ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

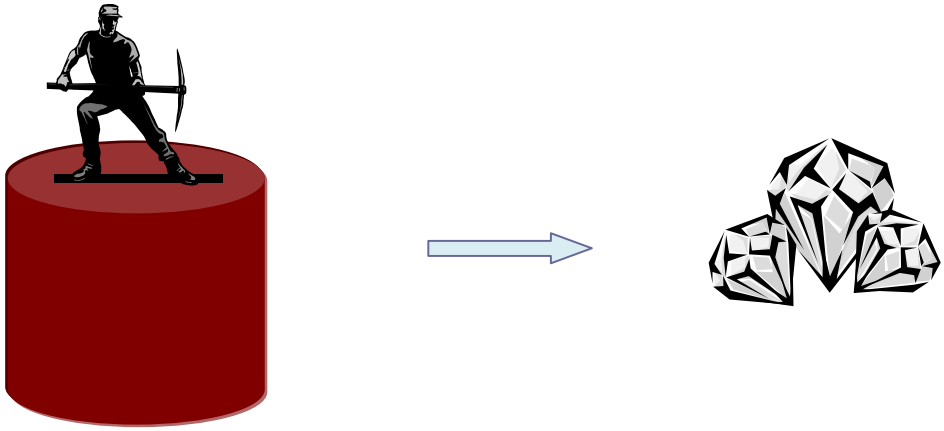
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

Μεταπτυχιακή Εργασία

Εξόρυξη γνώσης από ειδησεογραφικά δεδομένα και συσχετισμός με πραγματικά γεγονότα

Ειρήνη Ντούτση

Μηχανικός Η/Υ και Πληροφορικής



Επιβλέπων

Δημήτρης Χριστοδουλάκης, καθηγητής

Τριμελής επιτροπή

Δημήτρης Χριστοδουλάκης, καθηγητής

Ιωάννης Γαροφαλάκης, επίκουρος καθηγητής

Χαράλαμπος Ζαγούρας, καθηγητής

Πάτρα, Ιούλιος 2003

Αντί προλόγου

Κατ' αρχήν θα ήθελα να ευχαριστήσω τον επιβλέποντά μου καθ. **Δημήτρη Χριστοδουλάκη** για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα, για την τεχνική του καθοδήγηση και για τις συμβουλές σε ουσιαστικά ζητήματα που προέκυψαν κατά τη διάρκεια της μεταπτυχιακής εργασίας.

Επίσης θα ήθελα να ευχαριστήσω και τα υπόλοιπα μέλη της τριμελούς επιτροπής κ.κ. **Γαροφαλάκη Ιωάννη, επίκουρο καθηγητή** και **Χαράλαμπο Ζαγούρα, καθηγητή** για την εμπιστοσύνη που μου έδειξαν καθόλη τη διάρκεια της μεταπτυχιακής εργασίας.

Ευχαριστώ επίσης το συνάδελφο **Αθανάσιο Παπαγγελή** για την πολύ ενδιαφέρουσα κουβέντα που είχαμε – αυτό ήταν στην ουσία το έναυσμα για να ξεκινήσω. Πολλά ευχαριστώ και στον καθ. **Γιάννη Θεοδωρίδη**, η συζήτηση πάντα λύνει απορίες καθώς επίσης και στα παιδιά του Εργαστηρίου Βάσεων Δεδομένων, το **Μανόλη Τζαγκαράκη** και τη **Σοφία Στάμου**, για τα εποικοδομητικά τους σχόλια σχετικά με την εφαρμογή.

Επίσης, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για το ενδιαφέρον και τη στήριξη που μου έδειξαν καθ' όλη τη διάρκεια των μεταπτυχιακών μου σπουδών. Ειδικότερα ευχαριστώ το συνάδελφο και φίλο **Κώστα Γρατσία** που συν τις άλλοις ανέλαβε και το σχολιασμό του κειμένου της μεταπτυχιακής εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω το **Νίκο** για την ουσιαστική βοήθεια που μου προσέφερε σε θέματα της μεταπτυχιακής εργασίας, αλλά και για το γεγονός ότι στάθηκε δίπλα μου και με υποστήριξε στις δύσκολες στιγμές.

Πάτρα, Ιούλιος 2003

Ειρήνη Ντούτση

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

Μέρος Α

Μέρος Β

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

Μέρος Α

Μέρος Β

Εισαγωγικά

Η **Ανακάλυψη Γνώσης** μέσα από Βάσεις Δεδομένων *(Knowledge Discovery in Databases - KDD)* αποτελεί ένα ταχέως αναπτυσσόμενο επιστημονικό πεδίο που αποσκοπεί στην ανακάλυψη κρίσιμων "κρυμμένων" πληροφοριών μέσα από τις βάσεις δεδομένων. Οι εφαρμογές της ανακάλυψης γνώσης αυξάνονται συνεχώς λόγω της έκρηξης της πληροφορίας που παρατηρείται τα τελευταία χρόνια και της ανάγκης για ανακάλυψη χρήσιμων πληροφοριών από την πληθώρα των διαθέσιμων δεδομένων.

Ένα από τα βασικά βήματα στη διαδικασία ανακάλυψης γνώσης αποτελεί η **Εξόρυξη Γνώσης** *(Data Mining)* η οποία περιλαμβάνει κυρίως τις διαδικασίες και τα μέσα εξαγωγής προτύπων *(patterns)* από το σύνολο των δεδομένων. Μέχρι πρόσφατα η εξόρυξη γνώσης αφορούσε αποκλειστικά δομημένα δεδομένα (δηλαδή δεδομένα που είναι αποθηκευμένα σε βάσεις δεδομένων), τα τελευταία όμως χρόνια το ενδιαφέρον στράφηκε και σε μη δομημένα δεδομένα (π.χ. κείμενα, εικόνες, έγγραφα, web σελίδες) και ένας νέος κλάδος προέκυψε, η **Εξόρυξη Γνώσης από Κείμενα** *(Text Mining)*. Η στροφή αυτή είναι πολύ σημαντική καθώς η πλειοψηφία των δεδομένων σήμερα (περίπου το 80% του όγκου των δεδομένων) διατίθενται σε μη δομημένη μορφή και το ποσοστό αυτό αναμένεται να αυξάνεται συνεχώς λόγω της εξάπλωσης της χρήσης του διαδικτύου και της ηλεκτρονικής επικοινωνίας.

Το αντικείμενο της παρούσας μεταπτυχιακής εργασίας είναι ο σχεδιασμός και η υλοποίηση ενός πρωτότυπου συστήματος εξόρυξης γνώσης από ειδησεογραφικά δεδομένα και εν συνεχεία ο έλεγχος της απόδοσης του υλοποιηθέντος συστήματος σε περιπτώσεις πραγματικών προβλημάτων.

Ως πραγματικά προβλήματα χρησιμοποιήθηκαν 1) το πρόβλημα της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού με γνώμονα ειδησεογραφικά δεδομένα που αφορούν τις εσωτερικές και εξωτερικές εξελίξεις και στοιχεία σχετικά με το πλήθος των τουριστών του εξωτερικού στην Ελλάδα και 2) το πρόβλημα της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου (για τους σκοπούς της μεταπτυχιακής εργασίας χρησιμοποιήθηκε ο δικτυακός τόπος Flash - www.flash.gr)

Το κείμενο της μεταπτυχιακής εργασίας απαρτίζεται από 7 κεφάλαια τα οποία είναι ομαδοποιημένα σε δύο μέρη.

Στο **πρώτο μέρος** της μεταπτυχιακής εργασίας πραγματοποιείται μία σύντομη αναφορά στις έννοιες της Ανακάλυψης Γνώσης μέσα από Βάσεις Δεδομένων, της Εξόρυξης Γνώσης και της Εξόρυξης Γνώσης από Κείμενα. Ειδικότερα στο **πρώτο κεφάλαιο** παρουσιάζονται οι έννοιες της Ανακάλυψης Γνώσης μέσα από Βάσεις Δεδομένων και της Εξόρυξης Γνώσης (πως ορίζονται, ποιες είναι οι κατηγορίες της Εξόρυξης Γνώσης, ποιες τεχνικές χρησιμοποιούνται κ.λ.π). Στο **δεύτερο κεφάλαιο**

παρουσιάζεται η έννοια της Εξόρυξης Γνώσης από Κείμενα (πως ορίζεται, ποιες οι διαφορές της από την κλασσική Εξόρυξη Γνώσης, ποιες τεχνικές χρησιμοποιούνται κ.λ.π.).

Στο **δεύτερο μέρος** της μεταπτυχιακής εργασίας παρουσιάζεται με αναλυτικό τρόπο το σύστημα που αναπτύχθηκε για τους σκοπούς της εργασίας. Η παρουσίαση αυτή περιλαμβάνει:

- Μελέτη της αρχιτεκτονικής του συστήματος και ανάλυση των δομικών τμημάτων που το απαρτίζουν και του τρόπου επικοινωνίας μεταξύ των διαφόρων τμημάτων (**πρώτο κεφάλαιο**).

- Σύντομη αναφορά στο πρόβλημα της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού: ποιο είναι το πρόβλημα, τι δεδομένα έχουμε στη διάθεσή μας για την επίλυσή του κ.λ.π. (**δεύτερο κεφάλαιο**).

- Σύντομη αναφορά στο πρόβλημα της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου: ποιο είναι το πρόβλημα, τι δεδομένα έχουμε στη διάθεσή μας για την επίλυσή του κ.λ.π. (**τρίτο κεφάλαιο**).

- Παρουσίαση των πειραμάτων και των αντίστοιχων αποτελεσμάτων από την εφαρμογή του συστήματος στο πρόβλημα της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού και στο πρόβλημα της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου (**τέταρτο κεφάλαιο**)

- Παρουσίαση του περιβάλλοντος διεπαφής χρήστη της εφαρμογής μέσω αντιπροσωπευτικών οθονών από το τρέξιμο της εφαρμογής (**πέμπτο κεφάλαιο**).

Στον επίλογο παρουσιάζονται τα βασικά συμπεράσματα της μεταπτυχιακής εργασίας.

Στο τέλος της μεταπτυχιακής εργασίας υπάρχει αναλυτικά η βιβλιογραφία που χρησιμοποιήθηκε, στην οποία μπορεί να ανατρέξει κανείς για περισσότερες λεπτομέρειες καθώς και για θέματα τα οποία δεν κρίθηκε σκόπιμο να αναλυθούν στη συγκεκριμένη μεταπτυχιακή εργασία.

ΜΕΡΟΣ Α

Το θεωρητικό υπόβαθρο της μεταπτυχιακής εργασίας

1. Ανακάλυψη Γνώσης μέσα από Βάσεις Δεδομένων

1.1 Εισαγωγή

Τα τελευταία χρόνια με την εξάπλωση της χρήσης των υπολογιστών σε όλους τους τομείς της ζωής μας έχουν αυξηθεί σημαντικά οι δυνατότητές μας να παράγουμε και να συλλέγουμε πληροφορίες, γεγονός που οδήγησε στην συγκέντρωση μεγάλου όγκου πληροφορίας. Η αύξηση αυτή κάνει επιτακτική την ανάγκη εύρεσης νέων τεχνικών και εργαλείων που θα υποστηρίζουν την αυτόματη μετατροπή των υπό επεξεργασία πληροφοριών σε χρήσιμη γνώση.

Ένα νέο πεδίο έρευνας, η Ανακάλυψη Γνώσης μέσα από Βάσεις Δεδομένων *(Knowledge Discovery in Databases - KDD)* έχει καθιερωθεί ως το κατεξοχήν πεδίο για την ανακάλυψη κρίσιμων πληροφοριών που ενδέχεται να υπάρχουν «κρυμμένες» μέσα σε μεγάλες βάσεις δεδομένων.

Η ανακάλυψη γνώσης αποτελεί έναν ταχέως αναπτυσσόμενο τομέα, η εξέλιξη του οποίου κατευθύνεται τόσο από ερευνητικά ενδιαφέροντα όσο και από ισχυρές πρακτικές, οικονομικές και κοινωνικές ανάγκες. Στο κεφάλαιο αυτό θα παρουσιάσουμε τον ορισμό και τις βασικές έννοιες και μεθόδους που χρησιμοποιούνται στην Ανακάλυψη Γνώσης.

1.2 Ανακάλυψη Γνώσης

Θα δώσουμε ένα γενικό ορισμό της έννοιας της ανακάλυψης γνώσης μέσα από βάσεις δεδομένων.

Ορισμός

Η **Ανακάλυψη Γνώσης μέσα από Βάσεις Δεδομένων** *(KDD)* είναι μια μη-τετριμμένη διαδικασία για την αναγνώριση έγκυρων, νέων, χρήσιμων και εύκολα κατανοητών προτύπων από τα δεδομένα.

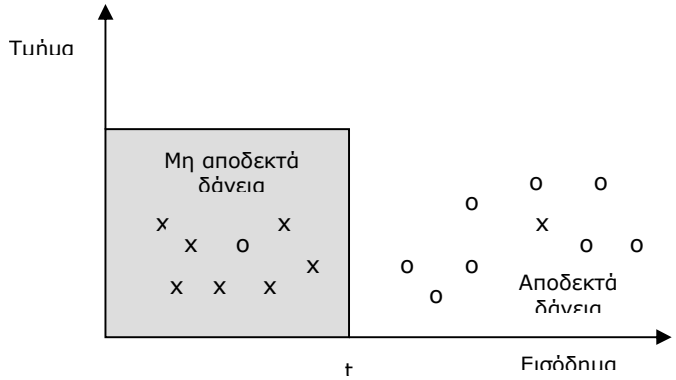
Ας δούμε όμως αναλυτικά τι σημαίνει κάθε επιμέρους όρος του ορισμού:

- Δεδομένα

Πρόκειται για ένα σύνολο παραδειγμάτων / στιγμιότυπων ενός προβλήματος που εμφανίζονται σε μια βάση δεδομένων. Για παράδειγμα, θα μπορούσε να είναι μια συλλογή εγγραφών από τη βάση δεδομένων μιας τράπεζας, όπου κάθε εγγραφή θα περιλάμβανε τρία πεδία (γνωρίσματα): το χρέος, το εισόδημα και την κατάσταση του δανείου των πελατών της τράπεζας.

- Πρότυπα *(patterns)*

Πρόκειται για εκφράσεις σε μια συγκεκριμένη γλώσσα οι οποίες περιγράφουν ένα υποσύνολο του συνόλου των παραδειγμάτων. Για παράδειγμα, η έκφραση '*Αν το εισόδημα<t, τότε ο πελάτης δεν μπορεί να εξοφλήσει το δάνειο*', θα μπορούσε να είναι ένα πρότυπο για κάποιο κατάλληλο κατώφλι *t* (Σχήμα 1.1).



Σχήμα 1.1 Ταξινόμηση των δεδομένων με χρήση κατάλληλου κατωφλίου *t*.Η σκιασμένη περιοχή του σχήματος είναι η περιοχή των μη αποδεκτών δανείων.

- KDD Διαδικασία

Πρόκειται για μια διαδικασία πολλών βημάτων που περιλαμβάνει την κατάλληλη προετοιμασία των δεδομένων, την αναζήτηση προτύπων και την αξιολόγηση της αποκτηθείσας γνώσης. Η KDD διαδικασία δεν είναι τετριμμένη καθώς εμπεριέχει κάποιο βαθμό αυτονομίας. Στο παράδειγμα του δανείου που αναφέραμε πιο πριν, ο υπολογισμός του μέσου όρου εισοδήματος του πελάτη αποτελεί πολύ χρήσιμο αποτέλεσμα, σε καμία όμως περίπτωση δεν αποτελεί ανακάλυψη γνώσης.

- Εγκυρότητα

Τα πρότυπα που προκύπτουν από τη διαδικασία ανακάλυψης γνώσης θα πρέπει να ισχύουν με κάποιο βαθμό βεβαιότητας και για νέα, άγνωστα στιγμιότυπα του προβλήματος. Για παράδειγμα, αν στο πρότυπο που απεικονίζεται στο Σχήμα 2.1 το κατώφλι μετακινηθεί προς τα δεξιά τότε το μέτρο βεβαιότητας θα μειωθεί καθώς περισσότερα αποδεκτά μέχρι πρότινος δάνεια θα ανήκουν πλέον στην περιοχή των μη αποδεκτών δανείων.

- Χρησιμότητα

Τα πρότυπα θα πρέπει να είναι χρήσιμα, δηλαδή να οδηγούν σε κάποιες χρήσιμες ενέργειες. Για παράδειγμα, αν η τράπεζα εκμεταλλευτεί τους κανόνες απόφασης του Σχήματος 2.1, θα πρέπει να πετύχει αύξηση των κερδών της.

- Κατανοησιμότητα

Τα πρότυπα θα πρέπει να είναι κατανοητά από τον ανθρώπινο παράγοντα, καθώς οι άνθρωποι είναι αυτοί που θα κληθούν να τα αξιοποιήσουν προκειμένου να εξάγουν χρήσιμα συμπεράσματα και να αποκτήσουν μια βαθύτερη κατανόηση των δεδομένων τους. Για την κατανόηση των προτύπων δε θα πρέπει να απαιτούνται εξειδικευμένες γνώσεις, αντιθέτως τα πρότυπα θα πρέπει να είναι πλήρως κατανοητά και να βοηθούν ακόμη και μη ειδικούς στην εξαγωγή χρήσιμων συμπερασμάτων.

Η Ανακάλυψη Γνώσης μέσα από Βάσεις Δεδομένων (KDD) αναφέρεται στη συνολική διαδικασία ανακάλυψης χρήσιμης πληροφορίας από τα δεδομένα. Ένα βήμα σ' αυτή τη διαδικασία αποτελεί και η Εξόρυξη Γνώσης *(Data Mining)*, της οποίας ο ορισμός ακολουθεί.

Ορισμός

Η **Εξόρυξη Γνώσης** αποτελεί ένα βήμα της KDD διαδικασίας και ορίζεται ως η διαδικασία της ανακάλυψης νέων πιθανώς κρυμμένων προτύπων και μοντέλων με αυτόματο ή ημιαυτόματο τρόπο με απώτερο στόχο την περιγραφή των δεδομένων μιας βάσης δεδομένων και την πρόβλεψη και εξήγηση νέων δεδομένων.

Η Εξόρυξη Γνώσης περιλαμβάνει κυρίως τις διαδικασίες και τα μέσα εξαγωγής προτύπων από το σύνολο των δεδομένων. Μέχρι πρόσφατα αφορούσε αποκλειστικά δομημένα δεδομένα (δηλαδή δεδομένα που είναι αποθηκευμένα σε βάσεις δεδομένων), τα τελευταία όμως χρόνια το ενδιαφέρον στράφηκε και σε μη δομημένα δεδομένα (π.χ. κείμενα, εικόνες, έγγραφα, web σελίδες) και ένας νέος κλάδος προέκυψε, η **Εξόρυξη Γνώσης από Κείμενα** *(Text Mining)*. Η στροφή αυτή είναι πολύ σημαντική καθώς η πλειοψηφία των δεδομένων σήμερα (περίπου το 80% του όγκου των δεδομένων) διατίθενται με τη μορφή μη δομημένων κειμένων.

1.3 Η KDD διαδικασία

Η KDD διαδικασία είναι μια αλληλεπιδραστική και επαναληπτική διαδικασία, η οποία περιλαμβάνει πλήθος βημάτων στα οποία χρειάζεται πολλές φορές να παρέμβει και ο άνθρωπος λαμβάνοντας κρίσιμες αποφάσεις.

Τα βασικά βήματα της KDD διαδικασίας είναι τα ακόλουθα (Σχήμα 1.2):

- *Ανάπτυξη και κατανόηση του πεδίου της εφαρμογής* συμπεριλαμβανόμενης οποιασδήποτε σχετικής προηγούμενης γνώσης για το πρόβλημα καθώς επίσης και των στόχων / προσδοκιών των τελικών χρηστών.

- *Δημιουργία του στοχευόμενου συνόλου δεδομένων (target data), το οποίο θα περιλαμβάνει* τα δεδομένα από τα οποία πρόκειται να εξαχθεί η γνώση. Το βήμα αυτό είναι εξαιρετικά κρίσιμο καθώς η ποιότητα των δεδομένων επηρεάζει την απόδοση του συστήματος ανακάλυψης γνώσης.

- *Καθαρισμός και επεξεργασία των δεδομένων (data cleaning)*. Το βήμα αυτό περιλαμβάνει βασικές λειτουργίες όπως η απομάκρυνση του θορύβου, η αντιμετώπιση του προβλήματος των δεδομένων με ελλιπείς τιμές κ.α.

- *Μείωση της ποσότητας των δεδομένων (data reduction).* Το βήμα αυτό περιλαμβάνει την εύρεση χρήσιμων χαρακτηριστικών για την αναπαράσταση των δεδομένων του προβλήματος ανάλογα με τους στόχους της ανακάλυψης γνώσης, τη μείωση του πλήθους αυτών των χαρακτηριστικών κ.α.

- *Επιλογή των εργασιών εξόρυξης γνώσης (data mining)* που θα χρησιμοποιηθούν για τις ανάγκες του προβλήματος, π.χ. ταξινόμηση, πρόβλεψη, ομαδοποίηση κ.α.

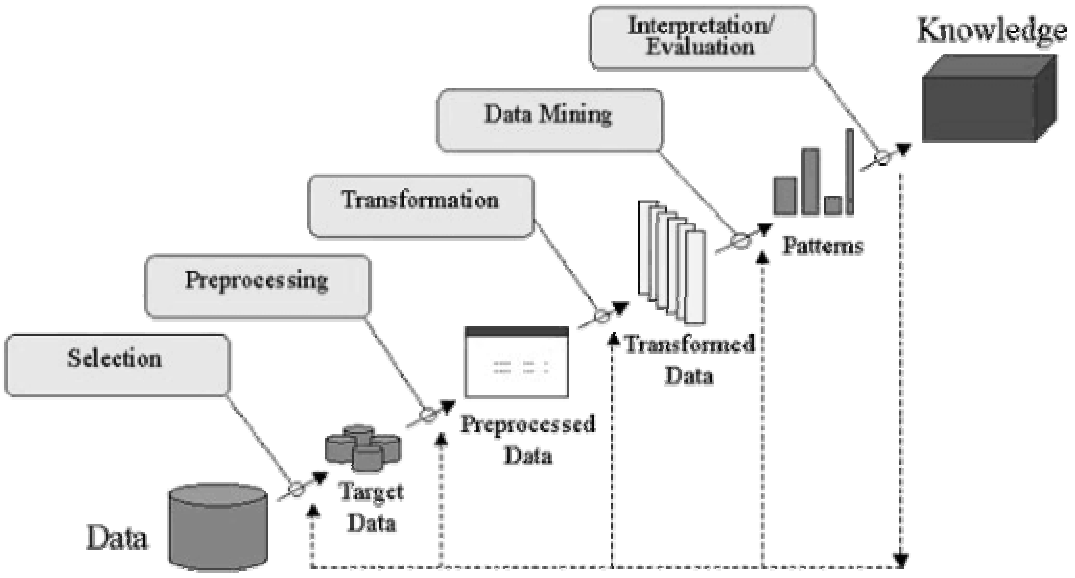
- *Επιλογή των αλγορίθμων εξόρυξης γνώσης (data mining)* που θα χρησιμοποιηθούν για την αναζήτηση προτύπων στα δεδομένα. Το βήμα αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου, την επιλογή των κατάλληλων παραμέτρων του μοντέλου κ.α.

- *Data Mining*: αναζήτηση στα δεδομένα των προτύπων που μας ενδιαφέρουν.

- *Ερμηνεία των προτύπων* που ανακαλύφθηκαν από την KDD διαδικασία - πιθανόν να χρειαστεί να επιστρέψουμε και πάλι σε κάποιο από τα παραπάνω βήματα.

- *Ενοποίηση της γνώσης που έχει εξαχθεί:* ενσωμάτωση αυτής της γνώσης στο σύστημα ή απλά κοινοποίησή της με την κατάλληλη τεκμηρίωση στα ενδιαφερόμενα μέλη. Το βήμα αυτό περιλαμβάνει και έλεγχο συγκρούσεων με την γνώση που επικρατούσε πριν.

Από τα διάφορα βήματα της KDD διαδικασίας αυτό που συγκεντρώνει το μεγαλύτερο ενδιαφέρον είναι το βήμα του data mining. Αυτό όμως, δε σημαίνει πως τα υπόλοιπα βήματα δεν είναι σημαντικά, αντιθέτως η επιτυχής τους διεκπεραίωση επηρεάζει την επιτυχία ολόκληρης της KDD διαδικασίας.



Σχήμα 1.2 Τα βήματα της KDD διαδικασίας.

1.4 Εξόρυξη γνώσης

Η εξόρυξη γνώσης *(data mining)* αναφέρεται στην εξαγωγή προτύπων από τα εξεταζόμενα δεδομένα ή στην προσαρμογή ήδη υπαρχόντων μοντέλων στα δεδομένα αυτά. Τα μοντέλα παίζουν το ρόλο της γνώσης που εξάγεται από το σύνολο των δεδομένων. Η απόφαση για το αν τα μοντέλα αυτά αντανακλούν ή όχι χρήσιμη γνώση είναι μέρος της συνολικής KDD διαδικασίας και συνήθως λαμβάνεται από κάποιον ανθρώπινο παράγοντα.

