τιμές του διανύσματος κυμαίνονται από 0 ως 1 και δηλώνουν το βάρος συγκεκριμένων λέξεων, οι οποίες διαμορφώνουν το προφίλ. Όσο μεγαλύτερη είναι μια τιμή τόσο περισσότερη σημασία αποκτά αυτή η λέξη για το προφίλ. Το μέγεθος του διανύσματος αποτελεί παράμετρο του πειράματος και είναι ίση με τον αριθμό των *μοναδικών* λέξεων που θα εξαχθούν - επιλεγούν από τα κείμενα της συλλογής. Τα χρωμοσώματα αρχικοποιούνται συνήθως με τυχαίες τιμές[3].

Ο αλγόριθμος ξεκινά εξετάζοντας το πρώτο θέμα(topic). Στη συνέχεια ένα - ένα τα κείμενα αξιολογούνται χρησιμοποιώντας τη μέθοδο του εσωτερικού γινομένου (Εξ. 1.1). Η μεγαλύτερη από τις τιμές που θα δώσουν τα χρωμοσώματα αποτελεί το σκορ του κειμένου. Ο αλγόριθμός επαναλαμβάνει την προηγούμενη διαδικασία για όλα τα κείμενα της συλλογής. Έπειτα συνεχίζει με τα υπόλοιπα θέματα μέχρι να ολοκληρωθούν και τα 23.

6.3 Εκτέλεση Πειραμάτων

6.3.1 Πείραμα αναφοράς

Το πρώτο και βασικό μας πείραμα (*baseline*) δεν υλοποιούσε κάποιο γενετικό αλγόριθμο αλλά τον αλγόριθμο του Ροσσιο. Ορίζουμε ένα ξεχωριστό προφίλ για κάθε θέμα ενδιαφέροντος και στόχος μας είναι να "μάθει" το κάθε προφίλ το θέμα που του αναλογεί. Επιλέξαμε, βάσει του μέσου όρου των TFIDF βαρών τους, τις καλύτερες 100, 500, 10000, 5000, 10000, 20000 λέξεις και εκτελέσαμε το πείραμα για κάθε ένα από τους παραπάνω διανυσματικούς χώρους. Επίσης εκτελέσαμε και ένα τελευταίο

[3]Οι τιμές κυμαίνονται και αυτές από 0 έως 1.

πείραμα συμπεριλαμβάνοντας όλες τις λέξεις που προέκυψαν από τη συλλογή των κειμένων (31978).

Σε κάθε περίπτωση το κάθε κείμενο αναπαριστάται ως ένα βεβαρημένο διάνυσμα, έχοντας για βάρη τις τιμές TFIDF των αντίστοιχων μοναδικών λεξεών του. Το διάνυσμα του προφίλ αρχικοποιείται θέτοντας όλα του τα βάρη ίσα με *μηδέν* (0). Το προφίλ εξετάζει όλα τα κείμενα με χρονολογική σειρά και στη συνέχει αξιολογεί το καθένα από αυτά, δίνοντάς του ένα σκορ συσχέτισης. Το σκορ συσχέτισης είναι το αποτέλεσμα του εσωτερικού γινομένου μεταξύ των διανυσμάτων του κειμένου και του προφίλ (Εξ. 1.1, σελ. 3).

Κάθε φορά που συναντάται σχετικό κείμενο ο αλγόριθμος προσαρμόζει τα βάρη του προφίλ κάνοντας χρήση της συνάρτησης που περιγράφηκε προηγουμένως (Εξ. 1.3, σελ. 5). Το πείραμά μας δεν περιλαμβάνει αρνητική ανάδραση του χρήστη και, επομένως, δεν μας απασχολεί το δεύτερο σκέλος της συνάρτησης. Για το πείραμα μας επιλέχθηκαν οι συντελεστές $\alpha = 0.95$ και $\beta = 0.25$ με τους οποίους επετεύχθησαν τα καλύτερα αποτελέσματα όπως αναφέρονται και στο [8].

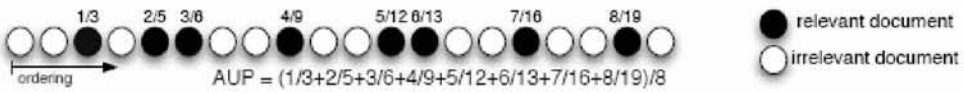
Αφού αποτιμηθούν και τα 21578 κείμενα της συλλογής, ταξινομούνται με φθίνουσα σειρά βάσει του σκορ τους, και υπολογίζεται το AUP (Average Uninterpolated Precision[4]). Το AUP ενός θέματος ορίζεται ως το άθροισμα της ακρίβειας (p) – το ποσοστό των σχετικών με το θέμα κειμένων – σε κάθε σημείο της ταξινομημένης λίστας όπου εμφανίζεται κάποιο σχετικό κείμενο, διαιρεμένης με το μέγεθος (N) του θέματος (Εξ. 6.2).

$$AUP = \sum \frac{p}{N} \tag{6.2}$$

Το Σχήμα 6.1 αναπαριστά τον υπολογισμό της τιμής AUP σε μια λίστα 20 κειμένων για ένα θέμα με 8 σχετικά κείμενα. Το AUP ενός θέματος είναι υψηλό όταν τα περισσότερα από τα κείμενα στην κορυφή της λίστας ανήκουν στο θέμα αυτό, πράγμα το οποίο σημαίνει ότι το προφίλ έχει αναθέσει σε αυτά τα κείμενα υψηλότερο σκορ

[4]Θα μπορούσε να μεταφραστεί ως "Μέση μη-παραμβληθείσα ακρίβεια". Στην συνέχεια της παρούσας εργασίας, για πρακτικούς λόγους, θα χρησιμοποιούμε τον όρο AUP.