Οι βασικοί στόχοι του data mining είναι η πρόβλεψη και η περιγραφή.

- Η **πρόβλεψη** *(prediction)* αναφέρεται στην πρόβλεψη της τιμής κάποιου συγκεκριμένου γνωρίσματος ενός προβλήματος και αφορά νέα στιγμιότυπα του προβλήματος.

- Η **περιγραφή** *(description)* αναφέρεται στην εύρεση εύκολα διερμηνεύσιμων προτύπων από τα δεδομένα κάποιου προβλήματος. Τα πρότυπα αυτά θα πρέπει να αποτελούν στην ουσία συμπαγείς και περιεκτικές αναπαραστάσεις των δεδομένων του προβλήματος.

Σήμερα υπάρχει πληθώρα αλγορίθμων οι οποίοι προέρχονται από διαφορετικά επιστημονικά πεδία όπως: Στατιστική, Αναγνώριση Προτύπων, Μηχανική Μάθηση, Βάσεις Δεδομένων κ.α. Παρά τη διαφορετικότητα των πεδίων, οι αλγόριθμοι αυτοί χαρακτηρίζονται από τα ακόλουθα κοινά στοιχεία:

- **το μοντέλο**

Υπάρχουν δύο παράγοντες που σχετίζονται με το μοντέλο:

- *Η λειτουργία του μοντέλου*, η οποία καθορίζει τις βασικές εργασίες που θα διεκπεραιωθούν κατά τη διάρκεια του data mining, π.χ. ταξινόμηση, ομαδοποίηση κ.α.

- *Ο τύπος αναπαράστασης του μοντέλου*, ο οποίος καθορίζει τόσο την προσαρμοστικότητα του μοντέλου στην αναπαράσταση των δεδομένων όσο και τη δυνατότητα ερμηνείας του μοντέλου με όρους κατανοητούς από τον άνθρωπο. Τυπικά, τα πιο πολύπλοκα μοντέλα προσαρμόζονται καλύτερα στα δεδομένα, αλλά ενδέχεται να είναι πιο δύσκολο να γίνουν κατανοητά και να προσαρμοστούν σε πραγματικά δεδομένα. Οι πιο γνωστές αναπαραστάσεις μοντέλων είναι τα δέντρα απόφασης, οι κανόνες, τα γραμμικά μοντέλα, τα γραφικά μοντέλα που βασίζονται σε πιθανότητες, τα νευρωνικά δίκτυα κ.ο.κ.

- **την αξιολόγηση του μοντέλου**

Η αξιολόγηση, η οποία γίνεται βάσει κάποιων κριτηρίων αξιολόγησης (π.χ. *maximum likelihood*), καθορίζει κατά πόσο ένα συγκεκριμένο μοντέλο και οι παράμετροι του προσαρμόζονται στα κριτήρια της KDD διαδικασίας. Η αξιολόγηση ενός μοντέλου περιλαμβάνει τόσο την εκτίμηση της εγκυρότητας των προτύπων που παράγονται από αυτό όσο και την εκτίμηση της ακρίβειας, της χρησιμότητας και της ευκολίας κατανόησης του μοντέλου.

- **τον αλγόριθμο αναζήτησης**

Αναφέρεται στον καθορισμό ενός αλγορίθμου για την εύρεση συγκεκριμένων μοντέλων και παραμέτρων, με βάση ένα σύνολο δεδομένων, μια οικογένεια μοντέλων και ένα κριτήριο αξιολόγησης. Οι αλγόριθμοι αναζήτησης χωρίζονται σε δύο τύπους:

- *Αλγόριθμοι αναζήτησης παραμέτρων*, οι οποίοι αναζητούν τις παραμέτρους εκείνες που θα βελτιστοποιήσουν το μοντέλο ως προς το κριτήριο αξιολόγησης. Εκτελούν την αναζήτηση λαμβάνοντας ως είσοδο το σύνολο των δεδομένων και την αναπαράσταση του μοντέλου.

- *Αλγόριθμοι αναζήτησης μοντέλου,* οι οποίοι εκτελούν μια επαναληπτική διαδικασία αναζήτησης ενός μοντέλου για την αναπαράσταση των δεδομένων. Για μία συγκεκριμένη αναπαράσταση μοντέλου εκτελείται η μέθοδος αναζήτησης παραμέτρων και εκτιμάται η ποιότητα του συγκεκριμένου μοντέλου.

1.5 Βασικές εργασίες εξόρυξης γνώσης

Οι μέθοδοι που χρησιμοποιούνται για την επίτευξη των στόχων του data mining εκτελούν κατά την εφαρμογή τους ένα σύνολο από εργασίες, οι βασικότερες εκ των οποίων είναι οι ακόλουθες:

- Ταξινόμηση *(Classification)*
- Τμηματοποίηση *(Clustering)*
- Εξαγωγή κανόνων συσχέτισης *(ψation rules extraction)*
- Πρόβλεψη *(Prediction)*

Στη συνέχεια αναλύουμε κάθε επιμέρους εργασία και παραθέτουμε ενδεικτικά παραδείγματα για καλύτερη κατανόηση.

1.5.1 Ταξινόμηση *(Classification)*

Δοθέντων

- ενός προβλήματος με *N* κλάσεις: C_1, C_2, …C_n όπου κάθε στιγμιότυπο του προβλήματος έχει *m* ιδιότητες (γνωρίσματα): A_1, A_2,… A_m
- και ενός συνόλου στιγμιότυπων του προβλήματος για τα οποία γνωρίζουμε εκ των προτέρων σε ποια κλάση ανήκουν - το σύνολο αυτό είναι γνωστό ως **σύνολο εκπαιδευτικών στιγμιότυπων** *(training set),*

το ζητούμενο είναι

- η δημιουργία ενός μοντέλου για την ταξινόμηση νέων άγνωστων στιγμιότυπων του προβλήματος. Με τον όρο ταξινόμηση εννοούμε την τοποθέτηση ενός στιγμιότυπου σε μία από τις προκαθορισμένες κλάσεις του προβλήματος.

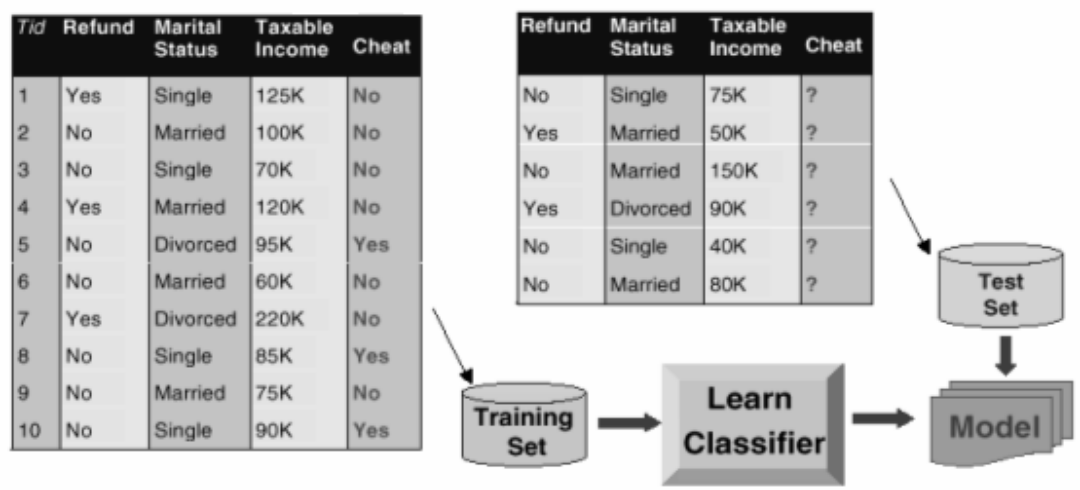
Η επιτυχής έκβαση της ταξινόμησης εξαρτάται από δύο βασικούς παράγοντες:

- *το σαφή καθορισμό των κλάσεων του προβλήματος* - οι κλάσεις είναι προκαθορισμένες και δεν μεταβάλλονται κατά τη διάρκεια της ταξινόμησης.
- *την «ποιότητα» του συνόλου των στιγμιότυπων εκπαίδευσης* – τα στιγμιότυπα αυτά θα πρέπει να είναι αντιπροσωπευτικά του προβλήματος.

Όπως έχουμε ήδη αναφέρει το σύνολο των εκπαιδευτικών στιγμιότυπων χρησιμοποιείται για την κατασκευή του ταξινομητή. Υπάρχει ωστόσο ένα ακόμη σύνολο στιγμιότυπων, το **σύνολο των στιγμιότυπων ελέγχου** *(test set)* βάσει του οποίου ελέγχεται η απόδοση του ταξινομητή. Με τον όρο απόδοση εννοούμε την ακρίβεια με την οποία ο ταξινομητής απαντά στο πρόβλημα της ταξινόμησης νέων άγνωστων στιγμιότυπων του προβλήματος. Η απόδοση ισούται με τον αριθμό των

στιγμιότυπων του συνόλου ελέγχου για τα οποία ο ταξινομητής προέβλεψε σωστά την κλάση προς το συνολικό αριθμό των στιγμιότυπων του συνόλου ελέγχου.

Στο ακόλουθο σχήμα (Σχήμα 1.3) διαφαίνεται ο ρόλος των δύο επιμέρους συνόλων.



Σχήμα 1.3 Η λειτουργία της ταξινόμησης

Στη συνέχεια παραθέτουμε κάποια ενδεικτικά παραδείγματα ταξινόμησης:

- **Παράδειγμα πολιτικής πίστωσης μιας τράπεζας**

Στην περίπτωση της πολιτικής πίστωσης, η τράπεζα θα ήθελε να γνωρίζει πότε μπορεί να δίνει δάνειο σε κάποιο πελάτη χωρίς να λαμβάνει μεγάλο ρίσκο.

Εφαρμόζοντας τη διαδικασία της ταξινόμησης το πρόβλημα ορίζεται ως εξής:

- *Οι πελάτες χωρίζονται στις κλάσεις: "άριστος", "καλός", "μέτριος" και "κακός" ανάλογα με την πιθανότητα κάθε πελάτη να εξοφλήσει το δάνειο.*
- *Κάθε πελάτης χαρακτηρίζεται από την ηλικία του, την εκπαίδευσή του, το ετήσιο εισόδημά του, κ.α.*
- *Έχουμε στη διάθεση μας δεδομένα πελατών που έχουν δανειστεί από την τράπεζα στο παρελθόν.*

Ένα ενδεχόμενο αποτέλεσμα της ταξινόμησης για τις κλάσεις "άριστος" και "καλός" θα μπορούσε να είναι το ακόλουθο:

Για κάθε πελάτη Ρ, με Ρ.πτυχίο = μεταπτυχιακό and Ρ.εισόδημα > 75,000 → Ρ.κλάση = άριστος

Για κάθε πελάτη Ρ, με Ρ.πτυχίο = πτυχίο πανεπιστημίου ή (Ρ.εισόδημα≥25,000 και Ρ.εισόδημα ≤ 75000) → Ρ.κλάση = καλός

Αξιοποιώντας το αποτέλεσμα μιας τέτοιας ταξινόμησης η τράπεζα αναμένεται να μειώσει το ρίσκο ο πελάτης στον οποίο χορήγησε κάποιο δάνειο να είναι ασυνεπής ως προς την εξόφληση του δανείου.

- **Παράδειγμα οργάνωσης διαφημιστικής καμπάνιας**

Στην περίπτωση της οργάνωσης μιας διαφημιστικής καμπάνιας, η εταιρία θα ήθελε να γνωρίζει ποιοι πελάτες είναι πιο πιθανό να απαντήσουν θετικά στην καμπάνια. Στόχος της εταιρίας είναι να προωθήσει την καμπάνια μόνο σε (πιθανά) ενδιαφερόμενα άτομα μειώνοντας έτσι το συνολικό κόστος.

Εφαρμόζοντας τη διαδικασία της ταξινόμησης το πρόβλημα ορίζεται ως εξής:

- *Οι πελάτες χωρίζονται στις κλάσεις: θετικοί και αρνητικοί αποδέκτες διαφημιστικών φυλλαδίων.*
- *Κάθε πελάτης χαρακτηρίζεται από το όνομά του, την ηλικία του, το επάγγελμά του κ.α.*
- *Έχουμε στη διάθεση μας δεδομένα πελατών που είχαν απαντήσει σε παλαιότερες διαφημιστικές καμπάνιες της εταιρίας.*

Ένα ενδεχόμενο αποτέλεσμα της ταξινόμησης για την κλάση των θετικών αποδεκτών διαφημιστικών φυλλαδίων θα μπορούσε να είναι το ακόλουθο:

Για κάθε πελάτη Ρ, με (Ρ.ηλικία>25 και Ρ.ηλικία<55) και Περιοχή = Ν. Προάστεια → Ρ.κλάση = θετικός αποδέκτης

Η εταιρία θα μπορούσε να αξιοποιήσει το αποτέλεσμα αποστέλλοντας το νέο διαφημιστικό υλικό μόνο στους θετικούς αποδέκτες μειώνοντας έτσι το συνολικό κόστος της διαφημιστικής καμπάνιας.

1.5.2 Τμηματοποίηση *(Clustering)*

Δοθέντων

- ενός προβλήματος όπου κάθε στιγμιότυπο του προβλήματος έχει *m* ιδιότητες (γνωρίσματα): $A_1, A_2, \ldots A_m$
- και ενός συνόλου στιγμιότυπων του προβλήματος

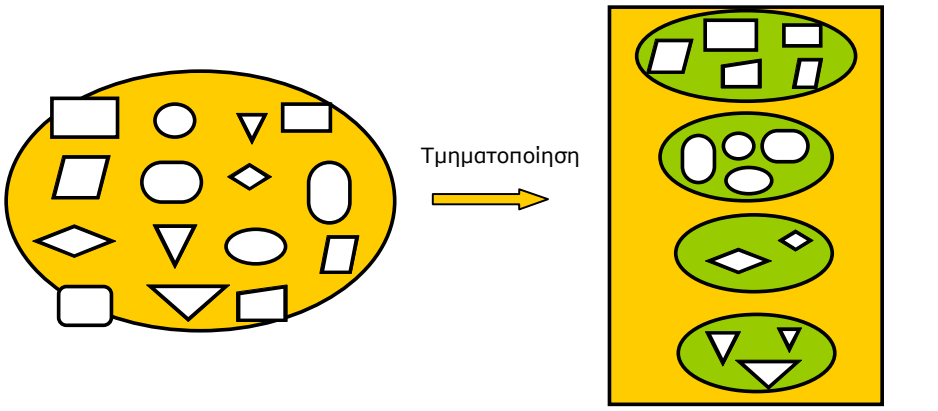
το ζητούμενο είναι

- ο διαχωρισμός των στιγμιότυπων του προβλήματος σε τμήματα *(clusters)*, έτσι ώστε στιγμιότυπα με παρόμοια χαρακτηριστικά να ανήκουν στο ίδιο τμήμα και

- η εύρεση του προφίλ κάθε τμήματος.

Η τμηματοποίηση είναι κατάλληλη για την εύρεση τμημάτων αντικειμένων με παρόμοια χαρακτηριστικά. Έτσι όταν θέλουμε να εξάγουμε κανόνες σχετικά με τη συμπεριφορά των αντικειμένων ενός συγκεκριμένου τμήματος δε χρειάζεται να εξετάσουμε τις ανεξάρτητες εγγραφές του συνόλου των δεδομένων, αρκεί να εξετάσουμε τα χαρακτηριστικά του συγκεκριμένου τμήματος. Η ιδέα / διαίσθηση είναι πως τα στοιχεία που ανήκουν στο ίδιο τμήμα θα συμπεριφέρονται με ενιαίο τρόπο, καθώς έχουν παρόμοια χαρακτηριστικά. Συνεπώς, ένας κανόνας που είναι έγκυρος για κάποιο από τα στοιχεία ενός τμήματος αναμένεται να είναι έγκυρος και για τα υπόλοιπα στοιχεία του τμήματος.

Στο ακόλουθο σχήμα (Σχήμα 1.4) παρατίθεται ένα γενικό παράδειγμα τμηματοποίησης των ετερογενών στιγμιότυπων ενός προβλήματος σε τμήματα με κοινά χαρακτηριστικά.



Σχήμα 1.4 Η λειτουργία της τμηματοποίησης

Η βασική διαφορά μεταξύ της ταξινόμησης και της τμηματοποίησης έγκειται στο γεγονός ότι στην ταξινόμηση οι κλάσεις είναι προκαθορισμένες, ενώ στην τμηματοποίηση δεν υπάρχουν προκαθορισμένες κλάσεις, τα στιγμιότυπα διασπόνται σε τμήματα βάσει της ομοιότητας που παρουσιάζουν μεταξύ τους ως προς τα γνωρίσματα της τμηματοποίησης. Συνεπώς, αν εφαρμόσουμε τμηματοποίηση σε ένα σύνολο δεδομένων, δεν υπάρχει κάποιο συγκεκριμένο σύνολο παραδειγμάτων το οποίο θα μπορούσε να μας υποδείξει ποιες είναι οι επιθυμητές σχέσεις που θα πρέπει να ισχύουν μεταξύ των δεδομένων.

Αρκετά συχνά η τμηματοποίηση χρησιμοποιείται και σαν πρώτο βήμα σε κάποια άλλη μορφή data mining εργασίας. Για παράδειγμα, μπορεί να χρησιμοποιηθεί σαν πρώτο βήμα στην προσπάθεια μερισμού της αγοράς. Αντί δηλαδή να προσπαθήσουμε να προσδιορίσουμε τι είδους διαφήμιση ταιριάζει καλύτερα σε κάθε πελάτη, μπορούμε να διασπάσουμε τους πελάτες σε τμήματα με βάση τις συνήθειές τους κατά την αγορά προϊόντων, να φτιάξουμε το προφίλ κάθε τμήματος και στη συνέχεια να προσδιορίσουμε το είδος της διαφήμισης που ταιριάζει καλύτερα στο κάθε τμήμα.

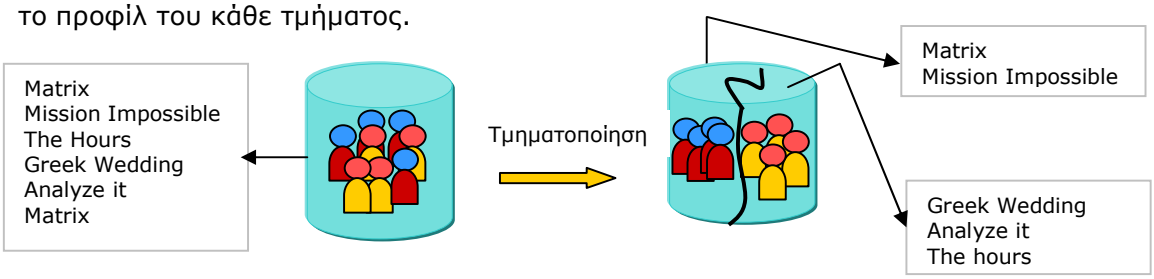
Στη συνέχεια παραθέτουμε κάποια ενδεικτικά παραδείγματα τμηματοποίησης:

- **Παράδειγμα τμηματοποίησης του πληθυσμού μιας χώρας σε διάφορα μορφωτικά επίπεδα**

Στην περίπτωση αυτή κάθε στιγμιότυπο του προβλήματος έχει ιδιότητες όπως ηλικία, τόπος κατοικίας, οικονομική κατάσταση, μόρφωση, κ.α., οπότε εφαρμόζοντας τμηματοποίηση μπορούμε να διασπάσουμε τον πληθυσμό με βάση τα γνωρίσματα αυτά και να βρούμε το προφίλ του κάθε τμήματος.

- **Παράδειγμα διαχωρισμού των χρηστών ενός δικτυακού τόπου με βάση τα κινηματογραφικά τους ενδιαφέροντα**

Στην περίπτωση αυτή κάθε στιγμιότυπο του προβλήματος έχει ιδιότητες όπως ηλικία, προηγούμενες προτιμήσεις σε ταινίες, επάγγελμα, μόρφωση, κ.α., οπότε μέσω της τμηματοποίησης μπορούμε να διασπάσουμε τους χρήστες σε τμήματα και να βρούμε το προφίλ του κάθε τμήματος.



Σχήμα 1.5 Τμηματοποίηση των χρηστών με βάση τα κινηματογραφικά τους ενδιαφέροντα

Έτσι όταν εμφανίζεται ένας νέος χρήστης, μπορούμε να βρούμε το πιο κοντινό στο χρήστη τμήμα χρηστών (ανάλογα με τις προηγούμενες προτιμήσεις του) και να του προτείνουμε ταινίες με βάση τις προτιμήσεις του τμήματος στο οποίο ανήκει. Η διαίσθησή μας είναι πως θα τον ενδιαφέρουν ταινίες που ενδιαφέρουν και τα υπόλοιπα μέλη του τμήματος.

1.5.3 Εξαγωγή κανόνων συσχέτισης *(Association rules)*

Οι κανόνες συσχέτισης είναι κατάλληλοι για την εύρεση συσχετίσεων μεταξύ διαφορετικών αντικειμένων. Ένας κανόνας συσχέτισης μεταξύ δύο αντικειμένων *A* και *B* δηλώνει πως η παρουσία του *A* σε κάποιο στιγμιότυπο του προβλήματος

συνεπάγεται και την παρουσία του *B* στο ίδιο στιγμιότυπο του προβλήματος και συμβολίζεται με *A → B.*

Η εξαγωγή των κανόνων συσχέτισης γίνεται με τη βοήθεια κάποιων αλγορίθμων, οι οποίοι αποδεικνύονται αρκετά αποδοτικοί. Μετά την ανάλυση και την εύρεση των κανόνων θα πρέπει να διαπιστωθεί κατά πόσο είναι έγκυροι και σημαντικοί για την εκάστοτε εφαρμογή. Για το σκοπό αυτό υπάρχουν δύο συντελεστές: η υποστήριξη *(support)* και η σιγουριά *(confidence)*.

- Η **υποστήριξη** *(support)* ισούται με το ποσοστό του συνόλου των στιγμιότυπων, έστω Ν το σύνολο των στιγμιότυπων, που ικανοποιεί το συνδυασμό Α και Β.

$$support = [AB]/N,$$

Έστω για παράδειγμα ο κανόνας συσχέτισης γ*άλα → κατσαβίδια*, αν υποθέσουμε πως μόνο το 0.001 όλων των αγορών περιλαμβάνει γάλα και κατσαβίδια, τότε η υποστήριξη του κανόνα συσχέτισης είναι χαμηλή. Συνήθως, οι επιχειρήσεις δεν ενδιαφέρονται για κανόνες με χαμηλή υποστήριξη, δεδομένου ότι αφορούν ένα πολύ μικρό ποσοστό των πελατών τους.

Από την άλλη αν το 50% των αγορών περιλαμβάνει γάλα και ψωμί, τότε η υποστήριξη για τον κανόνα συσχέτισης γ*άλα → ψωμί* είναι μεγάλη. Τέτοιοι κανόνες παρουσιάζουν ενδιαφέρον για τις επιχειρήσεις καθώς αφορούν ένα μεγάλο ποσοστό των πελατών.

- Η **σιγουριά** *(support)* ισούται με το ποσοστό του συνόλου των στιγμιότυπων για τα οποία όταν ισχύει το Α ισχύει και το Β.

$$confidence = [AB]/[A]$$

Για παράδειγμα, ο κανόνας συσχέτισης *ψωμί → γάλα* έχει μια σιγουριά 80% αν το 80% των αγορών που περιλαμβάνουν ψωμί περιλαμβάνει επίσης και γάλα. Για τις επιχειρήσεις ένας κανόνας με χαμηλή σιγουριά δεν παρουσιάζει ενδιαφέρον.

Να σημειώσουμε πως η σιγουριά του κανόνα *ψωμί → γάλα* μπορεί να διαφέρει από τη σιγουριά του κανόνα *γάλα → ψωμί*, παρόλο που και οι δύο κανόνες έχουν την ίδια υποστήριξη.

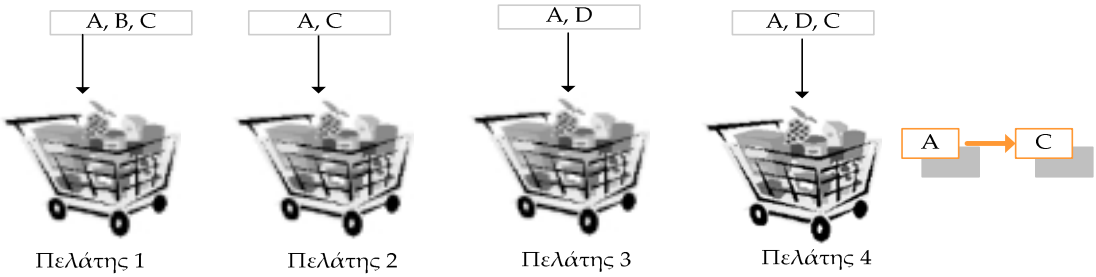
Στη συνέχεια παραθέτουμε κάποια ενδεικτικά παραδείγματα κανόνων συσχέτισης:

- **Παράδειγμα: σχεδιασμός καταλόγου σε καταστήματα**

Τα μαγαζιά λιανικής πώλησης ενδιαφέρονται να βρουν συσχετίσεις μεταξύ των διαφορετικών προϊόντων που αγοράζουν οι πελάτες του. Παραδείγματα τέτοιων συσχετίσεων θα μπορούσαν να είναι:

- Κάποιος που αγοράζει ψωμί είναι πολύ πιθανό να αγοράσει και γάλα (Σχήμα 1.6). Γνωρίζοντας ένα σουπερμάρκετ αυτόν τον κανόνα θα μπορούσε να τοποθετήσει σε

διπλανά ράφια το ψωμί και το γάλα, δεδομένου ότι τα δύο αυτά προϊόντα αγοράζονται συχνά μαζί.



Σχήμα 1.6 Παράδειγμα του κανόνα συσχέτισης A → C που μπορεί να προκύψει από τις αγορές των πελατών ενός σουπερμάρκετ.

- Κάποιος που αγοράζει το βιβλίο "Database System Concepts" είναι πολύ πιθανό να αγοράσει και το βιβλίο "Operating System Concepts". Γνωρίζοντας ένα online βιβλιοπωλείο αυτόν τον κανόνα θα μπορούσε να προτείνει το βιβλίο "Operating System Concepts" στους πελάτες του που αγοράζουν το βιβλίο "Database System Concepts", δεδομένου ότι τα δύο αυτά βιβλία αγοράζονται συχνά μαζί.

Ένας κανόνας συσχέτισης πρέπει να σχετίζεται με κάποιο πληθυσμό: ο πληθυσμός αυτός αποτελείται από ένα σύνολο στιγμιότυπων. Στην περίπτωση του σουπερμάρκετ για παράδειγμα, ο πληθυσμός προκύπτει από το ιστορικό των αγορών των πελατών του και κάθε στιγμιότυπο του προβλήματος αποτελείται από τα προϊόντα που αγόρασε ο πελάτης κατά της διάρκεια μιας αγοράς. Στην περίπτωση του online καταστήματος από την άλλη, ο πληθυσμός αποτελείται από όλους τους πελάτες του καταστήματος και κάθε στιγμιότυπο του προβλήματος περιλαμβάνει τις προτιμήσεις και τα προϊόντα που αγόρασε ο πελάτης καθ' όλη τη διάρκεια λειτουργίας του καταστήματος.

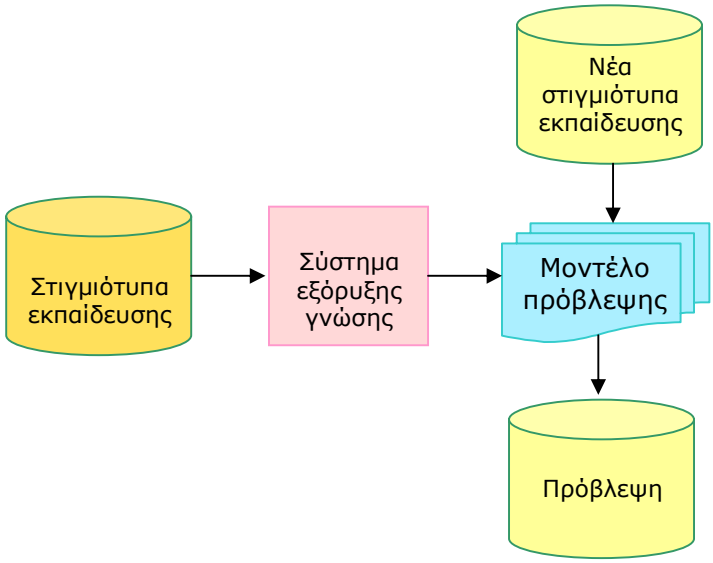
Παρατηρούμε λοιπόν ότι η έννοια του πληθυσμού καθορίζεται κάθε φορά από το πρόβλημα που καλούμαστε να αντιμετωπίσουμε. Έτσι στο πρώτο παράδειγμα επικεντρωνόμαστε στις επιμέρους αγορές ενός πελάτη ενώ στο δεύτερο επικεντρωνόμαστε στη συνολική αγοραστική εικόνα του πελάτη χωρίς να λαμβάνουμε υπόψη τα προϊόντα που αγοράστηκαν στις επιμέρους αγορές.

1.5.4 Πρόβλεψη *(Prediction)*

Δοθέντος ενός μοντέλου πρόβλεψης και ενός νέου στιγμιότυπου του προβλήματος, το ζητούμενο είναι η πρόβλεψη της τιμής ενός συγκεκριμένου γνωρίσματος του στιγμιότυπου αυτού (Σχήμα 1.7).

Το μοντέλο πρόβλεψης «χτίζεται» μέσω των παραδειγμάτων του συνόλου εκπαίδευσης (πρόκειται για παραδείγματα στα οποία η τιμή του προς πρόβλεψη

γνωρίσματος είναι γνωστή). Πέραν του συνόλου εκπαίδευσης υπάρχει και το σύνολο ελέγχου, το οποίο αποτελείται από παραδείγματα στα οποία η τιμή του προς πρόβλεψη γνωρίσματος είναι γνωστή – το σύνολο αυτό συνήθως ισούται αριθμητικά με το 1/3 των παραδειγμάτων του συνόλου εκπαίδευσης και χρησιμοποιείται για την αξιολόγηση του μοντέλου πρόβλεψης. Το σύνολο ελέγχου ελέγχει κατά κάποιο τρόπο την απόδοση του μοντέλου πρόβλεψης, δηλαδή την ακρίβεια με την οποία το μοντέλο πρόβλεψης προβλέπει την τιμή ενός άγνωστου γνωρίσματος στα νέα στιγμιότυπα του προβλήματος. Για να βρούμε την άγνωστη τιμή ενός γνωρίσματος σε κάποιο νέο στιγμιότυπο του προβλήματος, θα πρέπει να περάσουμε το στιγμιότυπο αυτό από το μοντέλο πρόβλεψης.



Σχήμα 1.7 Λειτουργία της πρόβλεψης

Συγκρίνοντας την πρόβλεψη με την ταξινόμηση που είδαμε στην αμέσως προηγούμενη ενότητα μπορούμε να πούμε πως η ταξινόμηση αποτελεί μια ειδική περίπτωση πρόβλεψης, καθώς αναφέρεται στην πρόβλεψη της κλάσης των στιγμιότυπων του προβλήματος.

Μερικά ενδεικτικά παραδείγματα πρόβλεψης είναι: πρόβλεψε την πιθανότητα ένας ασθενής να πάσχει από μία συγκεκριμένη ασθένεια, πρόβλεψε το πλήθος των αγορών που θα κάνει ένας νέος πελάτης στον πρώτο χρόνο κ.α.

1.6 Τεχνικές εξόρυξης γνώσης

Για την επιτυχή διεκπεραίωση των διαφόρων εργασιών data mining έχουν αναπτυχθεί πολλές τεχνικές. Κάποιες από τις πιο σημαντικές τεχνικές είναι οι ακόλουθες:

- Τα δέντρα απόφασης *(decision trees)*
- Τα νευρωνικά δίκτυα *(neural networks)*

- Οι Bayesian ταξινομητές
- Τμηματοποίηση διαμερισμού *(partitional clustering)*

Οι παραπάνω τεχνικές διαφέρουν ως προς την ακρίβεια και την κατανοησιμότητά τους. Στη συνέχεια αναλύουμε κάθε επιμέρους τεχνική.

1.6.1 Δέντρα απόφασης *(decision trees)*

Τα δέντρα απόφασης είναι πολύ ισχυρά εργαλεία που χρησιμοποιούνται ευρέως για τις περιπτώσεις της ταξινόμησης και της πρόβλεψης. Ένα δέντρο απόφασης αντιπροσωπεύει μια σειρά από IF THEN κανόνες ξεκινώντας από τη ρίζα του δέντρου και καταλήγοντας στα φύλλα του.

Οι εσωτερικοί κόμβοι ενός δέντρου απόφασης περιέχουν τα γνωρίσματα του προβλήματος, οι ακμές περιέχουν τις δυνατές τιμές των γνωρισμάτων και τα φύλλα περιέχουν τις πιθανές κλάσεις του προβλήματος. Απαραίτητο για την κατασκευή ενός δέντρου απόφασης είναι ένα σύνολο από στιγμιότυπα εκπαίδευσης, κάθε στιγμιότυπο του οποίου περιγράφεται από κάποια γνωρίσματα και την κλάση του προβλήματος στην οποία ανήκει.

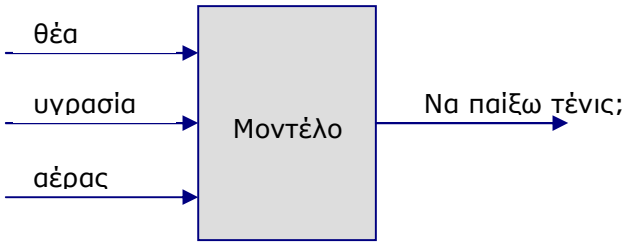
Η διαδικασία που ακολουθούν οι αλγόριθμοι κατασκευής ενός δέντρου απόφασης συνοψίζεται στα ακόλουθα: Ξεκινώντας από τη ρίζα του δέντρου ο αλγόριθμος διασπά το σύνολο των στιγμιότυπων εκπαίδευσης σε υποσύνολα με βάση τη βέλτιστη ιδιότητα *(best attribute)* του κόμβου – η βέλτιστη ιδιότητα ενός κόμβου καθορίζεται από κάποιο κριτήριο όπως το information gain, το gain ratio κ.ο.κ. Έτσι προκύπτει ένα πλήθος υποσυνόλων που το καθένα περιέχει λιγότερα παραδείγματα από το αρχικό σύνολο. Για καθένα απ' αυτά τα επιμέρους υποσύνολα εφαρμόζεται επαναληπτικά η παραπάνω διαδικασία χρησιμοποιώντας τα εναπομείναντα γνωρίσματα, οπότε η διάσπαση των στιγμιότυπων προχωρά και σταματά όταν όλα τα στιγμιότυπα του υποσυνόλου ανήκουν στην ίδια κλάση ή έχουν εξαντληθεί όλα τα γνωρίσματα. Στην ουσία πρόκειται για εφαρμογή της μεθόδου «Διαίρει και βασίλευε».

Εκτός από το σύνολο των στιγμιότυπων εκπαίδευσης υπάρχει και το σύνολο των στιγμιότυπων ελέγχου με βάση τα οποία ελέγχεται η απόδοση του δέντρου, δηλαδή η ακρίβεια με την οποία το κατασκευασμένο δέντρο απαντά στο πρόβλημα της ταξινόμησης. Στην περίπτωση αυτή δίνουμε ως είσοδο στο δέντρο τις τιμές των γνωρισμάτων του στιγμιότυπου ελέγχου και περιμένουμε ως απάντηση την τάξη του στιγμιότυπου. Το πλήθος των λανθασμένων απαντήσεων (δηλαδή τα στιγμιότυπα στα οποία το δέντρο απάντησε διαφορετική κλάση από την πραγματική) καθορίζει την ακρίβεια του δέντρου.

Η διαδικασία που ακολουθούμε προκειμένου να ταξινομήσουμε ένα νέο στιγμιότυπο του προβλήματος είναι η ακόλουθη: διατρέχουμε το δέντρο από τη ρίζα προς τα φύλλα ακολουθώντας τα κατάλληλα μονοπάτια. Κάθε φορά η επιλογή του μονοπατιού καθορίζεται εφαρμόζοντας τη συνθήκη ελέγχου κάθε κόμβου στις τιμές

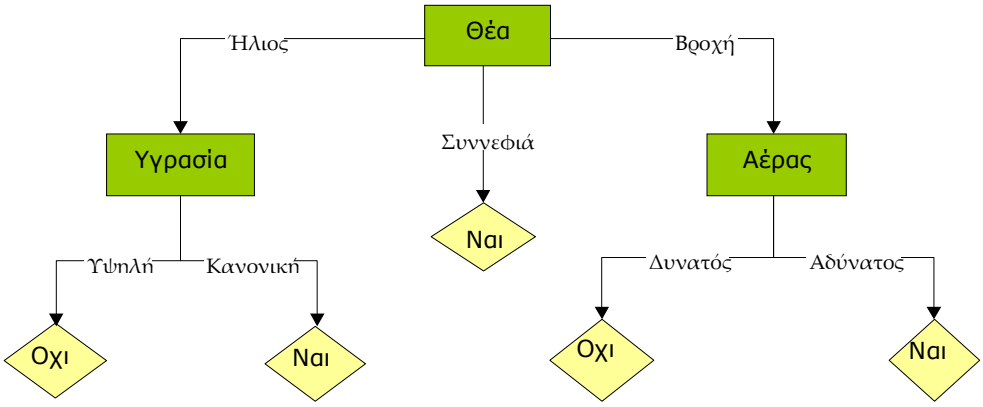
των γνωρισμάτων του προς ταξινόμηση στιγμιότυπου. Όταν καταλήξουμε σε κάποιο φύλλο, η κλάση αυτού είναι και η ζητούμενη κλάση του στιγμιότυπου.

Έστω για παράδειγμα το κλασσικό πρόβλημα που προσπαθεί να απαντήσει στο ερώτημα «Παίξε τένις» και το οποίο έχει δύο κλάσεις: «Ναι» και «Όχι» (Σχήμα 1.8). Η απάντηση στο πρόβλημα εξαρτάται από τους εξής παράγοντες: τη Θέα (με πιθανές τιμές: ήλιος, βροχή, συννεφιά), την Υγρασία (με πιθανές τιμές: υψηλή, κανονική) και τον Αέρα (με πιθανές τιμές :δυνατός, αδύνατος).



Σχήμα 1.8 Το πρόβλημα «Παίξε τένις»

Στο Σχήμα 1.9 φαίνεται το δέντρο απόφασης του προβλήματος. Περιέχει 3 εσωτερικούς κόμβους, σε κάθε κόμβο γίνεται έλεγχος ως προς κάποιο από τα γνωρίσματα του προβλήματος, ενώ στα φύλλα του περιέχονται οι κλάσεις του προβλήματος.



Σχήμα 1.9 Δέντρο απόφασης για το πρόβλημα «παίξε τένις»

Τα δέντρα απόφασης χρησιμοποιούνται ευρέως τόσο από την επιστημονική κοινότητα όσο και από τη βιομηχανία και αρκετοί αλγόριθμοι έχουν αναπτυχθεί για το σκοπό αυτό, όπως για παράδειγμα ο CART, ο ID3, ο C4.5, ο ITI κ.α.

Ένας από τους βασικούς λόγους για τους οποίους τα δέντρα απόφασης είναι τόσο δημοφιλή είναι, πέραν της ικανότητάς τους να απαντούν με ικανοποιητική ακρίβεια σε προβλήματα ταξινόμησης και πρόβλεψης, και η ευκολία με την οποία μπορούν να διατυπωθούν σε φυσική γλώσσα ή ακόμα και σε μια γλώσσα

προσπέλασης δεδομένων, όπως η SQL, έτσι ώστε να είναι εύκολα κατανοητά από τους ανθρώπους.

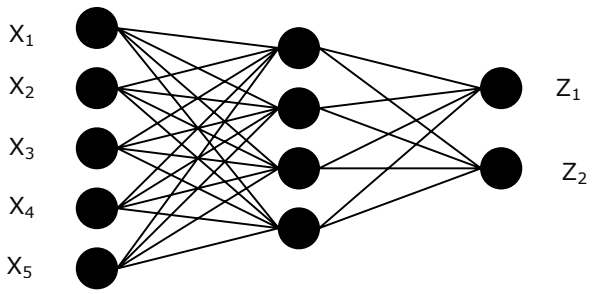
1.6.2 Νευρωνικά δίκτυα *(neural networks)*

Τα νευρωνικά δίκτυα αποτελούν μια πολύ δυνατή, γενικού σκοπού τεχνική η οποία μπορεί να εφαρμοστεί για πρόβλεψη, ταξινόμηση και τμηματοποίηση. Η ικανότητά τους να μαθαίνουν από τα δεδομένα μιμείται την ικανότητα των ανθρώπων να μαθαίνουν από τις εμπειρίες τους. Τα νευρωνικά δίκτυα αποτελούν μία προσέγγιση ανάπτυξης και εκτίμησης μαθηματικών δομών με δυνατότητα μάθησης και χρησιμοποιούνται για την εξαγωγή προτύπων και τον προσδιορισμό τάσεων οι οποίες είναι πολύ πολύπλοκες για να προσδιοριστούν από ανθρώπους ή από άλλες υπολογιστικές τεχνικές. Ένα εκπαιδευμένο νευρωνικό δίκτυο μπορεί να θεωρηθεί ως "ειδικός" για το πρόβλημα στο οποίο έχει εκπαιδευτεί και έτσι μπορεί να κάνει έγκυρες προβλέψεις για τα νέα στιγμιότυπα του προβλήματος.

Τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από στοιχεία επεξεργασίας (κόμβους) ανάλογους με τους νευρώνες στο ανθρώπινο μυαλό. Οι κόμβοι αυτοί διασυνδέονται μεταξύ τους σε ένα δίκτυο που μπορεί να αναγνωρίσει τα πρότυπα μόλις αυτά παρουσιαστούν μέσα σε ένα σύνολο δεδομένων. Δηλαδή, το δίκτυο μπορεί να μαθαίνει από την εμπειρία όπως ακριβώς κάνουν και οι άνθρωποι. Το σημείο αυτό διακρίνει τα νευρωνικά δίκτυα από τα παραδοσιακά προγράμματα υπολογιστών, τα οποία απλά ακολουθούν οδηγίες σύμφωνα με μια καλά ορισμένη σειρά.

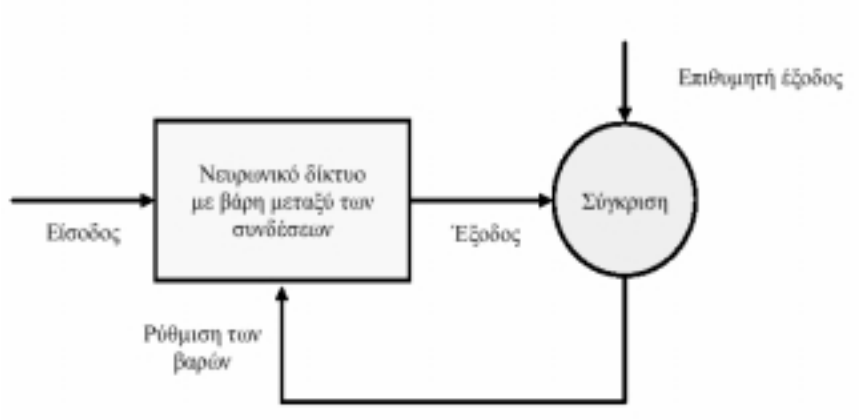
Η βασική μονάδα ενός νευρωνικού δικτύου είναι το perceptron, το οποίο παίρνει ως είσοδο ένα διάνυσμα πραγματικών τιμών, υπολογίζει ένα γραμμικό συνδυασμό των εισόδων και δίνει ως έξοδο 1 αν το αποτέλεσμα είναι μεγαλύτερο από κάποιο κατώφλι θ ή μηδέν διαφορετικά.

Τα νευρωνικά αποτελούνται από επιμέρους μονάδες που λειτουργούν παράλληλα (Σχήμα 2.10). Η συνάρτηση του δικτύου καθορίζεται ως επί το πλείστον από τις συνδέσεις μεταξύ των perceptrons. Μπορούμε να εκπαιδεύσουμε το νευρωνικό ώστε να εκτελεί μια συγκεκριμένη συνάρτηση ρυθμίζοντας τα βάρη μεταξύ των συνδέσεων.



Σχήμα 1.10 Η δομή ενός νευρωνικού δικτύου

Συνήθως τα νευρωνικά εκπαιδεύονται ώστε μια συγκεκριμένη είσοδος να οδηγεί σε μια συγκεκριμένη έξοδο, όπως φαίνεται στο Σχήμα 2.9. Στη συνέχεια το νευρωνικό ρυθμίζεται βάσει μιας σύγκρισης της τρέχουσας εξόδου με την επιθυμητή έξοδο, μέχρι να ταιριάξουν. Ο πιο δημοφιλής αλγόριθμος των νευρωνικών δικτύων είναι ο back propagation.



Σχήμα 1.11 Η λειτουργία ενός νευρωνικού δικτύου

Τα νευρωνικά δίκτυα είναι πολύ ισχυρά εργαλεία, με πολύ ικανοποιητική απόδοση ακόμη και σε μη κλασσικές περιπτώσεις data mining προβλημάτων. Επίσης, έχουν πολύ μεγάλη ανοχή σε ελλιπή δεδομένα ή δεδομένα με θόρυβο. Για το λόγο αυτό χρησιμοποιούνται ευρέως παρά το γεγονός ότι η εκπαίδευσή τους απαιτεί πολύ χρόνο και η ερμηνεία τους είναι δύσκολη - δεν είναι τόσο κατανοητά από τον άνθρωπο όσο για παράδειγμα τα δέντρα απόφασης.

1.6.3 Bayesian ταξινομητές

Πρόκειται για στατιστικούς ταξινομητές που μπορούν να υπολογίσουν την πιθανότητα ένα δοθέν στιγμιότυπο κάποιου προβλήματος να ανήκει σε μία από τις προκαθορισμένες κλάσεις του προβλήματος. Στηρίζονται στο θεώρημα του Bayes το οποίο διατυπώνεται ως εξής:

Έστω:

- *P* είναι η διαμοίραση πιθανότητας
- *D* είναι μια συλλογή στιγμιότυπων για τα οποία γνωρίζουμε την κλάση τους
- *h* είναι μια υπόθεση, όπως για παράδειγμα τα δεδομένα D να ανήκουν σε μία συγκεκριμένη κλάση *C*

Εάν γνωρίζουμε:

- *P(h)*, την a priori πιθανότητα η υπόθεση *h* να είναι σωστή

- *P(D)*, την πιθανότητα να παρατηρηθούν τα δεδομένα *D*
- *P(D|h)*, την posteriori πιθανότητα να παρατηρηθούν τα δεδομένα *D* με την προϋπόθεση ότι η υπόθεση *h* είναι σωστή

τότε το θεώρημα του Bayes προσφέρει μια μέθοδο υπολογισμού της posteriori πιθανότητας *P(h|D)*, δηλαδή της πιθανότητας να είναι σωστή η υπόθεση *h* δεδομένου ότι παρατηρούνται τα δεδομένα *D*. Δίνεται από τη σχέση:

$$P(h|D)=P(D|h)*P(h)/P(D)$$

Θεωρητικά οι Bayesian ταξινομητές έχουν το μικρότερο ρυθμό λάθους συγκρινόμενοι με τους υπόλοιπους ταξινομητές. Στην πράξη, ωστόσο, αυτό δεν ισχύει πάντα λόγω των σφαλμάτων που γίνονται στις υποθέσεις, όπως για παράδειγμα στην υπόθεση για την ανεξαρτησία ως προς την κατανομή των κλάσεων.

Επίσης, οι Bayesian ταξινομητές είναι χρήσιμοι και επειδή προσφέρουν μια θεωρητική αιτιολόγηση για άλλους ταξινομητές λόγω του θεωρήματος του Bayes. Για παράδειγμα, κάτω από συγκεκριμένες συνθήκες μπορεί να αποδειχθεί πως πολλά νευρωνικά δίκτυα παράγουν ως έξοδο την υπόθεση με τη μεγαλύτερη posteriori πιθανότητα όπως ακριβώς κάνουν και οι Bayesian ταξινομητές.

Οι Bayesian ταξινομητές παρουσιάζουν υψηλή απόδοση σε ακρίβεια και ταχύτητα όταν εφαρμόζονται σε μεγάλες βάσεις δεδομένων. Ένας πολύ απλός Bayesian ταξινομητής είναι ο Naïve Bayes ταξινομητής, η απόδοση του οποίου συγκρίνεται με την απόδοση ταξινομητών όπως τα δέντρα απόφασης και τα νευρωνικά δίκτυα.

1.6.3.1 Παράδειγμα ενός Bayesian ταξινομητή

Έστω ότι έχουμε στη διάθεσή μας στιγμιότυπα από τη βάση δεδομένων ενός καταστήματος με ηλεκτρονικά είδη (Πίνακας 1.1). Κάθε στιγμιότυπο αποτελείται από τα γνωρίσματα age, income, student, credit rating και ανήκει σε κάποια από τις δύο κλάσεις του προβλήματος (Yes, No).