απ' ότι στα υπόλοιπα.



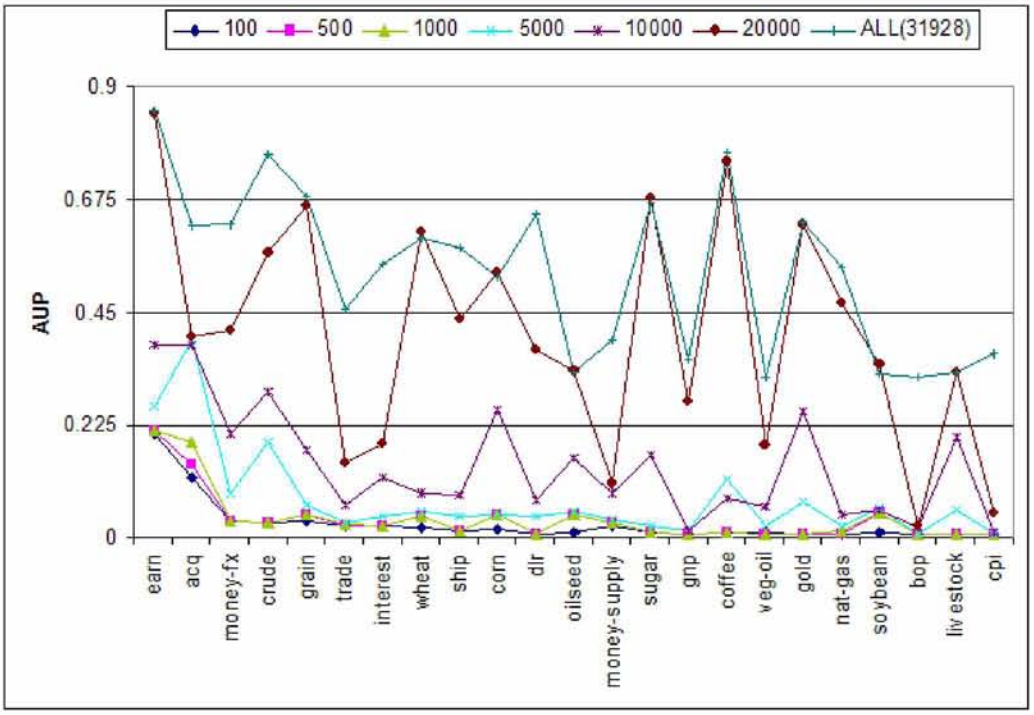
Σχημα 6.1: Υπολογισμός AUP

Τα αποτελέσματα του πειράματος αποτυπώνονται στο Σχήμα 6.2. Στον άξονα των X είναι τα 23 θέματα που μας ενδιαφέρουν (ταξινομημένα σε φθίνουσα σειρά μεγέθους) και σε αυτόν των Y οι τιμές AUP. Είναι φανερό ότι όσο περισσότερες είναι οι λέξεις τόσο μεγαλύτερες τιμές AUP επιτυγχάνονται. Όταν οι λέξεις είναι λίγες, μόνο όσα θέματα έχουν πολλά σχετικά ξεχωρίζουν ελάχιστα από τα υπόλοιπα, που κινούνται σε μηδενικά επίπεδα. Μια μικρή βελτίωση και στα υπόλοιπα θέματα παρουσιάζεται στις 1000 λέξεις. Όταν, όμως, οι λέξεις γίνονται 20000, και υπάρχει ικανοποιητικός αριθμός λέξεων για να περιγράψει τα θέματα, η βελτίωση γίνεται πιο έντονα αντιληπτή. Τα καλύτερα αποτελέσματα επιτυγχάνονται όταν για τη δημιουργία το διανυσματικού χώρου χρησιμοποιούνται όλες οι λέξεις. Αυτό απεικονίζεται ποιοτικά στο Σχήμα 6.3, όπου καταγράφεται ο μέσος όρος του AUP όλων των θεμάτων, σε κάθε κέκλο πειράματος. Όπως γίνεται αντιληπτό, για να καταφέρουμε να απεικονίσουμε αποτελεσματικά ένα μεγάλο εύρος θεμάτων απαιτείται ένας αρκετά πολυδιάστατος χώρος.

6.3.2 Εξελικτικά πειράματα

Εκτελούμε το προηγούμενο πείραμα χρησιμοποιώντας τώρα ένα γενετικό αλγόριθμο. Μιας και το πείραμα μας είναι αρκετά απλό και δεν περιλαμβάνει ούτε πολλαπλά θέματα ούτε εναλλαγές ενδιαφερόντων, αναμένουμε τουλάχιστον παρόμοια αποτελέσματα με αυτά του πειράματος αναφοράς.