RID	age	income	student	Credit_rating	Class: buys_computer
1	<=30	high	No	fair	No
2	<=30	high	No	excellent	No
3	31…40	high	No	fair	Yes
4	>40	medium	no	Fair	Yes
5	>40	low	yes	Fair	Yes
6	>40	low	yes	Excellent	No
7	31…40	low	yes	Excellent	Yes
8	<=30	medium	no	fair	No
9	<=30	low	yes	Fair	Yes
10	>40	medium	yes	fair	Yes

11	<=30	medium	yes	Excellent	Yes
12	31…40	medium	no	Excellent	Yes
13	31…40	high	yes	fair	Yes
14	>40	medium	no	Excellent	No

Πίνακας 1.1 Το σύνολο των στιγμιότυπων εκπαίδευσης του παραδείγματος

Έστω ότι θέλουμε να ταξινομήσουμε το ακόλουθο άγνωστο στιγμιότυπο του προβλήματος:

X=(age="<=30",income="medium",student="yes", credit_rating="fair")

Αρχικά υπολογίζουμε τις a priori πιθανότητες των δύο κλάσεων του προβλήματος:

P(buys_computer="yes") = 9/14 = 0.643

P(buys_computer="no") = 5/14 = 0.357

Στη συνέχεια, υπολογίζουμε τις υπό συνθήκη πιθανότητες για κάθε γνώρισμα για όλες τις κλάσεις του προβλήματος:

P(age="<=30"|buys_computer="yes") = 2/9 = 0.222

P(age="<=30"|buys_computer="no") = 3/5 = 0.600

P(income="medium"|buys_computer="yes") = 4/9 = 0.444

P(income="medium"|buys_computer="no") = 2/5 = 0.400

P(student="yes"|buys_computer="yes") = 6/9 = 0.667

P(student="yes"|buys_computer="no") = 1/5 = 0.200

P(credit_rating="fair"|buys_computer="yes") = 6/9 = 0.667

P(credit_rating="fair"|buys_computer="no") = 2/5 = 0.400

Χρησιμοποιώντας τις παραπάνω πιθανότητες υπολογίζουμε την πιθανότητα το άγνωστο στιγμιότυπο *X* να ανήκει σε κάποια από τις δύο κλάσεις του προβλήματος:

*P(X|buys_computer="yes") = 0.222 * 0.444 * 0.667 * 0.667=0.044*

*P(X|buys_computer="no") = 0.600 * 0.400 * 0.200 * 0.400=0.019*

*P(X|buys_computer="yes") * P(buys_computer="yes")=0.044*0.643=0.028*

*P(X|buys_computer="no") * P(buys_computer="no")=0.019 * 0.357=0.007*

Η κλάση με τη μεγαλύτερη πιθανότητα, εν προκειμένω η κλάση "Yes", είναι η απάντηση που ψάχνουμε.

1.6.3.2 Ο Απλός *(Naïve)* Bayes ταξινομητής

Πρόκειται για μια απλουστευμένη έκδοση του βασικού Bayes αλγορίθμου, η οποία χρησιμοποιείται για την ταξινόμηση των στιγμιότυπων ενός προβλήματος στις προκαθορισμένες κλάσεις του προβλήματος.

Η λειτουργία του Naive Bayes ταξινομητή συνοψίζεται στα ακόλουθα:

- Κάθε στιγμιότυπο X του προβλήματος αποτελείται από ένα σύνολο γνωρισμάτων $x_1,x_2,...x_n$, δηλαδή $X=<x_1,x_2,...x_n>$

- Έστω ότι το πρόβλημα έχει m κλάσεις, $c_1,c_2,...c_m$. Δοθέντος ενός άγνωστου στιγμιότυπου X του προβλήματος, για το οποίο δε γνωρίζουμε σε ποια κλάση ανήκει, ο ταξινομητής προβλέπει ότι το X ανήκει στην κλάση με τη μεγαλύτερη posteriori πιθανότητα. Ο ταξινομητής αναθέτει ένα άγνωστο στιγμιότυπο X του προβλήματος στην κλάση C_i αν και μόνο αν

  $$P(C_i|X) > P(C_j|X) \text{ για } 1 \le j \le m, j \ne i$$

  Έτσι μεγιστοποιείται η πιθανότητα $P(C_i|X)$, η οποία βάσει του θεωρήματος του Bayes δίνεται από τη σχέση:

  $$P(C_i|X)=P(X|C_i)*P(C_i)/P(X)$$

- Επειδή στον παραπάνω τύπο το $P(X)$ είναι σταθερό για όλα τα στιγμιότυπα, το μόνο που χρειάζεται να μεγιστοποιηθεί είναι η έκφραση $P(X|C_i)*P(C_i)$.

- Ο υπολογισμός του $P(X|C_i)$ είναι εξαιρετικά δαπανηρός και προκειμένου να μειώσουμε το υπολογιστικό κόστος υποθέτουμε ότι υπάρχει μια ανεξαρτησία ως προς την κατανομή των κλάσεων. Αυτό σημαίνει πως δοθείσας της κλάσης κάποιου στιγμιότυπου, οι τιμές των γνωρισμάτων του στιγμιότυπου είναι ανεξάρτητες μεταξύ τους, δηλαδή δεν υπάρχει εξάρτηση μεταξύ των γνωρισμάτων. Έτσι

  $$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

  Οι πιθανότητες $P(x_k|C_i)$ μπορούν να υπολογιστούν κατευθείαν από τα στιγμιότυπα εκπαίδευσης ως εξής:

  - Αν το γνώρισμα A_k είναι κατηγορηματικό *(categorical)*, τότε $P(x_k|C_i)=s_{ik}/s_i$, όπου s_{ik} είναι το πλήθος των στιγμιότυπων εκπαίδευσης της κλάσης C_i με

τιμή $A_k=x_k$ και s_i είναι το πλήθος των στιγμιότυπων εκπαίδευσης που ανήκουν στην κλάση C_i.

- Αν το γνώρισμα A_k είναι συνεχές, τότε υποθέτουμε ότι οι τιμές του ακολουθούν τη Gaussian κατανομή

$$P(x_k \mid C_i) = g(x_k, \mu_c, \sigma_c) = (1/\sqrt{2\pi\pi_{C_i}}) * e^{-(x_k - \mu_{C_i})/2\sigma_{C_i}^2}$$

όπου η συνάρτηση $g(x_k, \mu_c, \sigma_c)$ είναι η Gaussian συνάρτηση πυκνότητας για το γνώρισμα A_k.

- Προκειμένου να ταξινομήσουμε ένα νέο στιγμιότυπο X, υπολογίζουμε την πιθανότητα $P(X|C_i)*P(C_i)$ για κάθε κλάση C_i του προβλήματος. Το στιγμιότυπο X ανατίθεται στην κλάση C_i αν και μόνο αν

$$P(X|C_i)*P(C_i) > P(X|C_j)*P(C_j) \text{ για } 1 \le j \le m,\ j \ne i$$

Δηλαδή, το X ανατίθεται στην κλάση με τη μεγαλύτερη πιθανότητα.

1.6.4 Τμηματοποίηση διαμερισμού *(partitional clustering)*

Η τμηματοποίηση διαμερισμού αποτελεί ένα είδος τμηματοποίησης που βασίζεται στην άμεση αποσύνθεση του συνόλου των δεδομένων σε ένα σύνολο μη σχετιζόμενων τμημάτων *(clusters).* Το κριτήριο που εφαρμόζεται για την αποσύνθεση αυτή είναι η ελαχιστοποίηση κάποιων μέτρων ανομοιότητας μεταξύ των δειγμάτων μέσα σε κάθε ένα από τα τμήματα καθώς και η μεγιστοποίηση της ανομοιότητας μεταξύ διαφορετικών τμημάτων.

Μια από τις πιο συχνά χρησιμοποιούμενες μεθόδους της τμηματοποίησης διαμερισμού αποτελεί η μέθοδος K-Means, η οποία προσπαθεί να ελαχιστοποιήσει τη μέση τετραγωνική απόσταση των δεδομένων από τα πλησιέστερα κέντρα των τμημάτων.

$$E_k = \Sigma_k ||x_k - m_{c(xk)}||^2$$

όπου $c(x_k)$ είναι ο δείκτης του πλησιέστερου στο x_k κέντρου.

Ο αλγόριθμος K-Means χρησιμοποιεί σταθερό και δοθέντα εξ' αρχής αριθμό τμημάτων που θα δημιουργηθούν. Τα βήματά τού συνοψίζονται στα εξής:

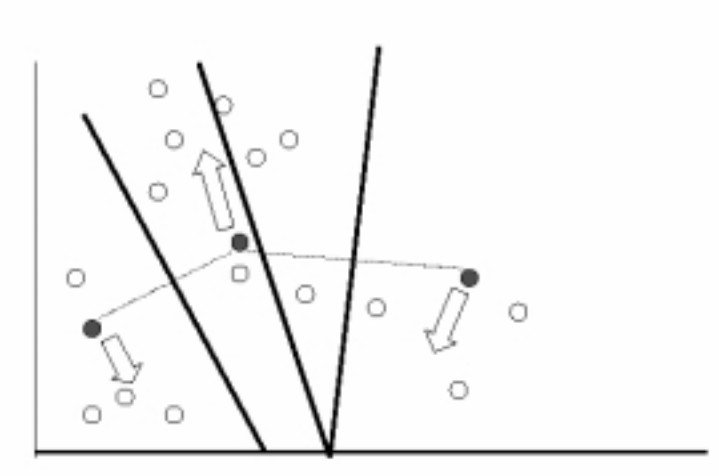
1. Θεώρησε ένα σύνολο από K σημεία-δεδομένα ως τα κέντρα των K τμημάτων (Σχήμα 1.12). Καθένα από τα κέντρα αντιπροσωπεύει ένα cluster.

2. Κάθε σημείο-δεδομένο αντιστοιχίζεται στο τμήμα του οποίου το κέντρο βρίσκεται πιο κοντά.

3. Υπολόγισε τα νέα κέντρα των τμημάτων χρησιμοποιώντας το μέσο όρο των σημείων τους.

4. Αντιστοίχισε κάθε σημείο-δεδομένο στο τμήμα του οποίου το κέντρο βρίσκεται πιο κοντά.

5. Επανάλαβε τα βήματα 3 και 4 έως ότου τα όρια των τμημάτων πάψουν να μεταβάλλονται ή η συνάρτηση *E* δεν μεταβάλλεται σημαντικά.

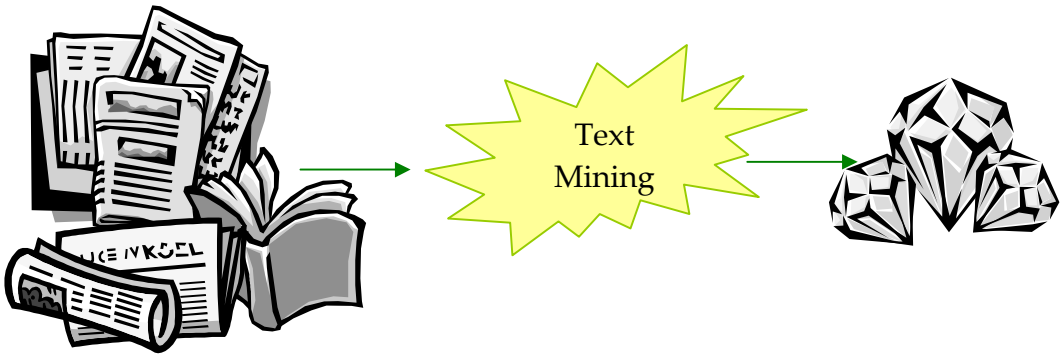


Σχήμα 1.12 Η λειτουργία της τμηματοποίησης διαμερισμού

2. Εξόρυξη Γνώσης από Κείμενα *(Text Mining)*

2.1 Εισαγωγή

Η Εξόρυξη Γνώσης *(Data Mining)* που είδαμε στο προηγούμενο κεφάλαιο περιλαμβάνει κυρίως τις διαδικασίες και τα μέσα εξαγωγής προτύπων από δομημένα δεδομένα (δηλαδή δεδομένα που είναι αποθηκευμένα σε βάσεις δεδομένων). Τα τελευταία, όμως, χρόνια με τη ραγδαία εξέλιξη του διαδικτύου και τη έκρηξη της πληροφορίας προέκυψε έντονη η ανάγκη εξόρυξης γνώσης από αδόμητα ή ημι-δομημένα δεδομένα όπως κείμενα, ιστοσελίδες, βιβλία, emails, chat rooms, εικόνες, έγγραφα κ.λ.π. Για το σκοπό αυτό δημιουργήθηκε ένας νέος κλάδος στην Εξόρυξη Γνώσης, η **Εξόρυξη Γνώσης από Κείμενα** *(Text Mining)* (Σχήμα 2.1). Η στροφή αυτή είναι πολύ σημαντική καθώς η πλειοψηφία των δεδομένων σήμερα (περίπου το 80% του όγκου των δεδομένων) διατίθενται με τη μορφή μη δομημένων κειμένων και το ποσοστό αυτό αναμένεται να αυξάνεται συνεχώς λόγω της εξάπλωσης της χρήσης του διαδικτύου και της ηλεκτρονικής επικοινωνίας.



Σχήμα 2.1 Η έννοια της εξόρυξης γνώσης από κείμενα

Τα δεδομένα που αποθηκεύονται σε μία βάση δεδομένων κειμένου είναι ημι-δομημένα. Για παράδειγμα ένα αρχείο κειμένου μπορεί να περιέχει λίγα δομημένα πεδία, π.χ. συγγραφέας, τίτλος, ημέρα δημοσίευσης κ.ο.κ., αλλά περιέχει επίσης και μεγάλα τμήματα μη δομημένου κειμένου, όπως για παράδειγμα η περίληψη και τα περιεχόμενά του.

Οι αρχικές προσπάθειες εξόρυξης γνώσης από κείμενα επικεντρώνονταν στην κατασκευή με το χέρι ενός συνόλου κανόνων. Για παράδειγμα, σε ένα πρόβλημα ταξινόμησης των αγγελιών μιας ιστοσελίδας σε διάφορες κατηγορίες ένας τέτοιος κανόνας θα μπορούσε να είναι "Αν το κείμενο περιέχει τη φράση expertise in Java" τότε η κατηγορία της αγγελίας είναι προγραμματιστής Η/Υ. Βέβαια η δημιουργία του πλήρους συνόλου των κανόνων απαιτεί πολύ καλή γνώση του πεδίου και αρκετούς ανθρώπινους πόρους σε κόπο και χρόνο. Ειδικά σήμερα που οι διαθέσιμες πληροφορίες είναι τόσο πολλές κάτι τέτοιο είναι σχεδόν αδύνατο.

Μια πιο αποδοτική προσέγγιση στο παραπάνω πρόβλημα είναι η χρήση Επιβλεπόμενης Μάθησης *(Supervised Learning)* για τη δημιουργία ενός ταξινομητή. Ο αλγόριθμος κατασκευής του ταξινομητή θα δέχεται ως είσοδο ένα σύνολο κειμένων για κάθε κλάση και θα βρίσκει μια αναπαράσταση ή κάποιους κανόνες για την ταξινόμηση νέων άγνωστων στιγμιότυπων του προβλήματος. Η προσέγγιση αυτή δημιουργεί ταξινομητές με αρκετά καλή απόδοση και με μικρότερο κόστος σε σχέση με την προηγούμενη προσέγγιση. Ένα από τα μειονεκτήματα αυτής της προσέγγισης αποτελεί το γεγονός ότι χρειάζεται ένα σύνολο στιγμιότυπων για τα οποία είναι γνωστή εκ των προτέρων η κλάση τους. Η ακρίβεια με την οποία τα στιγμιότυπα αυτά θα αποδοθούν σε κάθε κλάση επηρεάζει την απόδοση του ταξινομητή – όσο πιο καλό είναι το σύνολο εκπαίδευσης λοιπόν, τόσο καλύτερη θα είναι και η απόδοση του ταξινομητή.

2.2 Διαφορές εξόρυξης γνώσης και εξόρυξης γνώσης από κείμενα

Η εξόρυξη γνώσης από κείμενα μοιάζει πάρα πολύ με την κλασική εξόρυξης γνώσης καθώς και οι δύο σχετίζονται με τη διαχείριση γνώσης. Υπάρχει ωστόσο μία βασική διαφορά μεταξύ τους που αφορά τα δεδομένα που χρησιμοποιούνται σε κάθε περίπτωση.

Στην περίπτωση της εξόρυξης γνώσης στις βάσεις δεδομένων τα δεδομένα είναι δομημένα και περιγράφονται από μία συγκεκριμένη ενιαία δομή όπου κάθε στιγμιότυπο ενός προβλήματος ορίζεται μέσω ενός συγκεκριμένου και σταθερού συνόλου γνωρισμάτων. Αντιθέτως, στην περίπτωση της εξόρυξης γνώσης από κείμενα τα δεδομένα είναι ημι-δομημένα ή αδόμητα και δεν μπορούν να περιγραφούν μέσω κάποιου σταθερού συνόλου γνωρισμάτων.

Στη περίπτωση της εξόρυξης γνώσης από κείμενα, λοιπόν, υπάρχουν δύο βασικές προσεγγίσεις όσον αφορά την αναπαράσταση του κειμένου. Στην πρώτη προσέγγιση κρατάμε την πληροφορία (Ναι/ Όχι) για το αν κάποιο γνώρισμα (λέξη) εμφανίζεται σε ένα κείμενο. Έτσι, όταν παρουσιάζεται κάποιο νέο στιγμιότυπο του προβλήματος αυτό που ελέγχεται είναι η ύπαρξη των γνωρισμάτων (λέξεων) των στιγμιότυπων στις διάφορες κλάσεις του προβλήματος - η κλάση στην οποία εμφανίζονται οι περισσότερες λέξεις του κειμένου είναι η ζητούμενη κλάση.

Στη δεύτερη προσέγγιση για κάθε γνώρισμα κρατάμε τη συχνότητα εμφάνισής του σε ένα κείμενο. Έτσι η κλάση ενός νέου στιγμιότυπου απορρέει από τη συχνότητα εμφάνισης των λέξεων του κειμένου στις διάφορες κλάσεις του προβλήματος - η κλάση στην οποία εμφανίζονται οι περισσότερες και με μεγαλύτερη συχνότητα εμφάνισης λέξεις του κειμένου είναι η ζητούμενη κλάση.

2.3 Θέματα σχεδιασμού

Το βασικό θέμα κατά την εφαρμογή της εξόρυξης γνώσης σε κείμενα είναι η αναπαράσταση του κειμένου με τη μορφή γνωρισμάτων. Δοθείσας μιας συλλογής από δεδομένα κειμένου, οι περισσότερες τεχνικές εξόρυξης γνώσης από κείμενα εκτελούν εξόρυξη γνώσης στις λέξεις - κλειδιά που έχουν εξαχθεί από τα δεδομένα αυτά. Προκειμένου ωστόσο, η μέθοδος αυτή να είναι αποδοτική θα πρέπει να γίνει σωστή εξαγωγή των λέξεων – κλειδιά.

Στη συνέχεια θα παρουσιάσουμε κάποιες τέτοιες μεθόδους για εξόρυξη γνώσης από κείμενα.

2.3.1 Συσχέτιση με βάση τις λέξεις - κλειδιά

Η μέθοδος αυτή εντοπίζει τις επαναλαμβανόμενες λέξεις ή φράσεις κλειδιά ενός κειμένου και βρίσκει εν συνεχεία τη μεταξύ τους συσχέτιση.

Στο πρώτο στάδιο της μεθόδου γίνεται η προεπεξεργασία του κειμένου που περιλαμβάνει την ανάλυση του κειμένου, τον εντοπισμό της ρίζας των λέξεων και την απομάκρυνση ειδικών κατηγοριών λέξεων όπως άρθρα, σύνδεσμοι, αντωνυμίες κ.λ.π. Το επόμενο στάδιο είναι η εφαρμογή αλγορίθμων εξόρυξης γνώσης.

Σε μια βάση δεδομένων κειμένου μπορούμε να αντιμετωπίσουμε κάθε δεδομένο κειμένου ως μία δοσοληψία και τις λέξεις κλειδιά του κειμένου ως το σύνολο των αντικειμένων της δοσοληψίας. Δηλαδή, η βάση δεδομένων είναι της μορφής

{document_id, a_set_of_keywords}

Συνεπώς το πρόβλημα της συσχέτισης με βάση τις λέξεις - κλειδιά αντιστοιχίζεται στο κλασικό πρόβλημα εξαγωγής κανόνων συσχέτισης και μπορεί να επιλυθεί εφαρμόζοντας κάποιο σχετικό αλγόριθμο. Οι κανόνες συσχέτισης μπορούν να μας βοηθήσουν να βρούμε σύνθετες λέξεις, δηλαδή λέξεις που εμφανίζονται συνήθως μαζί, όπως για παράδειγμα “Stanford University”, “European Union” κ.λ.π. ή μη σύνθετες λέξεις, όπως για παράδειγμα [dollars, shares, exchange, total, commission, stake, securities].

Η μέθοδος που περιγράψαμε μπορεί να εφαρμοστεί είτε στο αρχικό κείμενο είτε στο σύνολο των λέξεων κλειδιών που εξάχθηκαν από το κείμενο. Τα βασικά της πλεονεκτήματα είναι δύο: (1) οι λέξεις και οι φράσεις κατηγοριοποιούνται αυτόματα χωρίς την παρεμβολή του ανθρώπινου παράγοντα και (2) το πλήθος των αποτελεσμάτων που δεν έχουν μεγάλη αξία μειώνεται σημαντικά.

2.3.2 Ταξινόμηση κειμένου

Η αυτόματη ταξινόμηση των κειμένων σε προκαθορισμένες κλάσεις / κατηγορίες αποτελεί μία από τις πιο σημαντικές εργασίες της εξόρυξης γνώσης από κείμενα,

ειδικά σήμερα που υπάρχει πληθώρα πληροφοριών σε μορφή κειμένου και η ανάγκη κατηγοριοποίησης αυτών των κειμένων είναι μεγάλη.

Το πρώτο στάδιο της μεθόδου είναι η εκπαίδευση του ταξινομητή μέσω ενός συνόλου κειμένων εκπαίδευσης. Εν συνεχεία ελέγχεται η απόδοση του ταξινομητή μέσω ενός συνόλου κειμένων ελέγχου. Για να ταξινομήσουμε ένα νέο αρχείο θα πρέπει να το περάσουμε από τον ταξινομητή ο οποίος και θα αποφανθεί για την κλάση / κατηγορία του.

Όπως βλέπουμε η διαδικασία είναι παρόμοια με αυτή της ταξινόμησης δομημένων δεδομένων. Υπάρχει ωστόσο μία βασική διαφορά. Τα δομημένα δεδομένα περιγράφονται από μία συγκεκριμένη ενιαία δομή - κάθε πλειάδα ορίζεται μέσω ενός συνόλου γνωρισμάτων. Για παράδειγμα, στην πλειάδα {sunny, warm, dry, not_windy, play_tennis}, η τιμή "sunny" αντιστοιχεί στο γνώρισμα weather_outlook, η τιμή "warm" αντιστοιχεί στο γνώρισμα temperature κ.ο.κ. Σε μια βάση δεδομένων κειμένου από την άλλη, δεν υπάρχει δομή με την έννοια που την χρησιμοποιήσαμε μόλις πριν. Δηλαδή το σύνολο των λέξεων – κλειδιών που προκύπτουν από ένα κείμενο δε μπορεί να αντιστοιχηθεί σε ένα σταθερό σύνολο γνωρισμάτων. Συνεπώς, για την περίπτωση των κειμένων δεν μπορούμε να χρησιμοποιήσουμε τους κλασικούς αλγόριθμους ταξινόμησης.

Μια αποδοτική μέθοδος ταξινόμησης είναι η ταξινόμηση με βάση τη συσχέτιση, η οποία στηρίζεται στην ταξινόμηση των κειμένων βάσει ενός συνόλου συχνά εμφανιζόμενων προτύπων. Τα βήματα που ακολουθεί η μέθοδος αυτή είναι τα ακόλουθα: Αρχικά εξάγονται οι λέξεις κλειδιά του συνόλου των κειμένων εκπαίδευσης. Στη συνέχεια εφαρμόζεται η μέθοδος της συσχέτισης με βάση τις λέξεις κλειδιά προκειμένου να εξαχθούν σχετιζόμενες λέξεις ή φράσεις που μπορούν να χρησιμοποιηθούν για να διαχωρίσουν (με όσο το δυνατόν μεγαλύτερη σιγουριά) τις κλάσεις μεταξύ τους. Στην ουσία, η μέθοδος αυτή παράγει ένα σύνολο από κανόνες συσχέτισης που σχετίζονται με κάθε κλάση του κειμένου. Οι κανόνες αυτοί μπορούν να χρησιμοποιηθούν για την ταξινόμηση νέων κειμένων λαμβάνοντας υπόψη και τη συχνότητα εμφάνισής τους.

2.4 Ταξινόμηση κειμένου με βάση το Naïve Bayes ταξινομητή

Ο Naive Bayes ταξινομητής, τον οποίο είδαμε να εφαρμόζεται σε κλασικά προβλήματα εξόρυξης γνώσης, μπορεί να εφαρμοστεί και στην περίπτωση της εξόρυξης γνώσης από κείμενα.

Σε μια τέτοια περίπτωση, κάθε κείμενο αποτελεί ένα στιγμιότυπο του προβλήματος και οι επιμέρους λέξεις του κειμένου αποτελούν τα γνωρίσματα του στιγμιότυπου. Η πιθανότητα ένα άγνωστο στιγμιότυπο X να ανήκει σε κάποια από τις προκαθορισμένες κλάσεις C_i του προβλήματος δίνεται από τη σχέση:

$$P(X \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i)$$

Η κλάση για την οποία μεγιστοποιείται η έκφραση $P(C_i) * \prod_{k=1}^{n} P(X_k \mid C_i)$ είναι η κλάση που ψάχνουμε. Στην ουσία, ο αλγόριθμος αποφαίνεται για την κλάση κάποιου νέου στιγμιότυπου (κειμένου) ελέγχοντας τη συχνότητα εμφάνισης των γνωρισμάτων (λέξεων) του στιγμιότυπου στο σύνολο των στιγμιότυπων εκπαίδευσης του προβλήματος.

Στη συνέχεια παραθέτουμε τον ψευδοκώδικα του αλγορίθμου (μέσω δύο συναρτήσεων). Η πρώτη συνάρτηση κατασκευάζει τον ταξινομητή μέσω του συνόλου των παραδειγμάτων εκπαίδευσης, ενώ η δεύτερη χρησιμοποιεί τον ταξινομητή για να προβλέψει την κλάση κάποιου νέου άγνωστου στιγμιότυπου του προβλήματος.

- **LEARN_NAIVE_BAYES_TEXT (Examples,V)**

Είσοδος

Examples: ένα σύνολο κειμένων για τα οποία γνωρίζουμε την κλάση τους.

V: το σύνολο όλων των πιθανών κλάσεων του προβλήματος.

Έξοδος

$P(w_k|v_j)$: η πιθανότητα η λέξη w_k να ανήκει στην κλάση v_j.

$P(v_j)$: η πιθανότητα εμφάνισης της κλάσης v_j.

Τα βήματα του αλγορίθμου συνοψίζονται στα ακόλουθα:

1) Συγκέντρωσε όλες τις λέξεις, τα σημεία στίξης και τις φράσεις που εμφανίζονται στο σύνολο *Examples*.

- Φτιάξε το λεξικό *(Vocabulary)*, το οποίο αποτελείται από τις διακριτές λέξεις και φράσεις που εμφανίζονται στα κείμενα του συνόλου *Examples*.

2) Υπολόγισε τις πιθανότητες $P(v_j)$ και $P(w_k|v_j)$ επαναλαμβάνοντας για κάθε κλάση τα ακόλουθα:

- Βρες το υποσύνολο $docs_j$ του συνόλου *Examples* τα κείμενα του οποίου έχουν ως κλάση την v_j.
- Υπολόγισε την πιθανότητα εμφάνισης της κλάσης v_j - ισούται με $|docs_j|/|Examples|$

- Βρες τη συνένωση $Text_j$ όλων των λέξεων και φράσεων των επιμέρους κειμένων του υποσυνόλου $docs_j$.
- Υπολόγισε το πλήθος *n* των λέξεων και φράσεων του $Text_j$
- Επανάλαβε για κάθε λέξη w_k του λεξικού
  - Βρες πόσες φορές η λέξη w_k εμφανίστηκε στο κείμενο $Text_j$, έστω n_k.
  - Υπολόγισε την πιθανότητα $P(w_k|v_j) = n_k+1/n + |Vocabulary|$

➢ **CLASSIFY_NAIVE_BAYES_TEXT (Doc)**

Είσοδος

Doc: το κείμενο που θέλουμε να ταξινομήσουμε.

Έξοδος

Η κλάση που προέβλεψε ο ταξινομητής για το κείμενο *Doc*.

Τα βήματα του αλγορίθμου συνοψίζονται στα ακόλουθα:

1) Βρες όλες τις λέξεις, φράσεις του κειμένου *Doc* που υπάρχουν στο Vocabulary (έστω *Positions*)

2) Επέστρεψε το V_{NB} που δίνεται από τη σχέση:

$$V_{NB} = argmax_{vj \in V} P(v_j) * \Pi_{i \in positions} P(w_i|v_j)$$

Ο αλγόριθμος που περιγράψαμε μόλις πριν χρησιμοποιήθηκε για την ταξινόμηση των ειδησεογραφικών άρθρων σε newsgroups. Στόχος ήταν η αυτόματη ανάθεση ενός άρθρου σε κάποιο newsgroup. Το πρόβλημα περιείχε 20 κλάσεις (όσα και τα newsgroups) και για κάθε κλάση χρησιμοποιήθηκαν 1000 άρθρα. Συνολικά, λοιπόν, συγκεντρώθηκαν 20.000 άρθρα εκ των οποίων τα 2/3 αποτέλεσαν το σύνολο των στιγμιότυπων εκπαίδευσης και τα υπόλοιπα το σύνολο των στιγμιότυπων ελέγχου. Η απόδοση του παραπάνω αλγορίθμου ήταν 89% ενώ η τυχαία πρόβλεψη είχε απόδοση μόλις 5%. Ο συγκεκριμένος αλγόριθμος χρησιμοποιήθηκε και στα πλαίσια της εφαρμογής μας.

ΜΕΡΟΣ Β

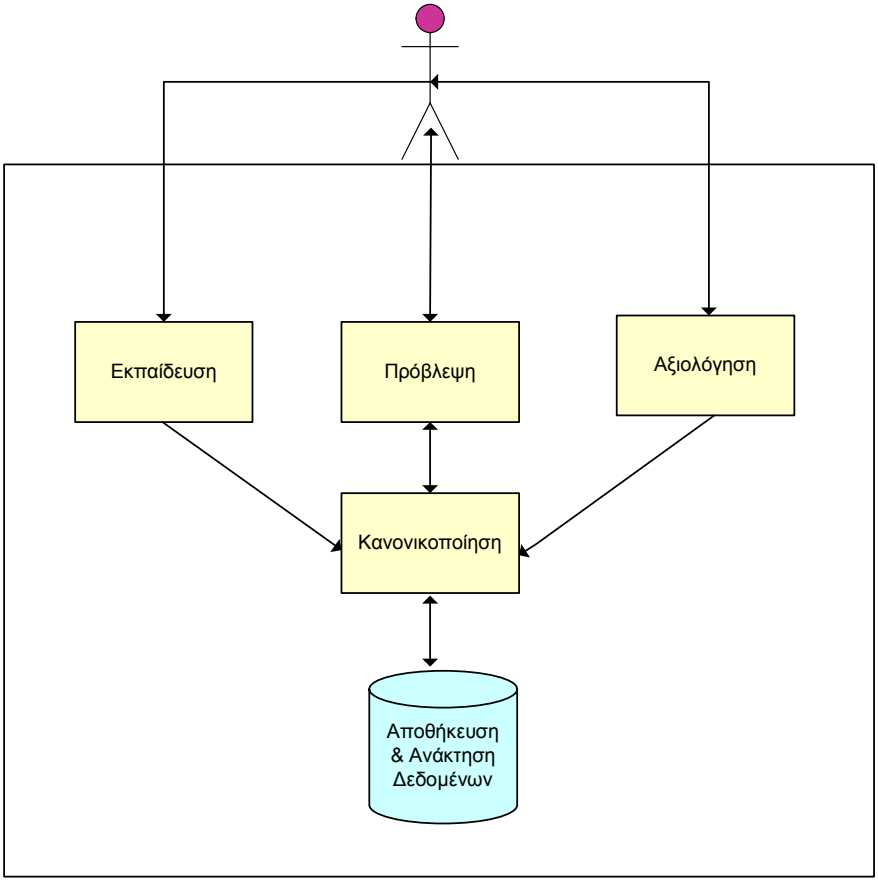
Το σύστημα TMPredictor (Text Mining Predictor)

1. Περιγραφή του συστήματος

Στο συγκεκριμένο κεφάλαιο παρουσιάζεται η αρχιτεκτονική του συστήματος[1] και γίνεται λεπτομερής ανάλυση των επιμέρους δομικών του στοιχείων.

1.1 Αρχιτεκτονική του συστήματος

Η αρχιτεκτονική του συστήματος απεικονίζεται στο επόμενο σχήμα (Σχήμα 1.1).



Σχήμα 1.1 Η αρχιτεκτονική του συστήματος

Όπως φαίνεται και από το παραπάνω σχήμα, το σύστημα αποτελείται από τα εξής πέντε δομικά τμήματα:

- **Εκπαίδευση**: Το τμήμα της εκπαίδευσης αφορά το "χτίσιμο" του ταξινομητή κάποιου συγκεκριμένου προβλήματος με βάση τα στιγμιότυπα του συνόλου εκπαίδευσης του προβλήματος.

[1] Για λόγους συντομίας το σύστημα θα αναφέρεται ως TMPredictor (Text Mining Predictor) για το υπόλοιπο του κειμένου

- **Πρόβλεψη**: Το τμήμα της πρόβλεψης αφορά την εύρεση της κλάσης νέων άγνωστων στιγμιότυπων κάποιου συγκεκριμένου προβλήματος.

- **Αξιολόγηση**: Στο τμήμα της αξιολόγησης ελέγχεται η απόδοση του ταξινομητή για κάποιο συγκεκριμένο πρόβλημα.

- **Κανονικοποίηση**:Το τμήμα της κανονικοποίησης αφορά τη μετατροπή των ρημάτων στο πρώτο ενικό πρόσωπο και των ουσιαστικών και επιθέτων στην ονομαστική πτώση ενικού αριθμού. Η κανονικοποίηση αφορά την ελληνική γλώσσα και επιτυγχάνεται μέσω του προγράμματος Normalizer. Στα πλαίσια της εφαρμογής μας χρησιμοποιούμε την κανονικοποίηση προκειμένου να μειώσουμε το πλήθος των λέξεων των διαφόρων στιγμιότυπων του προς επίλυση προβλήματος.

- **Αποθήκευση & Ανάκτηση δεδομένων**: Το τμήμα αυτό αφορά την αποθήκευση και ανάκτηση των απαραίτητων δεδομένων εισόδου για τα υπόλοιπα τμήματα του συστήματος καθώς επίσης και των αποτελεσμάτων που επιστρέφονται απ’ αυτά. Για το σκοπό χρησιμοποιείται μια βάση δεδομένων υλοποιημένη στο Σύστημα Διαχείρισης Βάσεων Δεδομένων Microsoft SQL Server 2000.

Στη συνέχεια αναλύουμε κάθε επιμέρους τμήμα του συστήματος.

1.2 Εκπαίδευση

Στη φάση της εκπαίδευσης “χτίζεται” το λεξικό του προβλήματος το οποίο περιλαμβάνει τις διακριτές σημαντικές λέξεις των παραδειγμάτων του συνόλου εκπαίδευσης κατανεμημένες στις διάφορες κλάσεις του προβλήματος. Το λεξικό περιλαμβάνει επίσης και τη συχνότητα εμφάνισης των λέξεων στις διάφορες κλάσεις του προβλήματος.

Η εκπαίδευση συνίσταται σε δύο κυρίως βήματα (για κάθε στιγμιότυπο του συνόλου εκπαίδευσης):

1. στην εύρεση της κλάσης του στιγμιότυπου εκπαίδευσης και

2. στον εμπλουτισμό του λεξικού της εφαρμογής με τις επιμέρους σημαντικές λέξεις του στιγμιότυπου εκπαίδευσης.

Όσον αφορά το πρώτο βήμα η κλάση του στιγμιότυπου εκπαίδευσης προκύπτει από το κείμενο της είδησης.

Όσον αφορά το δεύτερο βήμα η διαδικασία που ακολουθούμε είναι η ακόλουθη:

- κανονικοποιούμε το κείμενο μέσω του Normalizer, έτσι απομακρύνονται όλες οι μη σημαντικές λέξεις (π.χ. άρθρα, σύνδεσμοι κ.α.) και μετατρέπονται όλες οι λέξεις του κειμένου είτε στο πρώτο ενικό πρόσωπο οριστικής, εάν πρόκειται για ρήματα, είτε στην ονομαστική ενικού στην περίπτωση των ουσιαστικών, επιθέτων.

- αναλύουμε το κανονικοποιημένο κείμενο στις επιμέρους σημαντικές λέξεις που το αποτελούν.

- Για κάθε λέξη του κειμένου

    - εκτελούμε ένα ερώτημα στη βάση προκειμένου να ελέγξουμε αν υπάρχει ήδη εγγραφή για τη λέξη αυτή και την κλάση στην οποία ανήκει το στιγμιότυπο εκπαίδευσης. Ο πίνακας που χρησιμοποιούμε στην περίπτωση αυτή είναι ο dm_ vocabulary.

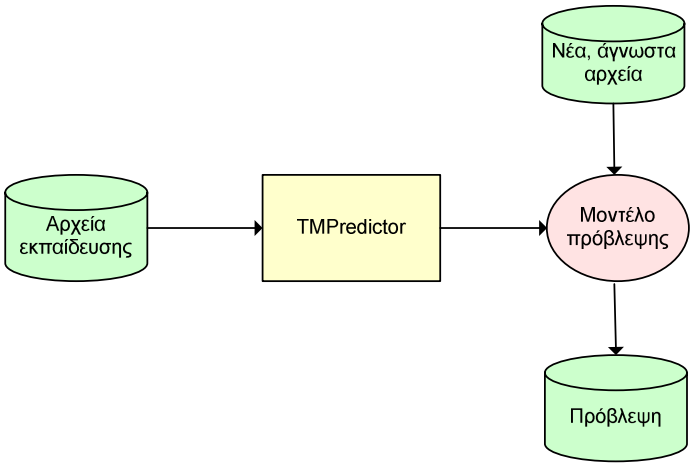
        - Αν δεν υπάρχει εγγραφή, εισάγουμε μια νέα εγγραφή στον πίνακα dm_vocabulary για τη λέξη αυτή αρχικοποιώντας με 1 το μετρητή των εμφανίσεων της λέξης στην εν λόγω κλάση,

        - ειδάλλως, ενημερώνουμε την ήδη υπάρχουσα εγγραφή αυξάνοντας κατά 1 το μετρητή των εμφανίσεων της λέξης στην εν λόγω κλάση.

Από τα παραπάνω φαίνεται πως το "χτίσιμο" του ταξινομητή του προβλήματος αποτελεί μια αυξητική *(incremental)* διαδικασία.

1.3 Πρόβλεψη

Στη φάση της πρόβλεψης προβλέπεται η κλάση ενός νέου άγνωστου στιγμιότυπου του προβλήματος με βάση το λεξικό που έχει δημιουργηθεί στη φάση της εκπαίδευσης. Όπως έχουμε ήδη αναφέρει η πρόβλεψη στηρίζεται στην πιθανότητα εμφάνισης του νέου στιγμιότυπου στις διάφορες κλάσεις του προβλήματος – η κλάση με τη μεγαλύτερη πιθανότητα είναι η ζητούμενη (Σχήμα 1.7).



Σχήμα 1.7 Η λειτουργία της πρόβλεψης

Η είσοδος του τμήματος της πρόβλεψης είναι το κείμενο της είδησης για το οποίο θέλουμε να προβλέψουμε την κλάση και η έξοδος του είναι η προβλεπόμενη κλάση. Τα βήματα που ακολουθούμε για την πρόβλεψη της κλάσης ενός νέου άγνωστου στιγμιότυπου είναι τα ακόλουθα:

- κανονικοποιούμε το κείμενο μέσω του Normalizer, έτσι απομακρύνονται όλες οι μη σημαντικές λέξεις (π.χ. άρθρα, σύνδεσμοι κ.α.) και μετατρέπονται όλες οι λέξεις του κειμένου είτε στο πρώτο ενικό πρόσωπο οριστικής, εάν πρόκειται για ρήματα, είτε στην ονομαστική ενικού στην περίπτωση των ουσιαστικών, επιθέτων.

- αναλύουμε το κανονικοποιημένο κείμενο στις επιμέρους σημαντικές λέξεις που το αποτελούν.

- Για κάθε κλάση *c* του προβλήματος επαναλαμβάνουμε τα εξής:

    - Βρίσκουμε την πιθανότητα εμφάνισης της κλάσης *c,* έστω *P(c)*. Η πιθανότητα αυτή ισούται με το πλήθος των στιγμιότυπων εκπαίδευσης που ανήκουν στην κλάση *c* προς το συνολικό πλήθος των στιγμιότυπων που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου πρόβλεψης του προβλήματος.

    - Για κάθε λέξη w_i του κειμένου επαναλαμβάνουμε τα εξής:

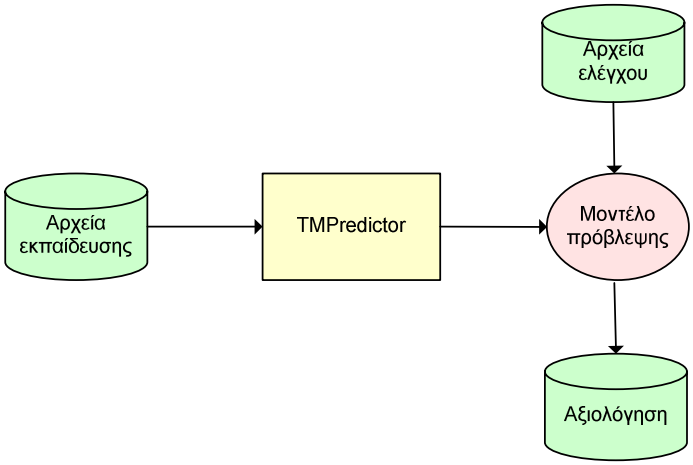
        - βρίσκουμε την πιθανότητα εμφάνισής της w_i στην κλάση *c,* έστω Prob(w_i|c). Η πιθανότητα αυτή ισούται με τον πλήθος των εμφανίσεων της λέξης w_i στην κλάση *c* συν ένα προς το πλήθος των λέξεων που εμφανίστηκαν στην κλάση c συν το μέγεθος του λεξικού.

    - Υπολογίζουμε για την κλάση *c* την πιθανότητα $\prod_{i \in \{1,\ldots,n\}} \Pr ob(w_i \mid c)$

- Από όλες τις κλάσεις επιλέγουμε εκείνη με τη μεγαλύτερη πιθανότητα, αυτή είναι και η προβλεπόμενη κλάση για το άγνωστο στιγμιότυπο του προβλήματος.

1.4 Αξιολόγηση

Στη φάση της αξιολόγησης ελέγχεται η απόδοση και η αξιοπιστία του μοντέλου στην πρόβλεψη της κλάσης νέων, άγνωστων στιγμιότυπων του προβλήματος. Η είσοδος αυτής της φάσης είναι το σύνολο των στιγμιότυπων ελέγχου του προβλήματος (Σχήμα 1.8).



Σχήμα 1.8 Η λειτουργία της αξιολόγησης

Η αξιολόγηση στηρίζεται στα στιγμιότυπα του συνόλου ελέγχου – πρόκειται για στιγμιότυπα που έχουν ακριβώς την ίδια δομή με τα στιγμιότυπα εκπαίδευσης.

Η αξιολόγηση (για κάθε αρχείο του συνόλου ελέγχου) περιλαμβάνει τα εξής βήματα:

- την εύρεση της πραγματικής κλάσης του αρχείου ελέγχου - η κλάση αυτή είναι γνωστή ως πραγματική κλάση.
- την πρόβλεψη της κλάσης του αρχείου ελέγχου - η κλάση αυτή είναι γνωστή ως προβλεπόμενη κλάση.
- τη σύγκριση της προβλεπόμενης κλάσης με την πραγματική κλάση του αρχείου.

Όσον αφορά την εύρεση της πραγματικής κλάσης ενός στιγμιότυπου ελέγχου, αυτή προκύπτει από το κείμενο της είδησης.

Όσον αφορά την εύρεση της προβλεπόμενης κλάσης ενός στιγμιότυπου ελέγχου τα βήματα που ακολουθούμε είναι τα ακόλουθα:

- κανονικοποιούμε το κείμενο μέσω του Normalizer, έτσι απομακρύνονται όλες οι μη σημαντικές λέξεις (π.χ. άρθρα, σύνδεσμοι κ.α.) και μετατρέπονται όλες οι λέξεις του κειμένου είτε στο πρώτο ενικό πρόσωπο οριστικής, εάν πρόκειται για ρήματα, είτε στην ονομαστική ενικού στην περίπτωση των ουσιαστικών, επιθέτων.

- αναλύουμε το κανονικοποιημένο κείμενο στις επιμέρους σημαντικές λέξεις που το αποτελούν.

- Για κάθε κλάση *c* του προβλήματος επαναλαμβάνουμε τα εξής:

    - Βρίσκουμε την πιθανότητα εμφάνισης της κλάσης *c,* έστω *P(c)*. Η πιθανότητα αυτή ισούται με το πλήθος των αρχείων εκπαίδευσης που ανήκουν στην κλάση *c* προς το συνολικό πλήθος των αρχείων που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου πρόβλεψης.

    - Για κάθε λέξη w_i του κειμένου επαναλαμβάνουμε τα εξής:

        - βρίσκουμε την πιθανότητα εμφάνισής της w_i στην κλάση *c,* έστω Prob(w_i|c). Η πιθανότητα αυτή ισούται με τον πλήθος των εμφανίσεων της λέξης w_i στην κλάση *c* συν ένα προς το πλήθος των λέξεων που εμφανίστηκαν στην κλάση c συν το μέγεθος του λεξικού.

    - Υπολογίζουμε για την κλάση *c* την πιθανότητα $\prod_{i \in \{1,\ldots,n\}} \Pr ob(w_i \mid c)$

- Από όλες τις κλάσεις επιλέγουμε εκείνη με τη μεγαλύτερη πιθανότητα, αυτή είναι και η προβλεπόμενη κλάση για το αρχείο ελέγχου.

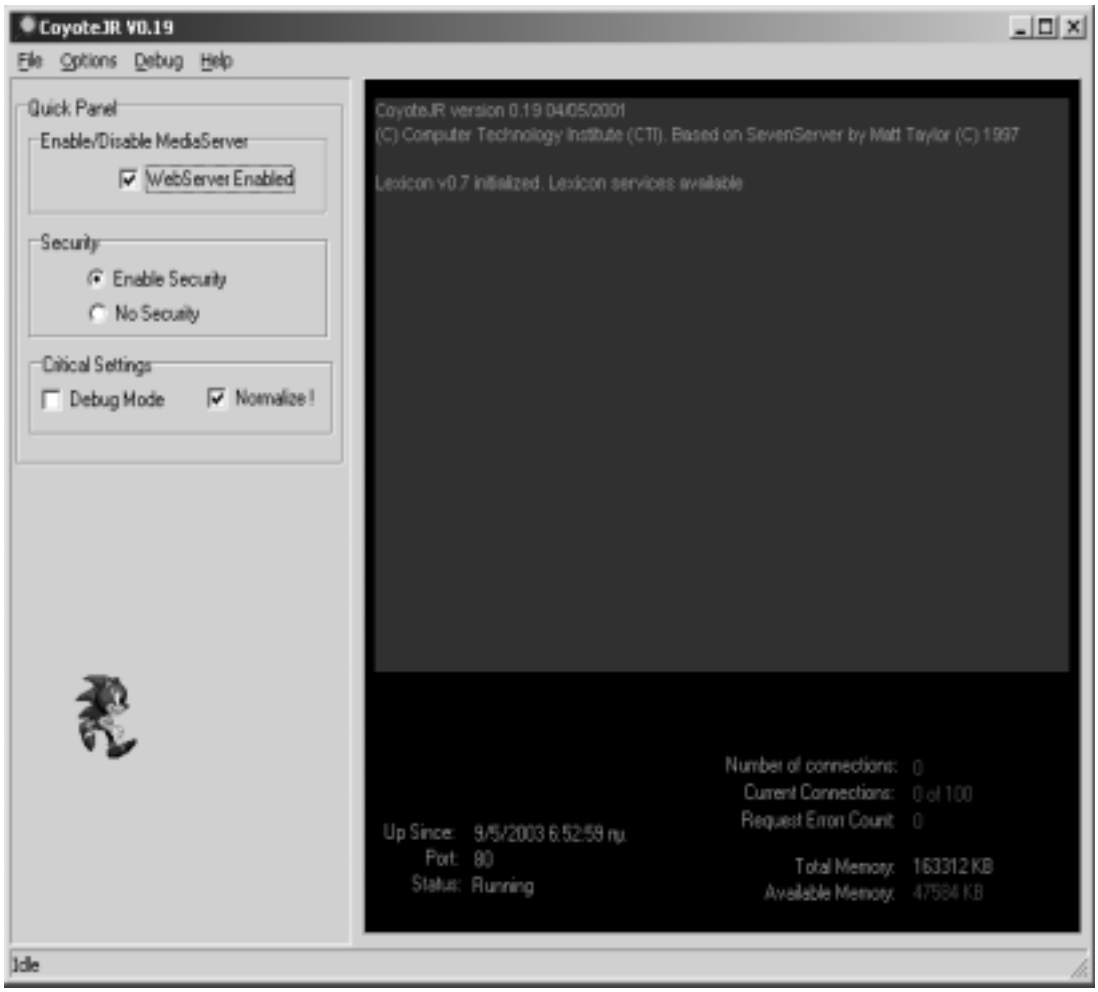
- Εισάγουμε μια νέα εγγραφή στον πίνακα dm_evaluateModel για το συγκεκριμένο αρχείο ελέγχου. Η εγγραφή αυτή περιλαμβάνει τον κωδικό του στιγμιότυπου για το οποίο έγινε η πρόβλεψη, την πραγματική κλάση του στιγμιότυπου, την προβλεπόμενη κλάση και τις πιθανότητες εμφάνισης του στιγμιότυπου στις διάφορες κλάσεις του προβλήματος.