Χρησιμοποιούμε όλες τις λέξεις των κειμένων (31,928) για καλύτερη κάλυψη των θεμάτων. Ορίζουμε έναν πληθυσμό από 100 προφίλ. Κάθε προφίλ είναι ένα βεβαρημένο διάνυσμα όπως προηγουμένως. Τα προφίλ αρχικοποιούνται με τυχαία βάρη

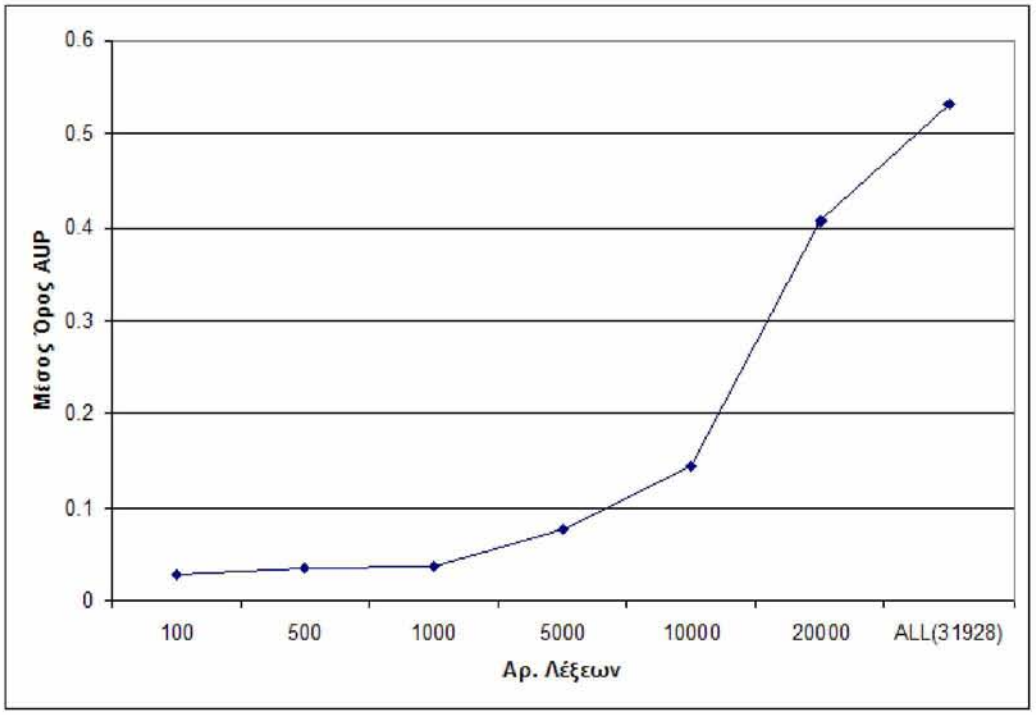


Σχημα 6.2: Πείραμα αναφοράς που χρησιμοποιεί τον αλγόριθμο του Rocchio

μεταξύ 0 και 1 και τιμή καταλληλότητας (fitness) ίση με 0. Ο ίδιος αρχικός πληθυσμός χρησιμοποιείται σε κάθε ένα από τα 23 θέματα. Οι διαδικασίες αξιολόγησης και ταξινόμησης των κειμένων γίνονται όπως στο πείραμα αναφοράς. Όταν εξετάζεται ένα σχετικό κείμενο, το προφίλ που έχει αναθέσει το μεγαλύτερο σκορ συσχέτισης αυξάνει το fitness του, προσθέτοντας σε αυτό το σκορ που υπολόγισε προηγουμένως[5].

Μετά την εξέταση κάθε σχετικού κειμένου λαμβάνουν χώρα οι γενετικές λειτουργίες του αλγορίθμου. Έπειτα, εάν το κείμενο που αξιολογείται ανήκει στη λίστα με τα σχετικά του υπό εξέταση θέματος κείμενα, τότε και μόνο τότε, ο πληθυσμός καλείται να δημιουργήσει μια νέα γενιά. Αφού αξιολογήσουν όλα τα χρωμοσώματα το κείμενο, ταξινομούνται με βάση το fitness τους και επιλέγονται τα καλύτερα από αυτά – 25% – για αναπαραγωγή. Η επιλογή των γονέων γίνεται με ντετερμινιστικό τρόπο καθώς το 1ο χρωμόσωμα διασταυρώνεται με το 2ο, το 3ο με το 4ο κ.ο.κ. Επιλέξαμε διασταύρωση ενός σημείου (single point crossover) και οι δύο απόγονοι που προκύπτουν υπόκεινται σε μετάλλαξη στο 5% των λέξεων τους, οι οποίες αντικαθίστανται από μία νέα τυχαία τιμή 0-1. Ο αλγόριθμός επαναλαμβάνει την προηγούμενη διαδικασία