Επαναλαμβάνουμε την παραπάνω διαδικασία για όλα τα στιγμιότυπα του συνόλου ελέγχου και βρίσκουμε το ποσοστό επιτυχίας των προβλέψεων. Το ποσοστό αυτό ισούται με το πλήθος των στιγμιότυπων ελέγχου για τα οποία το σύστημα έκανε σωστή πρόβλεψη προς το συνολικό πλήθος των στιγμιότυπων ελέγχου που χρησιμοποιήθηκαν για την αξιολόγηση του συστήματος.

1.5 Κανονικοποίηση

Όπως έχουμε ήδη αναφέρει η κανονικοποίηση αφορά την ελληνική γλώσσα και επιτυγχάνεται μέσω του προγράμματος Normalizer. Ο Normalizer στην περίπτωση των ρημάτων μετατρέπει τα ρήματα στο πρώτο ενικό πρόσωπο ενώ στην περίπτωση των ουσιαστικών, επιθέτων μετατρέπει τα ουσιαστικά, επίθετα στην ονομαστική πτώση ενικού αριθμού.

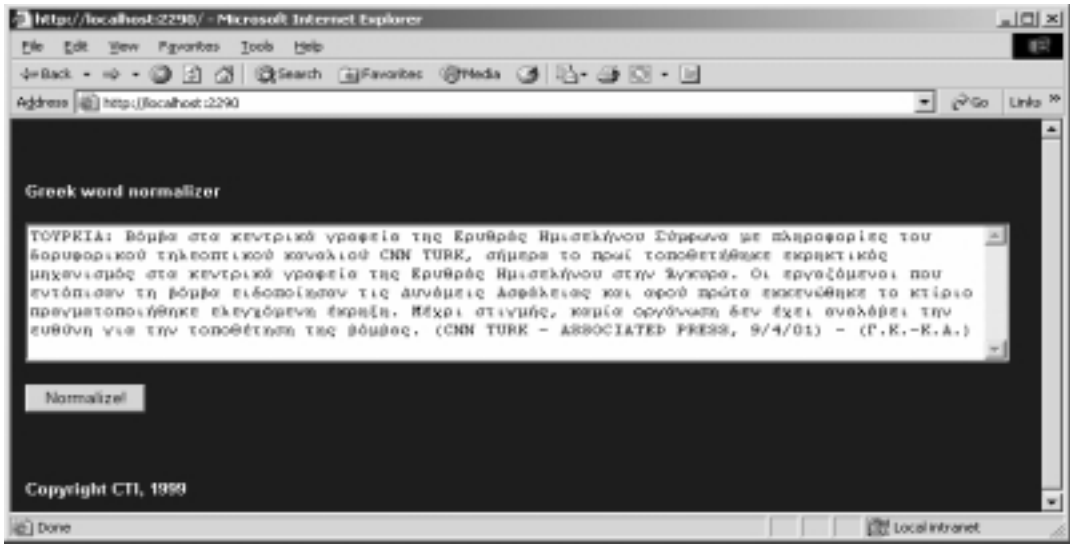
Ο Normalizer σχεδιάστηκε και υλοποιήθηκε από το Εργαστήριο Βάσεων Δεδομένων της Ερευνητικής Μονάδας 2 του Ερευνητικού και Ακαδημαϊκού Ινστιτούτου Τεχνολογίας Υπολογιστών (ΕΑΙΤΥ). Στο ακόλουθο σχήμα (Σχήμα 1.2) φαίνεται το γραφικό του περιβάλλον.



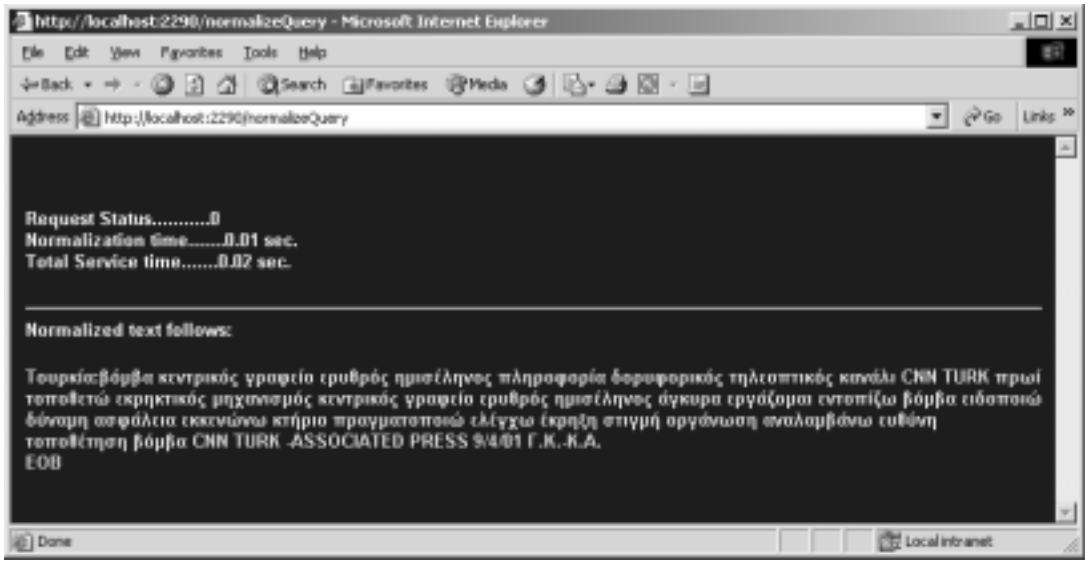
Σχήμα 1.2 Το γραφικό περιβάλλον του Normalizer

Ο Normalizer είναι ένα αυτόνομο πρόγραμμα, και ως αυτόνομο χρησιμοποιείται και στα πλαίσια της εφαρμογής μας. Δέχεται ως είσοδο το προς κανονικοποίηση

κείμενο (Σχήμα 1.3) και επιστρέφει στην έξοδό του το κανονικοποιημένο κείμενο (Σχήμα 1.4).

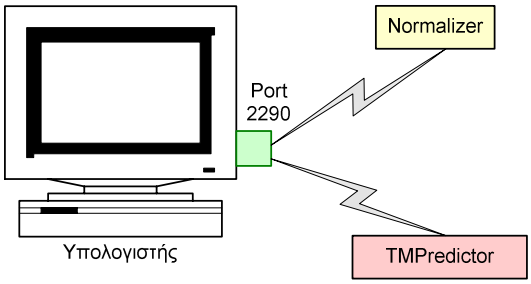


Σχήμα 1.3 Παράδειγμα χρήσης του Normalizer – κείμενο εισόδου



Σχήμα 1.4 Παράδειγμα χρήσης του Normalizer – κείμενο εξόδου

Η αμφίδρομη επικοινωνία του προγράμματος TMPredictor με τον Normalizer επιτυγχάνεται μέσω sockets. Ο Normalizer τρέχει στη θύρα 2290 του υπολογιστή στην οποία συνδέεται και το πρόγραμμα TMPredictor. Το πρόγραμμα TMPredictor τροφοδοτεί τον Normalizer με το προς κανονικοποίηση κείμενο (send request) και ο Normalizer με τη σειρά του επιστρέφει το κανονικοποιημένο κείμενο στο πρόγραμμα TMPredictor μέσω της ίδιας θύρας (get answer). Η επικοινωνία αυτή φαίνεται στο Σχήμα 1.5.



Σχήμα 1.5. Η επικοινωνία του προγράμματος TMPredictor με τον Normalizer

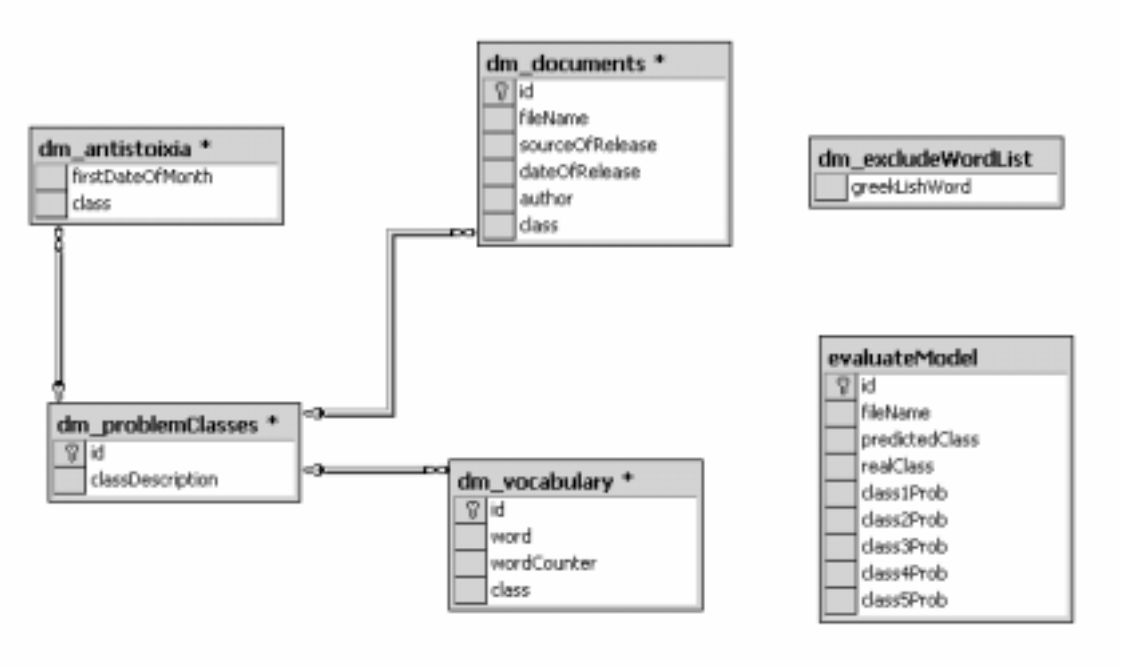
Ο Normalizer δέχεται http requests και μπορεί επομένως να λειτουργήσει και μέσω ενός φυλλομετρητή (web browser).

Ο Normalizer χρησιμοποιήθηκε αφενός μεν για να απομακρύνει τις μη σημαντικές λέξεις των στιγμιότυπων του προς επίλυση προβλήματος, π.χ. άρθρα, σημεία στίξης κ.α., αφετέρου δε για αποτρέψει τις πολλαπλές εγγραφές της ίδιας λέξης στο λεξικό του προβλήματος. Για παράδειγμα, οι λέξεις: "πόλεμος", "πολέμου", "πολέμων", "πόλεμοι" κ.ο.κ. δεν χρειάζεται να εμφανίζονται στο λεξικό ως τέσσερις διαφορετικές λέξεις όπου η καθεμία θα έχει συχνότητα εμφάνισης ένα, αρκεί μια εγγραφή στο λεξικό με τη λέξη "πόλεμος" και συχνότητα εμφάνισης τέσσερα. Με τη χρήση του Normalizer μειώθηκε σημαντικά το πλήθος των λέξεων του λεξικού .

1.6 Η βάση δεδομένων του συστήματος

Στη βάση δεδομένων αποθηκεύονται όλα τα δεδομένα που χρησιμοποιούνται από το σύστημα. Τα δεδομένα αυτά περιλαμβάνουν τόσο τις εισόδους των υπόλοιπων τμημάτων του συστήματος όσο και τις εξόδους αυτών των τμημάτων.

Το σχεσιακό μοντέλο της βάσης δεδομένων που υλοποιήθηκε για τις ανάγκες της εφαρμογής φαίνεται στο επόμενο σχήμα. (Σχήμα 1.6).



Σχήμα 1.6 Το σχεσιακό μοντέλο της βάσης δεδομένων

Στη συνέχεια αναλύουμε κάθε πίνακα της βάσης δεδομένων και περιγράφουμε τα γνωρίσματά του και το ρόλο τους.

- **Πίνακας dm_problemClasses**

Στον πίνακα dm_problemClasses αποθηκεύονται οι κλάσεις του προβλήματος. Κάθε πλειάδα του πίνακα dm_problemClasses περιλαμβάνει τον μοναδικό κωδικό της κλάσης και την λεκτική της περιγραφή.

Πεδίο	**Περιγραφή**
PK_ID	Το αναγνωριστικό της κλάσης - είναι μοναδικό για κάθε κλάση (πρωτεύον κλειδί)
fld_classDescription	Η λεκτική περιγραφή της κλάσης

Πίνακας 1.1 Περιγραφή του πίνακα dm_ problemClasses

- **Πίνακας dm_documents**

Στον πίνακα dm_documents αποθηκεύονται τα στιγμιότυπα του συνόλου εκπαίδευσης, δηλαδή τα αρχεία των ειδήσεων που χρησιμοποιούνται για την εκπαίδευση του ταξινομητή. Είναι σημαντικό να κρατάμε τα αρχεία που έχουμε χρησιμοποιήσει, προκειμένου να μην τα χρησιμοποιήσουμε ξανά. Αν συνέβαινε κάτι τέτοιο η αξιοπιστία του μοντέλου θα μειωνόταν σημαντικά.

Κάθε πλειάδα του πίνακα dm_documents περιλαμβάνει το μοναδικό κωδικό του αρχείου της είδησης, το όνομα του αρχείου, την πηγή έκδοσης, την ημερομηνία έκδοσης, το συγγραφέα και την κλάση στην οποία ανήκει η είδηση.

Πεδίο	**Περιγραφή**
PK_ID	Το αναγνωριστικό του αρχείου εκπαίδευσης, είναι μοναδικό για κάθε αρχείο εκπαίδευσης (πρωτεύον κλειδί).
Filename	Το όνομα του αρχείου
sourceOfRelease	Η πηγή (πρακτορείο, εφημερίδα, περιοδικό κ.λ.π.) από την οποία προήλθε η είδηση
dateOfRelease	Η ημερομηνία έκδοσης της είδησης
Author	Ο συγγραφέας της είδησης
Class	Η κλάση του αρχείου – είναι ξένο κλειδί στον πίνακα dm_ problemClasses.

Πίνακας 1.2 Περιγραφή του πίνακα dm_ documents

- **Πίνακας dm_vocabulary**

Στον πίνακα dm_vocabulary αποθηκεύονται σε κανονικοποιημένη μορφή οι διακριτές σημαντικές λέξεις που εμφανίζονται στα στιγμιότυπα του συνόλου εκπαίδευσης καθώς επίσης και ο αριθμός εμφανίσεων κάθε λέξης στις διάφορες κλάσεις του προβλήματος. Όπως έχουμε ήδη αναφέρει η κανονικοποίηση των λέξεων επιτυγχάνεται μέσω του Normalizer (βλέπε 1.2)

Κάθε πλειάδα του πίνακα dm_ vocabulary περιλαμβάνει ένα μοναδικό κωδικό για κάθε συνδυασμό λέξης και κλάσης, την κανονικοποιημένη λέξη, την κλάση του στιγμιότυπου εκπαίδευσης από το οποίο προήλθε η λέξη, και τη συχνότητα εμφάνισης της λέξης στην εν λόγω κλάση.

Πεδίο	**Περιγραφή**
PK_ID	Το αναγνωριστικό της λέξης - είναι μοναδικό για κάθε συνδυασμό λέξης και κλάσης (πρωτεύον κλειδί).
Word	Η κανονικοποιημένη λέξη (όπως προκύπτει μετά την επεξεργασία της από τον Normalizer).
Class	Η κλάση του στιγμιότυπου εκπαίδευσης από το οποίο προήλθε η

	λέξη – είναι ξένο κλειδί στον πίνακα dm_ problemClasses.
wordCounter	Το πλήθος εμφανίσεων της λέξης στην κλάση του στιγμιότυπου εκπαίδευσης.

Πίνακας 1.3 Περιγραφή του πίνακα dm_ vocabulary

- **Πίνακας dm_excludeWordList**

Στον πίνακα dm_excludeWordList αποθηκεύονται σε κανονικοποιημένη μορφή οι λέξεις που θέλουμε να εξαιρέσουμε από τη διαδικασία εξόρυξης γνώσης. Πρόκειται για κοινότυπες λέξεις, ονόματα ημερών, μηνών, ημερομηνίες, οι οποίες εξαιρέθηκαν επειδή θεωρήσαμε ότι είναι κοινότυπες και συνεπώς η συμβολή τους στη διαδικασία εξόρυξης γνώσης είναι μηδαμινή. Ο πίνακας αυτός δεν χρησιμοποιείται στην παρούσα έκδοση του συστήματος.

Κάθε πλειάδα του πίνακα dm_ excludeWordList περιλαμβάνει την κανονικοποιημένη λέξη η οποία εξαιρείται από τη διαδικασία εξόρυξης γνώσης.

Πεδίο	**Περιγραφή**
excludedWord	Η κανονικοποιημένη λέξη (όπως προκύπτει μετά την επεξεργασία της από τον Normalizer) που εξαιρείται από τη διαδικασία εξόρυξης γνώσης.

Πίνακας 1.4 Περιγραφή του πίνακα dm_ excludeWordList

- **Πίνακας dm_evaluateModel**

Ο πίνακας dm_evaluateModel χρησιμοποιείται για την αξιολόγηση του μοντέλου. Η αξιολόγηση γίνεται με βάση το ποσοστό των στιγμιότυπων ελέγχου για τα οποία το σύστημα προέβλεψε σωστά την κλάση. Για κάθε στιγμιότυπο του συνόλου ελέγχου που εξετάζουμε καταγράφουμε την πραγματική του κλάση (έτσι όπως προκύπτει από το ίδιο το κείμενο), την κλάση που προβλέπει το σύστημα και τις πιθανότητες που υπολογίζει το σύστημα για κάθε επιμέρους κλάση του προβλήματος.

Κάθε πλειάδα του πίνακα dm_ evaluateModel περιλαμβάνει το μοναδικό κωδικό του στιγμιότυπου ελέγχου, το όνομα του αρχείου της είδησης, την κλάση που προέβλεψε το σύστημα, την πραγματική κλάση του στιγμιότυπου, τις πιθανότητες εμφάνισης του στιγμιότυπου στις διάφορες κλάσεις του προβλήματος και την ονομασία του πειράματος. Στην παρούσα φάση τα γνωρίσματα που αναφέρονται στις πιθανότητες των διαφόρων κλάσεων του προβλήματος είναι σταθερού αριθμού (5), σε μία γενικότερη περίπτωση πάντως θα έπρεπε το πλήθος αυτό να είναι δυναμικό και να εξαρτάται από το πλήθος των κλάσεων του προβλήματος. Η ονομασία του πειράματος δεν είναι παρά μια λεκτική περιγραφή των πειραμάτων η οποία βοηθάει στο να

κρατείται το ιστορικό των πειραμάτων και να είναι πιο εύκολη η ανάκτηση των αποτελεσμάτων κάποιου συγκεκριμένου πειράματος του χρήστη.

Πεδίο	**Περιγραφή**
PK_ID	Το αναγνωριστικό του αρχείου ελέγχου, είναι μοναδικό για κάθε αρχείο ελέγχου (πρωτεύον κλειδί).
filename	Το όνομα του αρχείου
predictedClass	Η κλάση που προέβλεψε το σύστημα
realClass	Η πραγματική κλάση
class1Prob	Η πιθανότητα που υπολόγισε το σύστημα για την κλάση 1
class2Prob	Η πιθανότητα που υπολόγισε το σύστημα για την κλάση 2
class3Prob	Η πιθανότητα που υπολόγισε το σύστημα για την κλάση 3
class4Prob	Η πιθανότητα που υπολόγισε το σύστημα για την κλάση 4
class5Prob	Η πιθανότητα που υπολόγισε το σύστημα για την κλάση 5
experimentsName	Η λεκτική περιγραφή (ονομασία) των πειραμάτων

Πίνακας 1.5 Περιγραφή του πίνακα dm_ evaluateModel

- **Πίνακας dm_antistoixia**

Ο πίνακας αυτός είναι βοηθητικός και χρησιμοποιείται μόνο στην περίπτωση του προβλήματος της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού – συγκεκριμένα χρησιμεύει στην εύρεση της κλάσης των στιγμιότυπων εκπαίδευσης (και ελέγχου) του προβλήματος. Στο συγκεκριμένο πρόβλημα, η κλάση ενός στιγμιότυπου εκπαίδευσης προκύπτει με βάση την ημερομηνία έκδοσης της είδησης και το πλήθος των τουριστών εξωτερικού που ήρθαν στην Ελλάδα κατά τη διάρκεια του μήνα στον οποίο αναφέρεται η είδηση. Για παράδειγμα, αν ένα αρχείο εκπαίδευσης αναφέρεται σε μία είδηση που δημοσιεύτηκε την 1/5/2001 και τον 5^{ο} μήνα είχαμε στην Ελλάδα 157.000 τουρίστες εξωτερικού τότε το αρχείο ανήκει στην κλάση 3: (150-200].000 επισκέψεις".

Πεδίο	**Περιγραφή**

FK_class_ID	Η κλάση – είναι ξένο κλειδί στον πίνακα dm_ problemClasses.
fld_firstDateOfMonth	Όπως έχουμε ήδη αναφέρει γνωρίζουμε το πλήθος των τουριστών εξωτερικού ανά μήνα. Στο πεδίο αυτό αποθηκεύεται η πρώτη μέρα κάθε μήνα (η οποία καθορίζει μοναδικά το μήνα), π.χ. 1/1/2001 για το μήνα Ιανουάριο.

Πίνακας 1.6 Περιγραφή του πίνακα dm_antistoixia

1.7 Τεχνολογίες

Για τις ανάγκες της εφαρμογής χρησιμοποιήθηκαν οι ακόλουθες τεχνολογίες:

- Windows 2000 ως πλατφόρμα ανάπτυξης
- SQL Server ως Σύστημα Διαχείρισης Βάσεων Δεδομένων (DBMS)
- Delphi 7 ως γλώσσα ανάπτυξης

2. Εφαρμογή στο πρόβλημα της τουριστικής κίνησης

Στο κεφάλαιο αυτό θα μελετήσουμε την εφαρμογή του συστήματος TMPredictor στο πρόβλημα της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού.

2.1 Το πρόβλημα

Δοθέντων

- κάποιων αρχείων με ειδησεογραφικά δεδομένα που αφορούν τις εσωτερικές και εξωτερικές εξελίξεις (π.χ. σχέσεις της Ελλάδας με τις γείτονες χώρες αλλά και γενικότερες εξελίξεις στο διεθνή τομέα)

- και του αρχείου των επισκέψεων στην Ελλάδα από τουρίστες του εξωτερικού

καλούμαστε να μελετήσουμε κατά πόσο οι ειδήσεις, έτσι όπως περιγράφονται από τον ημερήσιο ηλεκτρονικό τύπο, αντανακλούν το γενικότερο κλίμα που επικρατεί στην Ελλάδα αλλά και διεθνώς και συνεπώς αποτελούν μια ένδειξη του μεγέθους της τουριστικής κίνησης από τουρίστες του εξωτερικού που θα δεχτεί η Ελλάδα (όλα τα διαθέσιμα δεδομένα αναφέρονται στο έτος 2001).

Η διαίσθηση μας λέει πως η θετική ή αρνητική τάση των ξένων τουριστών για την Ελλάδα "φωτογραφίζεται" στον ηλεκτρονικό τύπο μέσω της έντασης με την οποία παρουσιάζονται οι σχετικές ειδήσεις. Για παράδειγμα, αν στην Ελλάδα ή σε κάποια γειτονική της χώρα παρουσιαστεί κάποιο σοβαρό πρόβλημα (π.χ. πόλεμος, ασθένεια κ.λ.π.) θα εμφανιστούν στον τύπο πολλά σχετικά άρθρα σε έντονο ύφος και θα καλλιεργηθεί ένα κλίμα που θα αποτρέπει τους τουρίστες να επισκέπτονται τη χώρα μας. Άρα διαβάζοντας τον τύπο και αποκωδικοποιώντας τον τρόπο με τον οποίο παρουσιάζονται οι ειδήσεις μπορούμε να προβλέψουμε τις πιθανές επισκέψεις από ξένους τουρίστες στην Ελλάδα.

Η ιδέα ενός τέτοιου συστήματος δεν είναι καθόλου "ουτοπική", θυμίζει μάλιστα τον τρόπο με τον οποίο ένας άνθρωπος βγάζει συμπεράσματα σχετικά με την κρισιμότητα κάποιων καταστάσεων μελετώντας απλά τον ημερήσιο τύπο και κατανοώντας τον τρόπο με τον οποίο παρουσιάζονται οι ειδήσεις.

Στόχος μας είναι να "χτίσουμε" ένα μοντέλο πρόβλεψης με βάση τα ειδησεογραφικά δεδομένα και τις επισκέψεις ξένων τουριστών στην Ελλάδα και εν συνεχεία να χρησιμοποιήσουμε το μοντέλο αυτό προκειμένου να προβλέψουμε το πλήθος των επισκέψεων που ενδέχεται να σημειωθούν ως απόρροια μιας συγκεκριμένης είδησης.