[5] Δηλαδή, $fit_i = fit_{i-1} + \text{σκορ}$

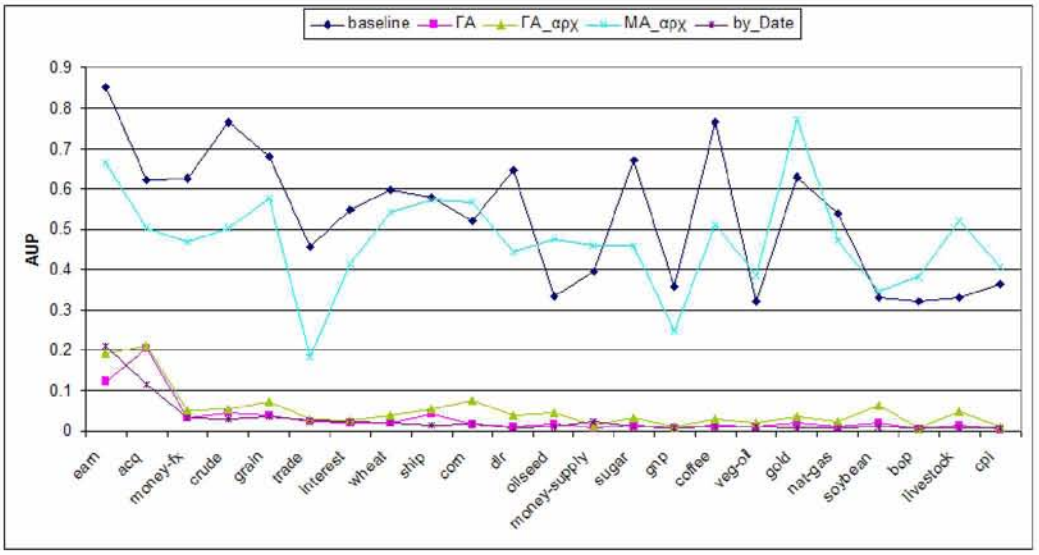


Σχημα 6.3: Μέσος όρος AUP του πειράματος αναφοράς

για όλα τα κείμενα της συλλογής. Έπειτα συνεχίζει με τα υπόλοιπα θέματα μέχρι να ολοκληρωθούν και τα 23 και υπολογίζει όπως προηγουμένως το AUP για το καθένα. Πραγματοποιούμε 50 επαναλήψεις της διαδικασίας αξιολόγησης και υπολογίζουμε 50 AUP για κάθε θέμα. Αυτό το επιλέξαμε για να έχει ο ΓΑ περισσότερα κείμενα για εκπαίδευση.

Επαναλάβαμε τα προηγούμενα πειράματα με δύο παραλλαγές του γενετικού αλγορίθμου. Στην πρώτη (ΓΑ_αρχ) ο πληθυσμός των προφίλ αρχικοποιείται θέτοντας το σταθμισμένο τους διάνυσμα ίσο με το διάνυσμα καθενός από τα 100 πρώτα σχετικά για κάθε θέμα κείμενα της συλλογής.

Η δεύτερη παραλλαγή (ΜΑ_αρχ) βασίζεται στην πρώτη αλλά την επεκτείνουμε προσθέτοντας δυνατότητες μάθησης. Πρόκειται για ένα ΜΑ που χρησιμοποιεί τον αλγόριθμο του Rocchio για να μετατοπίσει το διάνυσμα των πετυχημένων προφίλ προς αυτό του σχετικού κειμένου. Συγκεκριμένα, όποτε εξετάζεται ένα σχετικό κείμενο, τα βάρη του διανύσματος του προφίλ που έδωσε το μεγαλύτερο σκορ μεταβάλλονται σύμφωνα με την (1.3). Υιοθετούνται και σε αυτή την παραλλαγή οι ίδιοι συντελεστές μάθησης όπως στο πειραμα αναφοράς.

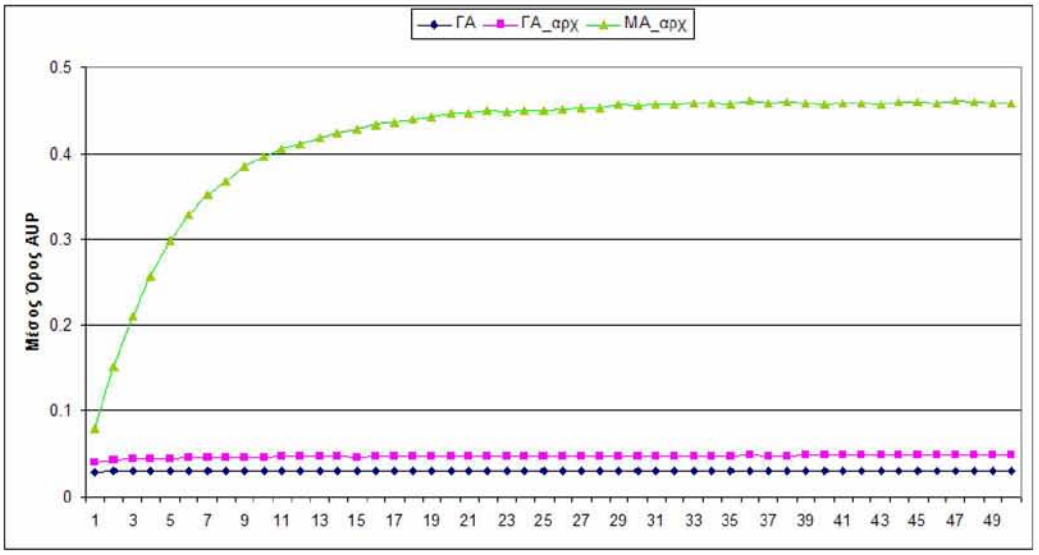


Σχημα 6.4: Συγκριτικά Αποτελέσματα

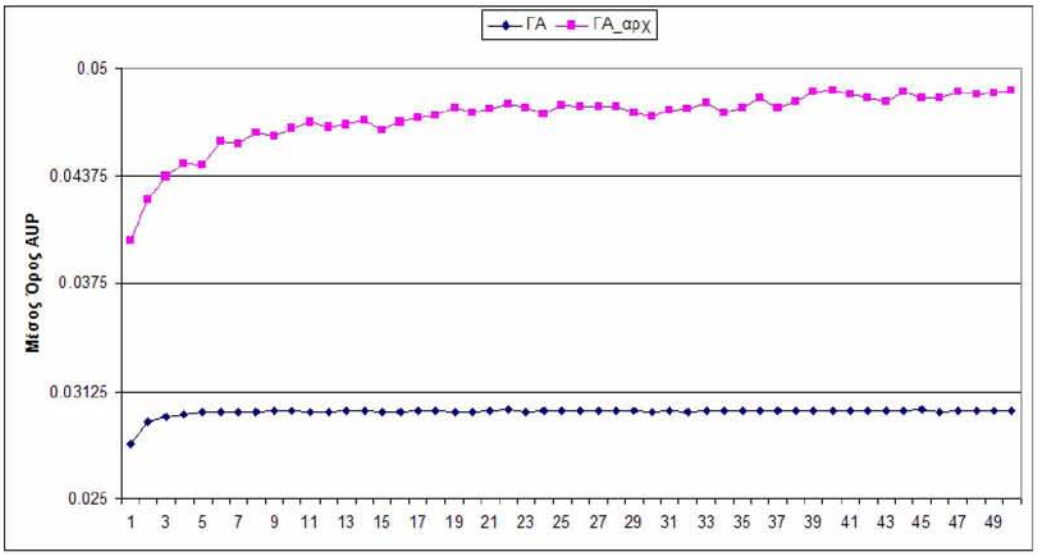
Στο Σχήμα 6.4 αποτυπώνεται για κάθε θέμα η μέγιστη τιμή AUP που επετεύχθη στη διάρκεια των 50 επαναλήψεων. Αυτή δεν είναι απαραίτητα η τιμή της τελευταίας επανάληψης (50η). Υιοθετούμε τη παραδοχή ότι έχουμε τη δυνατότητα να εντοπίσουμε τη μέγιστη τιμή κατά την εκτέλεση του εξελικτικού αλγορίθμου και να διακόψουμε τη διαδικασία. Περιλαμβάνονται, επίσης, στο διάγραμμα και οι τιμές του AUP των 23 θεμάτων, τις οποίες υπολογίσαμε έχοντας ταξινομήσει τα κείμενα *μόνο* με χρονολογική σειρά (by_Date), μη λαμβάνοντας υπόψη τα σκορ συσχέτισης.

Τα αποτελέσματα δεν είναι τα αναμενόμενα. Τα αποτελέσματα των τριών εξελικτικών αλγορίθμων είναι χειρότερα από αυτά του Rocchio. Ο ΓΑ στην απλή του μορφή έχει οριακά καλύτερα αποτελέσματα από αυτά τα των χρονολογικά ταξινομημένων κειμένων. Στη δεύτερη παραλλαγή οπου ο πληθυσμός των προφίλ αρχικοποιείται χρησιμοποιώντας σχετικά κείμενα, παρουσιάζεται μια μικρή βελτίωση στις τιμές AUP. Όταν ενισχύεται με τη συνάρτηση μάθησης επιτυγχάνει μια σημαντική βελτίωση στην απόδοσή του η οποία όμως δεν είναι αρκετή για να ξεπεράσει την απόδοση του Rocchio. Είναι καλύτερος μόνο για μερικά θέματα που περιέχουν λίγα σχετικά κείμενα στη συλλογή.

Το Σχήμα 6.5 παρουσιάζει το μέσο όρο των τιμών του AUP για τα 23 θέματα κατά τη διάρκεια των 50 επαναλήψεων. Σίγουρα ο μέσος όρος μπορεί να μην είναι μια ακριβής μέθοδος σύγκρισης, καθώς επηρεάζεται από τις ακραίες τιμές, αλλά μπορεί να



Σχημα 6.5: Μέση τιμή AUP κατά τη διάρκεια των επαναλήψεων



Σχημα 6.6: Σύγκριση των ΓΑ κατά τη διάρκεια των επαναλήψεων

μας δώσει μια σαφή εικόνα του τρόπου εξέλιξης του πληθυσμού στην όλη εξελικτική διαδικασία. Παρατηρούμε ότι ο ΜΑ_αρχ έχει μια ανοδική πορεία κατά τη διάρκεια των επαναλήψεων ενώ οι άλλοι 2 παραμένουν σχεδόν στάσιμοι. Εστιάζοντας ακόμα περισσότερο στους ΓΑ (Σχήμα 6.6) βλέπουμε ότι ο βασικός ΓΑ δεν εξελίσεται σχεδόν καθόλου. Αντίθετα, ο ΓΑ_αρχ ξεκινά με μεγαλύτερες τιμές AUP και σταδιακά εξελίσσεται έστω και σε πολύ μικρό βαθμό. Χρησιμοποιώντας όμως συνάρτηση μάθησης ο αλγόριθμος αποκτά μία πιο απότομη καμπύλη μάθησης η οποία βελτιώνεται σημαντικά και καταλήγει ασυμπτωτικά σε ένα όριο.