Για το σκοπό αυτό θα χρησιμοποιήσουμε το σύστημα TMPredictor την

αρχιτεκτονική του οποίου περιγράψαμε μόλις στο προηγούμενο κεφάλαιο. Μας ενδιαφέρουν οι ακόλουθες λειτουργίες του συστήματος:

1. η εκπαίδευση του μοντέλου πρόβλεψης μέσω των αρχείων των ειδησεογραφικών δεδομένων και των αντίστοιχων επισκέψεων τουριστών του εξωτερικού στην Ελλάδα.

2. η πρόβλεψη του πλήθους των επισκέψεων που θα σημειωθούν δοθέντος κάποιου νέου αρχείου είδησης.

3. η αξιολόγηση του συστήματος TMPredictor για το συγκεκριμένο πρόβλημα.

2.2 Τα ειδησεογραφικά δεδομένα

Τα ειδησεογραφικά δεδομένα για το πρόβλημα της πρόβλεψης της τουριστικής κίνησης αφορούν θέματα εσωτερικού και εξωτερικού που συνδέονται άμεσα ή έμμεσα με την Ελλάδα και αναφέρονται στο έτος 2001. Τα δεδομένα αυτά είναι διαθέσιμα σε ηλεκτρονική μορφή (αρχεία τύπου html).

Στη συνέχεια (Σχήμα 2.1) παραθέτουμε ένα παράδειγμα ενός τέτοιου ειδησεογραφικού αρχείου:

Ο Τουρκοκύπριος ηγέτης ΡΑΟΥΦ ΝΤΕΝΚΤΑΣ επανέλαβε σήμερα ότι δεν θα δεχθεί την ένταξη της Κύπρου στην ΕΕ "χωρίς την αναγνώριση δύο ανεξάρτητων κυρίαρχων κρατών στην Κύπρο". Μιλώντας σήμερα (6/6) σε μέλη της τουρκικής Εθνοσυνέλευσης που συμμετέχουν στη μεικτή κοινοβουλευτική επιτροπή Τουρκίας-ΕΕ, στην Άγκυρα, ο ΝΤΕΝΚΤΑΣ υποστήριξε ότι οι ελληνοκύπριοι προσπαθούν να πετύχουν δια της Ε.Ε., ότι δεν πέτυχαν με τη βία. "Συμμετέχουν στις εκ του σύνεγγυς συνομιλίες για να προωθήσουν τους δικούς τους σκοπούς και δεν προτίθενται να αναγνωρίσουν στους τουρκοκύπριους παρά δικαιώματα μειονότητας", είπε ο ΝΤΕΝΚΤΑΣ. Αν η Κυπριακή Δημοκρατία ενταχθεί στην ΕΕ, ο τουρκικός στρατός θα θεωρείται ως στρατός κατοχής, είπε ακόμα ο ΝΤΕΝΚΤΑΣ σε μία προσπάθεια να προκαλέσει συναισθηματικές αντιδράσεις και να ελέγξει την κριτική που για πρώτη φορά ασκείται δημόσια στην Τουρκία για την πολιτική που ακολουθεί στο Κυπριακό. "Οι ελληνοκύπριοι χρησιμοποιούν την Κύπρο για να εμποδίσουν την ένταξη της Τουρκίας στην ΕΕ", είπε ο ΝΤΕΝΚΤΑΣ. Ο ελληνοτουρκικός πόλεμος είναι αναπόφευκτος αν η Ελλάδα προσαρτήσει την Κύπρο, είπε ακόμα ο τουρκοκύπριος ηγέτης, που θεωρεί ότι η ένταξη της Κύπρου 7στη ΕΕ, ισοδυναμεί με ένωση με την Ελλάδα. Η Κύπρος πρέπει να ενταχθεί στην ΕΕ μαζί με την Τουρκία, αφού διευθετηθεί το θέμα των ιδιοκτησιών και επικρατήσει μια βιώσιμη ατμόσφαιρα στο νησί, τόνισε ο ίδιος. Πρόσθεσε επίσης ότι "αν οι ελληνοκύπριοι και οι Έλληνες πάρουν τον έλεγχο του νησιού" θα επαναληφθούν τα όσα έγιναν στην Κρήτη κατά την περίοδο της Οθωμανικής Αυτοκρατορίας. Σχολιάζοντας τις αποφάσεις του Ελσίνκι που συνδέουν την έναρξη των ενταξιακών διαπραγματεύσεων της Τουρκίας με την ΕΕ με τη λύση του Κυπριακού, ο ΝΤΕΝΚΤΑΣ, είπε ότι "η Μεγάλη Βρετανία πολεμούσε με τους Ιρλανδούς αλλά κανένας δεν της ζήτησε να λύσει τη διαμάχη για να ενταχθεί στην ΕΕ ". (ΑΠΕ, 6/6/01) - (Θ.Κ.)

Πηγή είδησης

Ημ/νια έκδοσης

Συγγραφέας

Σχήμα 2.1 Παράδειγμα ενός αρχείου ειδησεογραφικών δεδομένων για το πρόβλημα της τουριστικής κίνησης

Τα ακατέργαστο αρχεία των ειδήσεων για το πρόβλημα της τουριστικής κίνησης ήταν πάρα πολλά, δυστυχώς όμως δεν μπορούσαν να χρησιμοποιηθούν όλα καθώς δεν ακολουθούσαν μία κοινή σύνταξη. Επειδή το αρχικό πλήθος των αρχείων ήταν πολύ μεγάλο, χρειάστηκε να φτιαχτεί ένα πρόγραμμα το οποίο θα αποφαινόταν για το κατά πόσο ένα αρχείο είναι κατάλληλο για να χρησιμοποιηθεί για την εφαρμογή μας ή όχι. Η καταλληλότητα αναφέρεται αποκλειστικά στη δομή του αρχείου και όχι στο περιεχόμενο της είδησης. Έτσι απορρίψαμε όσα αρχεία δεν ακολουθούσαν την κοινή τυποποίηση: είδηση (πηγή είδησης, ημερομηνία έκδοσης είδησης) – (συγγραφέας είδησης). Το τελικό πλήθος των αρχείων που χρησιμοποιήθηκαν για τις ανάγκες του συστήματος είναι 1860.

2.3 Τα δεδομένα των επισκέψεων

Το αρχείο των επισκέψεων τουριστών του εξωτερικού στην Ελλάδα αφορά το έτος 2001 και περιλαμβάνει τον αριθμό των επισκέψεων που σημειώθηκαν στην Ελλάδα για κάθε μήνα του 2001 (Πίνακας 2.1).

Ιανουάριος	Φεβρουάριος	Μάρτιος	Απρίλιος	Μάιος	Ιούνιος	Ιούλιος	Αύγουστος	Σεπτέμβριος	Οκτώβριος	Νοέμβριος	Δεκέμβριος
60	67	73	130	157	181	249	353	171	128	95	93

Πίνακας 2.1 Οι μηνιαίες επισκέψεις τουριστών του εξωτερικού για το έτος 2001 (τα ποσά του πίνακα αναφέρονται σε χιλιάδες)

Κλάσεις επισκέψεων

Παρατηρώντας το αρχείο των επισκέψεων εύκολα διαπιστώνει κανείς πως υπάρχουν μήνες με παραπλήσιο αριθμό επισκέψεων (π.χ. Ιανουάριος, Φεβρουάριος). Για το λόγο αυτό σκεφτήκαμε να φτιάξουμε κάποιες κλάσεις επισκέψεων όπου κάθε κλάση θα χαρακτηρίζεται από ένα διάστημα επισκέψεων (ελάχιστος αριθμός επισκέψεων – μέγιστος αριθμός επισκέψεων) και θα περιλαμβάνει, ενδεχομένως, παραπάνω του ενός μήνες (Πίνακας 2.2).

Κλάση	**Διάστημα επισκέψεων (ελάχιστος αριθμός – μέγιστος αριθμός)**
1	[0 - 80] επισκέψεις
2	(80 - 150] επισκέψεις
3	(150 - 200] επισκέψεις
4	(200 – 400] επισκέψεις

Πίνακας 2.2 Οι κλάσεις επισκέψεων (τα ποσά της δεύτερης στήλης του πίνακα αναφέρονται σε χιλιάδες)

Η κατανομή των αρχείων των ειδησεογραφικών δεδομένων στις διάφορες κλάσεις επισκέψεων για το συγκεκριμένο πρόβλημα φαίνεται στον ακόλουθο πίνακα (Πίνακας 2.3).

Κλάση	**Σύνολο αρχείων της κλάσης**
[0 - 80] επισκέψεις	406
(80 - 150] επισκέψεις	407
(150 - 200] επισκέψεις	545
(200 – 400] επισκέψεις	502

Πίνακας 2.3 Η κατανομή των αρχείων στις κλάσεις επισκέψεων (τα ποσά της πρώτης στήλης του πίνακα αναφέρονται σε χιλιάδες)

Τα αρχεία κάθε κλάσης διασπάστηκαν σε δύο σύνολα:

- Το σύνολο των στιγμιότυπων εκπαίδευσης, το οποίο περιλαμβάνει τα 2/3 των αρχείων της κλάσης και χρησιμοποιείται για τη δημιουργία του ταξινομητή.
- Το σύνολο των στιγμιότυπων ελέγχου, το οποίο περιλαμβάνει το υπόλοιπο 1/3 των αρχείων της κλάσης και χρησιμοποιείται για την αξιολόγηση του ταξινομητή.

Η κατανομή των αρχείων κάθε κλάσης στα σύνολα εκπαίδευσης και ελέγχου φαίνεται στον ακόλουθο πίνακα (Πίνακας 2.4).

Κλάση	Σύνολο αρχείων εκπαίδευσης	Σύνολο αρχείων ελέγχου
[0 - 80] επισκέψεις	270	136
(80 - 150] επισκέψεις	271	136
(150 - 200] επισκέψεις	363	182
(200 – 400] επισκέψεις	328	164

Πίνακας 2.4 Η κατανομή των αρχείων κάθε κλάσης στα σύνολα εκπαίδευσης και ελέγχου

Η κλάση ενός αρχείου του συνόλου εκπαίδευσης ή ελέγχου προκύπτει από την ημερομηνία δημοσίευσης της είδησης. Για παράδειγμα, όσον αφορά το κείμενο του Σχήματος 2.1 η ημερομηνία δημοσίευσής του είναι η 06/06/2001 άρα ανήκει στον 6^{ο} μήνα του 2001 στον οποίο όπως βλέπουμε από τον Πίνακα 2.1 σημειώθηκαν στην Ελλάδα 181.000 επισκέψεις από ξένους τουρίστες. Οι 181.000 επισκέψεις ανήκουν στην κλάση: (150 - 200].000 επισκέψεις όπως φαίνεται από τον Πίνακα 2.2. Άρα η κλάση του αρχείου είναι η κλάση 3: (150-200].000 επισκέψεις.

3. Εφαρμογή στο πρόβλημα της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου

Στο κεφάλαιο αυτό θα μελετήσουμε την εφαρμογή του συστήματος TMPredictor στο πρόβλημα της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου ενημέρωσης σε διάφορες προκαθορισμένες κατηγορίες. Ο δικτυακός τόπος που χρησιμοποιήθηκε στα πλαίσια της μεταπτυχιακής εργασίας είναι το Flash (www.flash.gr) το οποίο αποτελεί ένα πολύ δημοφιλή τρόπο ενημέρωσης μέσω διαδικτύου.

3.1 Το πρόβλημα

Το Flash αποτελεί ένα δημοφιλή δικτυακό τόπο ενημέρωσης για μια πληθώρα θεμάτων: πολιτικά, οικονομικά, αθλητικά κ.α. Στόχος μας είναι να φτιάξουμε έναν ταξινομητή ο οποίος θα εκπαιδευτεί μέσω των ειδήσεων του Flash έτσι ώστε να μπορεί να προβλέπει την κατηγορία στην οποία ανήκει μια νέα άγνωστη είδηση. Το πρόβλημα είναι σημαντικό για κάθε δικτυακό τόπο καθώς δέχεται ειδήσεις από διάφορες πηγές και θα πρέπει να αποφασίζει κάθε φορά σε ποια από τις κατηγορίες του θα τις εντάσσει.

Για τους σκοπούς της εργασίας μας επιλέξαμε 4 κατηγορίες του Flash:

- Οικονομία (http://financial.flash.gr/)
- Τεχνολογία (http://tech.flash.gr/)
- Αθλητισμός (http://sportnews.flash.gr/)
- Αυτοκίνητο (http://automoto.flash.gr/)

Για κάθε κατηγορία κατεβάσαμε 210 ειδήσεις από το Flash εκ των οποίων τα 2/3 χρησιμοποιήθηκαν ως στιγμιότυπα εκπαίδευσης και το υπόλοιπο 1/3 ως στιγμιότυπα ελέγχου.

Ο στόχος μας είναι να "χτίσουμε" ένα μοντέλο πρόβλεψης με βάση τα ειδησεογραφικά δεδομένα και τις κλάσεις τους και εν συνεχεία να χρησιμοποιήσουμε το μοντέλο αυτό προκειμένου να προβλέψουμε την κατηγορία στην οποία θα ανήκει μία νέα άγνωστη είδηση.

Για το σκοπό αυτό χρησιμοποιούμε όπως και στην περίπτωση του προβλήματος της τουριστικής κίνησης το σύστημα TMPredictor. Οι λειτουργίες του συστήματος που μας ενδιαφέρουν είναι οι ακόλουθες:

1. η εκπαίδευση του μοντέλου πρόβλεψης μέσω των αρχείων των ειδησεογραφικών δεδομένων και των αντίστοιχων κλάσεών τους.

2. η πρόβλεψη της κατηγορίας στην οποία ανήκει μια νέα άγνωστη είδηση.

3. η αξιολόγηση του συστήματος TMPredictor για το συγκεκριμένο πρόβλημα.

3.2 Τα ειδησεογραφικά δεδομένα

Τα ειδησεογραφικά δεδομένα που χρησιμοποιήσαμε για τους σκοπούς της εφαρμογής προέρχονται από τον δικτυακό τόπο Flash. Πιο συγκεκριμένα

- για την κλάση *Οικονομία* τα δεδομένα προέρχονται από τη διεύθυνση: http://financial.flash.gr/
- για την κλάση *Τεχνολογία* τα δεδομένα προέρχονται από τη διεύθυνση: http://tech.flash.gr/
- για την κλάση *Αθλητισμός* τα δεδομένα προέρχονται από τη διεύθυνση: http://sportnews.flash.gr/
- για την κλάση *Αυτοκίνητο* τα δεδομένα προέρχονται από τη διεύθυνση: http://automoto.flash.gr/

Τα ειδησεογραφικά δεδομένα τα κατεβάσαμε μέσω του προγράμματος Offline Explorer και εν συνεχεία τα επεξεργαστήκαμε για να αφαιρέσουμε τα διάφορα hmtl tags και να πάρουμε μόνο το κείμενο της είδησης.

Στη συνέχεια (Σχήμα 3.1 – 3.4) παραθέτουμε κάποια παραδείγματα αρχείων ειδησεογραφικών δεδομένων για κάθε κλάση του προβλήματος της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου.

Στα 900 εκατ. ευρώ ανήλθαν τελικώς οι συνολικές προσφορές που υποβλήθηκαν από ελληνικές και ξένες τράπεζες αλλά και ασφαλιστικές εταιρείες για το ομολογιακό δάνειο μειωμένης εξασφάλισης της Εθνικής τράπεζας. Το τελικό ποσό που άντλησε η Εθνική ανήλθε στα 750 εκατ. ευρώ λόγω των αυξημένων προσφορών. Σημειώνεται ότι ο αρχικός στόχος ήταν να αντληθούν περί τα 500 εκατ. ευρώ. Αξίζει να σημειωθεί ότι το ύψος των προσφορών αποτέλεσε ρεκόρ για τα ελληνικά δεδομένα, όπως επίσης η απόδοση, η οποία ήταν ιδιαίτερα συμφέρουσα για την Εθνική τράπεζα και διαμορφώθηκε στο 4,25%.

Αξιοσημείωτη ήταν και η γεωγραφική διασπορά της ομολογιακής έκδοσης, καθώς το 50% απορροφήθηκε από αμοιβαία κεφάλαια και ασφαλιστικές εταιρείες της Ευρώπης.

Επίσης στα τέλη του μηνός λήγει παλαιότερη ομολογιακή έκδοση, ύψους 220 εκατ. ευρώ, την οποία η Εθνική θα ανανεώσει.

Τα υπόλοιπα κεφάλαια που θα αντλήσει η Εθνική, δηλαδή τα 540 εκατ. ευρώ, θα κατευθυνθούν στην επέκταση της τράπεζας στο retail αλλά και σε εξαγορές μικρών τραπεζικών ιδρυμάτων στο εξωτερικό.

Η άντληση φθηνών κεφαλαίων θα επιτρέψει στην Εθνική τράπεζα να εστιασθεί στην λιανική τραπεζική και ειδικά στα καταναλωτικά και επιχειρηματικά δάνεια μικρομεσαίων επιχειρήσεων. Επίσης θα δώσει την δυνατότητα να προχωρήσει σε εξαγορές μικρών τραπεζών στα Βαλκάνια αλλά και στις ΗΠΑ καθώς και σε άλλες επιχειρηματικές κινήσεις στο εσωτερικό.

Σημειώνεται ότι με τα κεφάλαια αυτά ο δείκτης κεφαλαιακής επάρκειας μπορεί να αυξηθεί έως και 2,5 % υψηλότερα. Την τρέχουσα περίοδο ο δείκτης κεφαλαιακής επάρκειας διαμορφώνεται στα 10,4% και με τα νέα κεφάλαια που θα προστεθούν θα αγγίξει το 12,8%..

Σχήμα 3.1 Παράδειγμα ενός αρχείου ειδησεογραφικών δεδομένων για την κλάση *Οικονομία*

Τρία σημαντικά προβλήματα ασφαλείας, το σημαντικότερο από τα οποία δίνει σε τρίτους τη δυνατότητα να αποκτήσουν πλήρη έλεγχο του υπολογιστή-στόχου, εντοπίστηκαν πρόσφατα στις ρουτίνες Java του Microsoft Virtual Machine που βρίσκεται στις περισσότερες πρόσφατες εκδόσεις των Windows και του Internet Explorer της Microsoft.

Για να εκμεταλλευτεί τα προβλήματα αυτά, ένας κακόβουλος τρίτος πρέπει να κατασκευάσει μια κατάλληλα διαμορφωμένη σελίδα web την οποία θα πρέπει να τοποθετήσει σε ένα site ή να στείλει μέσω ηλεκτρονικού ταχυδρομείου (HTML mail) στον υπολογιστή-στόχο.

Το πρώτο πρόβλημα οφείλεται στον ελλιπή τρόπο με τον οποίο ελέγχονται οι αιτήσεις για το φόρτωμα και την εκτέλεση μιας DLL στον υπολογιστή-client από τα Java Database Connectivity (JDBC) classes, τα οποία δίνουν τη δυνατότητα σε προγράμματα Java (Java applets) να συνδέονται και να χρησιμοποιούν δεδομένα από μια πληθώρα πηγών δεδομένων όπως αρχεία ή SQL Server databases.

Το δεύτερο πρόβλημα οφείλεται στον ελλιπή τρόπο ελέγχου των δεδομένων τα οποία δίνονται ως input στα Java Database Connectivity (JDBC) classes.

Το τρίτο πρόβλημα οφείλεται στον ελλιπή τρόπο με τον οποίο πιστοποιείται η δυνατότητα πρόσβασης των διαφόρων προγραμμάτων Java (Java applets) σε αρχεία XML από τα XML Support classes τα οποία δίνουν δυνατότητες XML σε προγράμματα Java. Το πρόβλημα αυτό είναι και το χειρότερο από τα τρία καθώς, ένας κακόβουλος τρίτος θα μπορούσε να αποκτήσει πλήρη πρόσβαση στον υπολογιστή-στόχο.

Ταυτόχρονα με την ανακοίνωση του προβλήματος (Q329077), κυκλοφόρησε και το σχετικό patch, που σύμφωνα με την εταιρεία, είναι μεγάλης σημασίας και πρέπει να εγκατασταθεί αμέσως, το οποίο είναι διαθέσιμο μέσω του site Windows Update.

Σχήμα 3.2 Παράδειγμα ενός αρχείου ειδησεογραφικών δεδομένων για την κλάση *Τεχνολογία*

Η ισχύς εν τη... «Ενώσει». Αυτή τη φορά δεν υπήρξαν αποδοκιμασίες (ούτε για τον Μπάγεβιτς, ούτε για τον Γεωργάτο) στη Νέα Φιλαδέλφεια. Μόνο ικανοποίηση για τη νίκη-βάλσαμο στις πολλές πληγές που έχουν δημιουργηθεί στην «Ενωση» μετά τις ραγδαίες εξελίξεις στα διοικητικά.

Ο «Ντούσκο» προτίμησε να σταθεί στην ουσία. «Βρεθήκαμε δυο φόρες πίσω στο σκορ. Αν και σε γενικές γραμμές παίξαμε καλά, είχαμε πρόβλημα στην άμυνα όπου κάναμε εύκολα λάθη τα οποία μπορούσαμε να αποφύγουμε ειδικά στο δεύτερο μέρος. Είχαμε αντίπαλο μια πολύ καλή και δυνατή ομάδα με καλούς επιθετικούς που μας δημιουργούσαν προβλήματα. Το παιχνίδι μπορούσε να λήξει με δικό μας τέταρτο γκολ, όμως και το Αιγάλεω μπορούσε να πετύχει τρίτο γκολ. Το πιο σημαντικό για εμάς είναι ότι πήραμε τους τρεις βαθμούς της νίκης».

Από την πλευρά του ο Γιώργος Χατζάρας αφού ευχήθηκε στην ΑΕΚ να ξεπεράσει γρήγορα την κρίση και να βγει από το αδιέξοδο είπε για τονν αγώνα: «Καταφέραμε να αιφνιδιάσουμε και να προηγηθούμε νωρίς στο σκορ. Χάσαμε ευκαιρίες και φέραμε την αντίπαλό μας σε δύσκολη θέση. Το δεύτερο ημίχρονο έκρυβε εκπλήξεις και μας δυσκόλεψε πολύ το γεγονός ότι μείναμε με δέκα παίκτες».

Ο Δημήτρης Ναλιτζής τόνισε: «Οταν δεχτήκαμε το πρώτο γκολ νιώσαμε ότι ήρθαν τα πάνω-κάτω. Χάναμε εύκολα τη μπάλα, δεν είχαμε αυτοσυγκέντρωση αλλά καταφέραμε να ανατρέψουμε την εις βάρος μας κατάσταση. Στο δεύτερο ημίχρονο ήμασταν αποφασιστικοί και κατά διαστήματα παίξαμε καλά. Μάλιστα, θα μπορούσαμε να πετύχουμε παραπάνω γκολ. Αυτό που μετράει περισσότερο για εμάς είναι η νίκη. Στην ανατροπή βοήθησε και η αποβολή του Χλωρού».

Ο Αβραάμ Συμεωνίδης, ο οποίος άνοιξε το σκορ για το Αιγάλεω: «Καταφέραμε να προηγηθούμε δύο φορές. Παλέψαμε το παιχνίδι και προσπαθήσαμε να πάρουμε το θετικό αποτέλεσμα. Δυστυχώς, όμως, στο τέλος δεν τα καταφέραμε και σε αυτό συνέβαλε και το ότι μείναμε με δέκα παίχτες».

***Για το καλό κλίμα που επικρατούσε στις εξέδρες μίλησε και ο Γιάννης Γρανίτσας: «Ευτυχώς σήμερα είχαμε ησυχία στο γήπεδο και καταλαβαίνετε το εννοώ. Βοήθησε και η συμβολή της original, η οποία ήταν αυτή που έπρεπε και ελπίζω να είναι και στην συνέχεια. Έχουμε μιλήσει και το έχουν καταλάβει» τόνισε ο πρόεδρος της ερασιτεχνικής ΑΕΚ και πρόσθεσε: «Εγώ δεν μιλάω με τους ποδοσφαιριστές γιατί δεν πηγαίνω στα αποδυτήρια. Δεν είναι δουλειά της διοίκησης να πηγαίνει στα αποδυτήρια, αυτή είναι δουλειά του προπονητή. Μοναδική που πήγα στα αποδυτήρια στα 7 χρόνια που είμαι στο μπάσκετ ήταν όταν η ΑΕΚ πήρε το πρωτάθλημα και πήγα να πανηγυρίσω».