6.4 Ερμηνεία πειραματικών αποτελεσμάτων

Προσπαθώντας να ερμηνεύσουμε τα παραπάνω αποτελέσματα των πειραμάτων, μπορούμε να ισχυριστούμε ότι ο βασικός ΓΑ αδυνατεί να "μάθει" το θέμα ενδιαφέροντος διότι δεν μπορεί να δημιουργήσει, τυχαία, το σωστό συνδυασμό βαρών - λέξεων. Σε αντίθεση με τον αλγόριθμο του Rocchio που μπορεί να μεταβάλλει επιλεκτικά τα βάρη των λέξεων ο ΓΑ δεν διαθέτει κάποιον παρόμοιο μηχανισμό. Προσπαθεί να προσαρμόσει το πληθυσμό των προφίλ, που έχουν αρχικοποιηθεί με τυχαία βάρη, στο θέμα ενδιαφέροντος δημιουργώντας τυχαίους συνδυασμούς βαρών[6] και τυχαίες μεταλλάξεις στα διανύσματα των προφίλ. Μετά από λίγες γενιές, στις πρώτες επαναλήψεις, το προφίλ έχει ήδη δημιουργήσει τον καλύτερο συνδυασμό βαρών αλλά δεν μπορεί να βελτιωθεί περαιτέρω (Σχήμα 6.6).

Αρχικοποιώντας τον πληθυσμό με βάρη προερχόμενα από τα διανύσματα σχετικών με το θέμα κειμένων, ο αλγόριθμος ξεκινά από καλύτερη βάση. Οι λέξεις που έχουν μεγάλο βάρος (σημαντικότερες) στα σχετικά κείμενα έχουν μεγάλο βάρος και στο αρχικό προφίλ. Δίνεται στον αλγόριθμο η δυνατότητα για παραγωγή καλύτερων συνδυασμών από τις πρώτες επαναλήψεις και περισσότερες πιθανότητες βελτίωσης.

Με τη ενσωμάτωση της συνάρτησης μάθησης ο αλγόριθμος (MA_αρχ) έχει τη δυνατότητα να βελτιώσει τα βάρη των λέξεων σύμφωνα με την ανάδραση του χρήστη. Εντούτοις, όμως, αποδεικνύεται χειρότερος από το Rocchio, που χρησιμοποιήθηκε στο πείραμα αναφοράς, κι ας έχει αξιοποιήσει για την εκπαίδευση του περισσότερα κείμενα. Μια πιθανή εξήγηση είναι ότι λόγω της διαδικασίας διασταύρωσης των διανυσμάτων των προφίλ, καταστρέφεται ό,τι έχει καταφέρει να μάθει ο αλγόριθμος μέχρι εκείνο το σημείο. Επίσης η χρήση 100 προφίλ, σε αντίθεση με το ένα του Rocchio, τα οποία διαμοιράζονται τα σχετικά κείμενα αποτελεί πιθανή αιτία. Τώρα, το γεγονός ότι παρουσιάζει καλύτερες τιμές AUP στα θέματα με τα λιγότερα σχετικά κείμενα από τα 23, μπορεί να αποδοθεί στα 50 επαναληπτικά βήματα του αλγορίθμου και, συνεπώς, στο μεγαλύτερο αριθμό κειμένων που επεξεργάζεται.

[6] μέσω της διαδικασίας αναπαραγωγής

6.5 Συμπεράσματα

Τα παραπάνω ευρήματα είναι ενδιαφέροντα από πολλές απόψεις. Είναι γεγονός πως η υλοποίηση των παραπάνω πειραμάτων είναι σχετικά απλή, αλλά όλες οι τεχνικές που χρησιμοποιήθηκαν εμφανίζονται στη σχετική λογοτεχνία. Δεν ήταν στόχος μας απλώς να υλοποιήσουμε υπάρχουσες τεχνικές και να τις συγκρίνουμε. Ο σκοπός της τρέχουσας εργασίας είναι να τονίσει την πολυπλοκότητα του ΠΦΠ και να αποδείξει ότι το πρόβλημα δεν μπορεί να αντιμετωπιστεί εύκολα με έναν πληθυσμό σταθμισμένων διανυσμάτων σε ένα πολυδιάστατο διανυσματικό χώρο. Ο πολυδιάστατος χαρακτήρας του ΠΦΠ παραμελείται συνήθως στη βιβλιογραφία. Δεν έχουμε βρει ρητή αναφορά, στη βιβλιογραφία, για τον αριθμό των λέξεων-κλειδιών που χρησιμοποιούν.

Όπως δηλώνει και ο τίτλος της εργασίας, αυτή είναι μια προσπάθεια για να διερευνηθεί η εφαρμογή ΓΑ και ΜΑ στο ΠΦΠ και τα πειραματικά μας αποτελέσματα δείχνουν σαφώς ότι αυτό είναι ένα πολυδιάστατο πρόβλημα το οποίο δεν διευκολύνει την εφαρμογή των εν λόγω εξελικτικών προσεγγίσεων.

Οι επιδόσεις των εξελικτικών πειραμάτων, θα μπορούσαν να βελτιωθούν, ενδεχομένως, επιλέγοντας διαφορετικές τεχνικές για την αξιολόγηση καταλληλότητας, την επιλογή και μάθηση. Η αλήθεια είναι ότι δοκιμάσαμε μερικές ακόμα συναρτήσεις καταλληλότητας όπως επίσης υλοποιήσαμε τη διαδικασία επιλογής με ρουλεττε ωηεελ, αλλά τα αποτελέσματα δεν έδειξαν κάτι διαφορετικό και γι' αυτό δεν τα συμπεριλάβαμε όλα σε αυτήν εργασία Θωρούμε ότι οι μεγάλες διαφορές στις επιδόσεις μεταξύ των ΓΑ και του πειράματος αναφοράς απαιτούν πιο ριζοσπαστικές εναλλακτικές λύσεις.

Οι εξελικτικοί αλγόριθμοι του πειράματος αν και είχαν περισσότερα κείμενα να επεξεργαστούν (λόγω των επαναλήψεων) τελικώς υπολείπονταν του αλγόριθμου μάθησης του πειράματος αναφοράς. Αυτό ίσως σημαίνει ότι πρέπει να εξεταστούν εναλλακτικές λύσεις εκτός της αναπαράστασης του προφίλ ως ένα σταθμισμένο διάνυσμα. Θα πρέπει να βρεθούν αναπαραστάσεις που δεν απαιτούν πολυδιάστατους χώρους.

Το ΠΦΠ, αναμφίβολα, είναι μια δύσκολη πρόκληση και μέχρι στιγμής καμία προτεινόμενη λύση δεν έχει καταφέρει να παράγει επιτυχής εφαρμογές που να ανταποκρίνονται στις απαιτήσεις και τις ιδιαιτερότητες του πραγματικού κόσμου.

Σε καμία περίπτωση δεν απορρίπτουμε την εξελικτική προσέγγιση στο ΠΦΠ. Πιστεύουμε ότι οι εμπνευσμένοι από τη βιολογία αλγόριθμοι ταιριάζουν στην πολυπλοκότητα και τη δυναμική του προβλήματος. Ωστόσο, τέτοιες προσεγγίσεις παρέμειναν έξω από τη επικρατούσα τάση της έρευνας στο ΠΦΠ και δεν έχουν αξιολογηθεί μεθοδικά.

Οι πειραματικές μεθοδολογίες που παρουσιάζονται στην παρούσα εργασία αλλά και στο [8] προτείνουν μια νέα τεχνική για την αξιολόγηση και τον έλεγχο συστημάτων ΠΦΠ, η οποία αντικατοπτρίζει ακριβέστερα πραγματικές καταστάσεις.

Κεφάλαιο 7

Επίλογος

Ο μεγάλος όγκος των διαθέσιμων πληροφοριών αποτελεί κεντρικό ζήτημα του Διαδικτύου. Μέχρι σήμερα δεν έχει παρουσιαστεί κάποια λύση που να βρίσκει εφαρμογή στον πραγματικο κόσμο και να μπορεί να διακρίνει και επιλέγει εξατομικευμένες πληροφορίες. Η προσαρμογή προφίλ είναι ένα δύσκολο πρόβλημα με δικά του χαρακτηριστικά και απαιτήσεις. Αν και το ΠΦΠ δεν είναι κλασσικό πρόβλημα βελτιστοποίησης χρησιμοποιούνται, για την προσαρμογή του προφίλ, εξελικτικοί αλγόριθμοι. Οι εξελικτικοί αλγόριθμοι ορίζουν ένα πληθυσμό από προφίλ, που αναπαριστούν τα συνολικά ενδιαφέροντα του χρήστη, ο οποίος εξελίσσεται με βάση την ανάδραση του χρήστη (feedback).

Παρ' όλα τα θετικά αποτελέσματα που αναφέρονται, θίξαμε σε αυτή την εργασία προβλήματα διαστάσεων (dimensionality) που προκύπτουν από τη χρησιμοποίηση σταθμισμένων διανυσμάτων για την αναπαράσταση των προφίλ και των κειμένων. Όταν οι διαστάσεις του διανυσματικού χώρου αυξάνονται τότε η μετρικές μέθοδοι που βασίζονται στην απόσταση των διανυσμάτων χάνουν την διακριτική τους ικανότητα. Επίσης, όταν αυξάνονται οι διαστάσεις ή οι λέξεις του διανύσματος τότε οι συνδυασμοί των σταθμισμένων λέξεων αυξάνοντα εκθετικά. Εϊναι, τότε, πολύ δύσκολο τυχαίες γενετικές λειτουργίες να παράγουν το σωστό συνδυασμό σταθμισμένων λέξεων για να αναπαραστήσουν επιτυχώς τα ενδιαφέροντα του χρήστη.

Εκτελέσαμε μια σειρά πειραμάτων σε ένα πολυδιάστατο χώρο και συγκρίναμε τρεις παραλλαγές των γενετικών αλγορίθμων με τον αλγόριθμο μάθησης του Rocchio (πείραμα αναφοράς). Παρ' όλο που το πειραματικό πλαίσιο ήταν απλό και δεν ανταποκρινόταν πλήρως στην πολυπλοκότητα και στη δυναμική του ΠΦΠ, οι ΓΑ απέδωσαν αρκετά χειρότερα από τον αλγόριθμο του Rocchio. Ακόμα κι όταν ενσωματώθηκε στους ΓΑ συνάρτηση μάθησης τα αποτελέσματα δεν προσέγγισαν αυτά του πειράματος αναφοράς.

Επιφυλασσόμαστε για την υλοποίηση κι άλλων εναλλακτικών τεχνικών αξιολόγησης της καταλληλότητας, επιλογής "γονέων" και άλλων γενετικών πράξεων, αλλά τα μέχρι στιγμής πειραματικά αποτελέσματα υποδηλώνουν ότι απαιτούνται πιο δραστικές λύσεις για να αντιμετωπιστεί το εγγενές πρόβλημα των πολλών διαστάσεων (dimensionality) που παρουσιάζουν οι διανυσματικές αναπαραστάσεις. Ίσως, οι ΓΑ και οι ΜΑ να μην είναι κατάλληλοι για την αντιμετώπιση του ΠΦΠ, και θα έπρεπε άλλες βιολογικά εμπνευσμένες λύσεις να εξερευνηθούν π.χ Τεχνητό Ανοσοποιητικό σύστημα ή Swarm Intelligence. Πιστεύουμε, τέλος, πως η αναζήτηση εναλλακτικών προσεγγίσεων στο ΠΦΠ θα πρέπει να πραγματοποιηθεί προς την κατεύθυνση όπου ο αλγόριθμος είναι ο μόνος υπεύθυνος για τον καθορισμό των κανόνων ανάπτυξης μιας αναπαράστασης προφίλ και όχι η ίδια η αναπαράσταση προφίλ. Η εξατομίκευση της δικτυακής πληροφορίας απαιτεί αποδοτικό ΠΦΠ και οι εμπνευσμένοι από τη βιολογία αλγόριθμοι είναι υποψήφιοι για να δώσουν λύση σ' αυτό άλυτο, πολύπλοκο και δυναμικό πρόβλημα.

Bibliography

[1] R. Dawkins. *The Selfish Gene.* Oxford University Press, 1990.

[2] G. Desjardins and R. Godin. Combining relevance feedback and genetic algorithm in an internet information filtering engine. In *RIAO 2000*, 2000.

[3] Gary William Flake. *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation.* The MIT Press, Cambridge, Massachusetts, 1998.

[4] Pablo Moscato. On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. Technical Report C3P Report 826, California Institute of Technology, 1989.

[5] Nikolaos Nanas and Anne De Roeck. Multimodal dynamic optimisation: from evolutionary algorithms to artificial immune systems. In *Proc. of the 6th International Conference on Artificial Immune Systems*, pages 13–24, 2007.

[6] Nikolaos Nanas and Anne De Roeck. A review of evolutionary and immune inspired information filtering. *Natural Computing*, 2007.

[7] Nikolaos Nanas, Manolis Vavalis, and Anne De Roeck. What happened to content based information filtering? In Leif Azzopadi, Gabrieall Kazai, Stephen Robertson, Stefan Ruger, Milad Shokouhi, Dawei Song, and Emine Yilmaz, editors, *International Conference on the Theory of Information Retrieval*, Lecture Notes in Computer Science, pages 249–256, 2009.

[8] Nikolaos Nanas, Manolis Vavalis, and Lefteris Kellis. Immune learning in a dynamic information environment. In *8th International Conference on Artificial Immune Systems*, pages 192–205, 2009.

[9] M. Porter. An algorithm for suffix stripping. *Program*, 14(3), 1980.

[10] Y. Seo and B. Zhang. A reinforcement learning agent for personalized information filtering. In *Intelligent User Interfaces*, pages 248–251, New Orleans, LA, 2000.

[11] B. D. Sheth. *A Learning Approach to Personalized Information Filtering.* Master of Science, Massachusetts Institute of Technology, 1994.

[12] D. R. Tauritz, J. N. Kok, and I. G. Sprinkhuizen-Kuyper. Adaptive information filtering using evolutionary computation. *Information Sciences*, 122(2–4):121–140, 2000.

[13] Daniel Remy Tauritz. *Adaptive Information Filtering: concepts and algorithms.* PhD thesis, Leiden University, 2002.

[14] Geoffrey I. Webb, Michael J. Pazzani, and Daniel Billsus. Machine learning for user modeling. *User Modeling and User-Adapted Interaction*, 11:19–29, 2001.