Σχήμα 3.3 Παράδειγμα ενός αρχείου ειδησεογραφικών δεδομένων για την κλάση *Αθλητισμός*

Alfa Romeo 147 Μπορεί η δίλιτρη έκδοση της Alfa Romeo 147 να αποδίδει 150 ίππους, ωστόσο το πλαίσιο του αυτοκινήτου, η ιστορία της ιταλικής φίρμας και οι γρήγορες εκδόσεις των ανταγωνιστικών μοντέλων (Audi S3 και τα επερχόμενα Ford Focus RS, Civic Type-R), κάνουν επιτακτική την ανάγκη τοποθέτησης ενός δυνατότερου κινητήρα. Την ανάγκη αυτή διαπίστωσε και η ίδια Alfa Romeo, η οποία σχεδιάζει το 2002 να λανσάρει στην αγορά μια σπορ έκδοση της 147 που θα εφοδιάζεται με ένα εξακύλινδρο κινητήρα με απόδοση από... 220 έως 250 ίππους!

Η ιταλική εταιρεία υποστηρίζει ότι το πλαίσιο της 147 είναι σε θέση να αντέξει τα παραπάνω άλογα, ωστόσο κρίνει απαραίτητη τη χρήση ενός εξελιγμένου συστήματος traction control, προκειμένου να περάσουν οι ίπποι στο δρόμο μέσω των εμπρός τροχών. Υπολογίζεται ότι η 147 GTA, όπως θα ονομάζεται η κορυφαία έκδοση, θα επιταχύνει από στάση στα 100 χλμ/ώρα, ξεπερνώντας έτσι σε επιδόσεις τον ανταγωνισμό.

Αν και η φιλοσοφία κατασκευής της 147 GTA στηρίζεται στην οδηγική ευχαρίστηση, δεν θα πρόκειται για ένα γυμνό μοντέλο από πλευράς εξοπλισμού (βλέπε 106 Rallye). Αντίθετα, θα προσφέρει πολυτέλεια χωρίς συμβιβασμούς και το επίπεδο εξοπλισμού θα έιναι αναβαθμισμένο σε σχέση με τις ταπεινότερες εκδόσεις. Η τιμή της νέας έκδοσης προσδιορίζεται στις 22.000 λίρες (12 εκατομμύρια δραχμές).

Alfa Romeo 156 Παράλληλα, η ιταλική εταιρεία εξελίσσει και μια GTA έκδοση της 156, η οποία θα εφοδιάζεται με ένα εξακύλινδρο σύνολο, χωρητικότητας 3.2 λίτρων, το οποίο θα αποδίδει 250 ίππους, ενώ η μετάδοση θα γίνεται μέσω ενός εξατάχυτου μηχανικού κιβωτίου. Επιπλέον, η 156 GTA μπορεί να διατίθεται και με το κιβώτιο διαδοχικών σχέσεων (Selespeed). Η τιμή της υπολογίζεται κοντά στις 26.000 λίρες, δηλαδή 14.2 εκατομμύρια δραχμές.

Σχήμα 3.4 Παράδειγμα ενός αρχείου ειδησεογραφικών δεδομένων για την κλάση *Αυτοκίνητο*

3.3 Οι κλάσεις του προβλήματος

Όπως αναφέραμε ήδη για τους σκοπούς της μεταπτυχιακής επιλέξαμε τέσσερις κατηγορίες ειδήσεων του Flash οι οποίες και αποτελούν τις κλάσεις του προβλήματος (Πίνακας 3.1).

Κλάση	Περιγραφή κλάσης
1	Οικονομία
2	Τεχνολογία
3	Αθλητισμός
4	Αυτοκίνητο

Πίνακας 3.1 Οι κλάσεις του προβλήματος

Η κατανομή των αρχείων των ειδησεογραφικών δεδομένων που χρησιμοποιήθηκαν στα πλαίσια της εφαρμογής στις διάφορες κλάσεις του προβλήματος φαίνεται στον ακόλουθο πίνακα (Πίνακας 3.2).

Κλάση	Σύνολο αρχείων της κλάσης
Οικονομία	210
Τεχνολογία	210
Αθλητισμός	210
Αυτοκίνητο	210

Πίνακας 3.2 Η κατανομή των αρχείων στις κλάσεις του προβλήματος

Τα αρχεία κάθε κλάσης διασπάστηκαν σε δύο σύνολα:

- Το σύνολο των στιγμιότυπων εκπαίδευσης, το οποίο περιλαμβάνει τα 2/3 των αρχείων της κλάσης και χρησιμοποιείται για τη δημιουργία του ταξινομητή.
- Το σύνολο των στιγμιότυπων ελέγχου, το οποίο περιλαμβάνει το υπόλοιπο 1/3 των αρχείων της κλάσης και χρησιμοποιείται για την αξιολόγηση του ταξινομητή.

Η κατανομή των αρχείων κάθε κλάσης στα σύνολα εκπαίδευσης και ελέγχου φαίνεται στον ακόλουθο πίνακα (Πίνακας 3.3).

Κλάση	**Σύνολο αρχείων εκπαίδευσης**	**Σύνολο αρχείων ελέγχου**
Οικονομία	140	70
Τεχνολογία	140	70
Αθλητισμός	140	70
Αυτοκίνητο	140	70

Πίνακας 3.3 Η κατανομή των αρχείων κάθε κλάσης στα σύνολα εκπαίδευσης και ελέγχου

Η κλάση ενός αρχείου του συνόλου εκπαίδευσης ή ελέγχου προκύπτει από την διεύθυνση (url) από την οποία προέρχεται. Για παράδειγμα, ένα αρχείο που προέρχεται από την διεύθυνση: www.financial.flash.gr ανήκει στην κατηγορία *Οικονομία*.

4. Πειράματα και αποτελέσματα

Στο κεφάλαιο αυτό θα περιγράψουμε τα πειράματα που τρέξαμε για τις περιπτώσεις των δύο προβλημάτων, του προβλήματος της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού και του προβλήματος της ταξινόμησης των ειδήσεων του δικτυακού τόπου Flash (www.flash.gr), καθώς επίσης και τα αποτελέσματα αυτών των πειραμάτων.

Ο βασικός στόχος αυτού του κεφαλαίου είναι να δούμε και να σχολιάσουμε τα αποτελέσματα της εφαρμογής του συστήματος TMPredictor σε πραγματικά προβλήματα.

4.1 Το πρόβλημα της πρόβλεψης της τουριστικής κίνησης

Όπως έχουμε ήδη αναφέρει το πρόβλημα της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού αφορά την πρόβλεψη της κλάσης κάποιας νέας είδησης σε κάποια από τις προκαθορισμένες κλάσεις του προβλήματος. Τα δεδομένα του προβλήματος αποτελούνται από ειδησεογραφικά δεδομένα για το έτος 2001 και τις αντίστοιχες επισκέψεις τουριστών του εξωτερικού στην Ελλάδα (ανά μήνα του 2001).

Για το πρόβλημα χρησιμοποιήθηκαν συνολικά 3800 αρχεία ειδησεογραφικών δεδομένων, εκ των οποίων τα 2/3 χρησιμοποιήθηκαν για τη δημιουργία του μοντέλου πρόβλεψης (τα στιγμιότυπα αυτά αποτελούν το σύνολο των στιγμιότυπων εκπαίδευσης) και τα υπόλοιπα 1/3 χρησιμοποιήθηκαν για την αξιολόγηση του μοντέλου πρόβλεψης (τα στιγμιότυπα αυτά αποτελούν το σύνολο των στιγμιότυπων ελέγχου). Η κατανομή των στιγμιότυπων στις διάφορες κλάσεις του προβλήματος και στα σύνολα εκπαίδευσης και ελέγχου φαίνεται στον ακόλουθο πίνακα (Πίνακας 4.1)

Κλάση	Σύνολο αρχείων της κλάσης	Σύνολο αρχείων εκπαίδευσης	Σύνολο αρχείων ελέγχου
[0 - 80] επισκέψεις	916	611	305
(80 - 150] επισκέψεις	978	735	243
(150 - 200] επισκέψεις	1118	687	431
(200 – 400] επισκέψεις	986	628	358

Πίνακας 4.1 Η κατανομή των αρχείων στις διάφορες κλάσεις του προβλήματος της τουριστικής κίνησης και στα σύνολα εκπαίδευσης και ελέγχου (τα ποσά της πρώτης στήλης του πίνακα αναφέρονται σε χιλιάδες).

.

Αρχικά λοιπόν, χτίστηκε το μοντέλο πρόβλεψης μέσω των στιγμιότυπων του συνόλου εκπαίδευσης (σύνολο 2661 στιγμιότυπα εκπαίδευσης) και στη συνέχεια αξιολογήθηκε η απόδοσή του μέσω των στιγμιότυπων του συνόλου ελέγχου (σύνολο 1337 στιγμιότυπα). Τα αποτελέσματα των πειραμάτων φαίνονται στον ακόλουθο πίνακα (Πίνακας 4.2).

Κλάση	**Σύνολο σωστά προβλεπόμενων αρχείων**	**Σύνολο λάθος προβλεπόμενων αρχείων**	**Ποσοστό επιτυχίας %**
[0 - 80] επισκέψεις	133	172	43,61
(80 - 150] επισκέψεις	122	121	50,21
(150 - 200] επισκέψεις	90	341	20,88
(200 – 400] επισκέψεις	252	106	70,39

Πίνακας 4.2 Τα αποτελέσματα των πειραμάτων για το πρόβλημα της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού (τα ποσά της πρώτης στήλης του πίνακα αναφέρονται σε χιλιάδες)

Παρατηρούμε λοιπόν πως η μέση απόδοση του μοντέλου για το συγκεκριμένο πρόβλημα είναι 46,26%. Να υπενθυμίσουμε πως η απόδοση ορίζεται ως το πηλίκο του πλήθους των σωστά ταξινομημένων ειδήσεων του συνόλου ελέγχου προς το πλήθος των ειδήσεων του συνόλου ελέγχου. Αν και η απόδοση που επιτεύχθηκε μέσω του συστήματος δεν είναι ικανοποιητική είναι σαφώς καλύτερη από την απόδοση στην περίπτωση της τυχαίας επιλογής μίας εκ των τεσσάρων κλάσεων του προβλήματος (η απόδοση αυτή ισούται με 25%).

Υπάρχουν αρκετοί λόγοι για τη χαμηλή απόδοση του συστήματος στο εν λόγω πρόβλημα – στη συνέχεια θα αναφερθούμε εκτενέστερα σ’ αυτούς. Ο πιο σημαντικός λόγος είναι η ποιότητα των διαθέσιμων ειδησεογραφικών δεδομένων και η χαμηλή τους συσχέτιση με το πρόβλημα που εξετάζουμε. Για παράδειγμα, στο σύνολο εκπαίδευσης ανήκουν και στιγμιότυπα που αφορούν ειδήσεις που δεν σχετίζονται άμεσα με την Ελλάδα και κατά συνέπεια η επίδρασή τους στο συγκεκριμένο πρόβλημα είναι μηδαμινή. Ένα τέτοιο παράδειγμα στιγμιότυπου εκπαίδευσης φαίνεται στο Σχήμα 4.1.

ΒΡΕΤΑΝΙΑ - ΑΡΓΕΝΤΙΝΗ Χωρίς αναφορές στον Πόλεμο των Φόκλαντς η ιστορική επίσκεψη του ΤΟΝΙ ΜΠΛΕΡ στην Αργεντινή Σύντομη, αλλά ιστορική, ήταν η χτεσινή επίσκεψη (01/8) του ΤΟΝΙ ΜΠΛΕΡ στην Αργεντινή, η πρώτη Βρετανού Πρωθυπουργού μετά τον Πόλεμο των Φόκλαντς, το 1982. Κατά τη διάρκεια των επαφών του δεν έθιξε το ζήτημα της κυριαρχίας των νησιών, αλλά εξέφρασε την υποστήριξή του στις προσπάθειες της Αργεντινής να ξεπεράσει την σοβαρή οικονομική κρίση. Εξάλλου "ιστορική" χαρακτήρισε ο Πρόεδρος της Αργεντινής, ΦΕΡΝΑΝΤΟ ΝΤΕ ΛΑ ΡΟΥΑ, την πρώτη αυτή επίσκεψη Βρετανού Πρωθυπουργού στην Αργεντινή. Το Λονδίνο και το Μπουένος Αΐρες είχαν συμφωνήσει να μη συζητηθεί κατά τη συνάντηση των δύο πολιτικών το ζήτημα της κυριαρχίας των Νήσων Φόκλαντς, τα οποία η Αργεντινή ονομάζει Μαλβίνες και τα διεκδικεί από το 1833. Ωστόσο, την ώρα που περνούσε τα σύνορα προς την Αργεντινή, ο ΜΠΛΕΡ, ήρθε αντιμέτωπος με μια πινακίδα όπου αναγραφόταν: "Las Malvinas son Argentinas" (Οι Μαλβίνες ανήκουν στην Αργεντινή). Σε κοινή συνέντευξη Τύπου που παραχώρησαν κατά τη λήξη της επίσκεψής του, ο ΜΠΛΕΡ δήλωσε ότι δεν έχει τίποτε "χρήσιμο να προσθέσει" για το θέμα αυτό, το οποίο θεωρεί πλέον παρελθόν. "Ό,τι συνέβη στο παρελθόν είναι παρελθόν. Η Αργεντινή ήταν τότε δικτατορία. Σήμερα είναι δημοκρατία", προσθέτοντας ότι τώρα είναι σημαντικό για τη Βρετανία να κοιτάξει το μέλλον και να αναπτύξει καλές εργασιακές σχέσεις με τον πρώην εχθρό της. Φοβούμενος ότι η οικονομική κρίση που μαστίζει την Αργεντινή μπορεί να συμβάλει σε μια ευρύτερη οικονομική ύφεση, ο ΜΠΛΕΡ έσπευσε να εκφράσει την επιδοκιμασία του για τα αντιλαϊκά μέτρα λιτότητας που έλαβε αυτή την εβδομάδα ο Πρόεδρος της Αργεντινής. Μετά τη σύντομη αυτή επίσκεψή του στην Αργεντινή, που διήρκεσε λιγότερο από δύο ώρες, ο ΤΟΝΙ ΜΠΛΕΡ επέστρεψε στο Φος ντε Ιγκουατσού στη γειτονική Βραζιλία για να αναπαυθεί λίγες ώρες προτού αναχωρήσει για το Μεξικό, που αποτελεί τον τελευταίο σταθμό της περιοδείας του στη Λατινική Αμερική. (ΑΠΕ - ΙΝ.GR, 2/8/01) - (Α.Γ.)

Σχήμα 4.1 Παράδειγμα ενός στιγμιότυπου εκπαίδευση που δεν σχετίζεται με το προς επίλυση πρόβλημα

Η καταλυτική επιρροή της ποιότητας του συνόλου εκπαίδευσης (και ελέγχου) στην απόδοση του συστήματος θα φανεί πιο καλά στο πρόβλημα της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου, το οποίο και παρουσιάζουμε ευθύς αμέσως.

4.2 Το πρόβλημα της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου

Όπως έχουμε ήδη αναφέρει το πρόβλημα της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου αφορά στην ταξινόμηση μιας νέας άγνωστης είδησης σε κάποια από τις προκαθορισμένες κλάσεις του προβλήματος. Για τις ανάγκες της μεταπτυχιακής εργασίας χρησιμοποιήθηκε ο δικτυακός τόπος Flash (www.flash.gr) που αποτελεί ένα δημοφιλή τόπο ενημέρωσης για διάφορα θέματα. Πιο συγκεκριμένα χρησιμοποιήσαμε τέσσερις κατηγορίες από το Flash, τις εξής: *Οικονομία*, *Τεχνολογία*, *Αθλητισμός* και

Αυτοκίνητο.

Για καθεμία από τις παραπάνω κατηγορίες κατεβάστηκαν *(download)* από 210 αρχεία ειδησεογραφικών δεδομένων. Για την κατηγορία *Οικονομία* τα δεδομένα προήλθαν από τη διεύθυνση: http://financial.flash.gr/, για την κατηγορία *Τεχνολογία* από τη διεύθυνση: http://tech.flash.gr/, για την κατηγορία *Αθλητισμός* από τη διεύθυνση: http://sportnews.flash.gr/ και για την κατηγορία *Αυτοκίνητο* από τη διεύθυνση: http://automoto.flash.gr/.

Συνολικά, λοιπόν, χρησιμοποιήθηκαν 840 αρχεία ειδησεογραφικών δεδομένων, εκ των οποίων τα 2/3 χρησιμοποιήθηκαν για τη δημιουργία του μοντέλου πρόβλεψης (τα στιγμιότυπα αυτά αποτελούν το σύνολο των στιγμιότυπων εκπαίδευσης) και τα υπόλοιπα 1/3 χρησιμοποιήθηκαν για την αξιολόγηση του μοντέλου πρόβλεψης (τα στιγμιότυπα αυτά αποτελούν το σύνολο των στιγμιότυπων ελέγχου). Η κατανομή των στιγμιότυπων στις διάφορες κλάσεις του προβλήματος και στα σύνολα εκπαίδευσης και ελέγχου φαίνεται στον ακόλουθο πίνακα (Πίνακας 4.3)

Κλάση	**Σύνολο αρχείων της κλάσης**	**Σύνολο αρχείων εκπαίδευσης**	**Σύνολο αρχείων ελέγχου**
Οικονομία	210	140	70
Τεχνολογία	210	140	70
Αθλητισμός	210	140	70
Αυτοκίνητο	210	140	70

Πίνακας 4.3 Η κατανομή των αρχείων στις διάφορες κλάσεις του προβλήματος της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου και στα σύνολα εκπαίδευσης και ελέγχου

Αρχικά λοιπόν, χτίστηκε το μοντέλο πρόβλεψης μέσω των στιγμιότυπων του συνόλου εκπαίδευσης (σύνολο 560 στιγμιότυπα εκπαίδευσης) και στη συνέχεια αξιολογήθηκε η απόδοσή του μέσω των στιγμιότυπων του συνόλου ελέγχου (σύνολο 280 στιγμιότυπα). Τα αποτελέσματα των πειραμάτων φαίνονται στο ακόλουθο σχήμα (Πίνακας 4.4).

Κλάση	Σύνολο σωστά προβλεπόμενων αρχείων	Σύνολο λάθος προβλεπόμενων αρχείων	Ποσοστό επιτυχίας %
Οικονομία	64	6	91,43
Τεχνολογία	63	7	90
Αθλητισμός	64	6	91,43
Αυτοκίνητο	63	7	90

Πίνακας 4.4 Τα αποτελέσματα των πειραμάτων για το πρόβλημα της ταξινόμησης των ειδήσεων του δικτυακού τόπου Flash.

Παρατηρούμε λοιπόν πως η μέση απόδοση του μοντέλου για το συγκεκριμένο πρόβλημα είναι 90,21%. Και πάλι η μέση απόδοση ορίζεται ως το πλήθος των σωστά ταξινομημένων ειδήσεων του συνόλου ελέγχου προς το πλήθος των ειδήσεων του συνόλου ελέγχου (σωστά ταξινομημένων και μη). Η απόδοση αυτή είναι σαφώς καλύτερη (σχεδόν διπλάσια) από την απόδοση στην περίπτωση του προβλήματος της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού.

4.3 Αξιολόγηση αποτελεσμάτων

Είδαμε λοιπόν ότι στην περίπτωση του προβλήματος της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού η απόδοση του συστήματος είναι χαμηλή (46,26%), σε αντίθεση με την περίπτωση του προβλήματος της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου όπου η απόδοση είναι πολύ καλή (90,21%).

Η βασική αιτία για τη διαφορά αυτή είναι η ποιότητα των στιγμιότυπων που χρησιμοποιήθηκαν σε κάθε πρόβλημα για τη δημιουργία των συνόλων εκπαίδευσης και ελέγχου. Στην περίπτωση του προβλήματος της τουριστικής κίνησης τα στιγμιότυπα των ειδησεογραφικών δεδομένων περιλαμβάνουν και ειδήσεις που δεν έχουν καμία σχέση ή σχετίζονται ελάχιστα με το συγκεκριμένο πρόβλημα. Ένα παράδειγμα μιας τέτοιας είδησης παρουσιάζεται στο Σχήμα 4.1 – πρόκειται για μία είδηση που αφορά την επίσκεψη του Βρετανού πρωθυπουργού στην Αργεντινή και η οποία δεν έχει καμία σχέση με το πρόβλημα της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού. Δυστυχώς στο εν λόγω πρόβλημα υπάρχουν πολλά τέτοια παραδείγματα "άσχετων" ειδήσεων. Αντιθέτως, στην περίπτωση του προβλήματος της ταξινόμησης των ειδήσεων του δικτυακού τόπου Flash τα σύνολα εκπαίδευσης και ελέγχου είναι πιο προσεκτικά επιλεγμένα και σχετίζονται άμεσα με το προς εξέταση πρόβλημα.

Άρα λοιπόν, η ποιότητα των στιγμιότυπων του συνόλου εκπαίδευσης καθορίζει άμεσα την απόδοση του μοντέλου πρόβλεψης – με τον όρο ποιότητα εννοούμε τη σχετικότητα των στιγμιότυπων του συνόλου εκπαίδευσης με το προς εξέταση πρόβλημα. Όσο πιο αντιπροσωπευτικά του προβλήματος είναι τα στιγμιότυπα του συνόλου εκπαίδευσης (και ελέγχου) τόσο καλύτερη θα είναι και η απόδοση του ταξινομητή. Προφανώς, μπορεί να υπάρχει κάποιος θόρυβος στο σύνολο των στιγμιότυπων εκπαίδευσης, δηλαδή κάποια στιγμιότυπα που δεν σχετίζονται ή σχετίζονται ελάχιστα με το πρόβλημα, αρκεί βέβαια ο θόρυβος αυτός να μην είναι κανόνας και να μην αφορά την πλειοψηφία των στιγμιότυπων του συνόλου εκπαίδευσης.

Μια άλλη αιτία για τη χαμηλή απόδοση του συστήματος στην περίπτωση του προβλήματος των παραβιάσεων στον ελληνικό εναέριο χώρο είναι και η κατανομή των στιγμιότυπων του συνόλου εκπαίδευσης στις διάφορες κλάσεις του προβλήματος.

Στη περίπτωση, λοιπόν, του προβλήματος της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού οι κλάσεις του προβλήματος καθορίζονται με βάση το πλήθος των επισκέψεων που σημειώθηκαν ανά μήνα κατά τη διάρκεια του έτους 2001. Και η κλάση ενός στιγμιότυπου εκπαίδευσης προκύπτει έμμεσα από την ημερομηνία έκδοσης της είδησης και το πλήθος των επισκέψεων που σημειώθηκαν στον αντίστοιχο μήνα.

Η απόδοση της ίδιας κλάσης σε όλα τα στιγμιότυπα ενός μήνα δεν είναι πολύ δίκαιη καθώς μπορεί σε κάποιες μέρες του μήνα οι επισκέψεις να είναι πιο έντονες ενώ σε κάποιες άλλες όχι. Μια πιο σωστή προσέγγιση θα ήταν να καθορίσουμε τις κλάσεις με βάση τις επισκέψεις που σημειώθηκαν ανά μέρα (ή έστω σε χρονικό διάστημα μικρότερο του μήνα). Ωστόσο, ακολουθήθηκε η προσέγγιση ανά μήνα επειδή δεν υπήρχαν διαθέσιμες οι επισκέψεις ανά ημέρα.

Αν πάντως υπήρχαν διαθέσιμες οι επισκέψεις ανά ημέρα τότε μπορούσαν να ομαδοποιηθούν οι μέρες με βάση τις επισκέψεις που σημειώθηκαν σε κάθε μέρα και εν συνεχεία να προκύψουν οι κλάσεις οι οποίες πλέον δε θα αφορούν ολόκληρους μήνες αλλά διαστήματα των μηνών ή ακόμα και μεμονωμένες μέρες.

Το πρόβλημα αυτό δεν υφίσταται στην περίπτωση του προβλήματος της ταξινόμησης των ειδήσεων του δικτυακού τόπου Flash καθώς οι κατηγορίες του προβλήματος προκύπτουν με άμεσο τρόπο από τις διάφορες πραγματικές κατηγορίες ειδήσεων του Flash. Η απόδοση ενός στιγμιότυπου σε κάποια κλάση προκύπτει άμεσα από τη διεύθυνση από την οποία προέρχεται το στιγμιότυπο. Για παράδειγμα ένα στιγμιότυπο που προέρχεται από τη διεύθυνση: http://financial.flash.gr/ αποδίδεται στην κλάση *Οικονομία*. Δηλαδή η αντιστοίχιση των στιγμιότυπων στις διάφορες κλάσεις του προβλήματος έγινε με βάση το περιεχόμενό τους και όχι με κάποιο έμμεσο τρόπο όπως έγινε στην περίπτωση του προβλήματος της πρόβλεψης της

τουριστικής κίνησης, όπου για την αντιστοίχιση χρησιμοποιήθηκε η ημερομηνία έκδοσης της είδησης.

Πιστεύουμε λοιπόν πως με ένα καλύτερο σύνολο στιγμιότυπων εκπαίδευσης και ελέγχου και μια πιο δίκαια κατανομή των στιγμιότυπων στις κλάσεις του προβλήματος η απόδοση του συστήματος για το πρόβλημα της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού θα αυξηθεί σημαντικά.

Στη συνέχεια αξίζει να δούμε κάποια παραδείγματα σωστής και λανθασμένης πρόβλεψης – τα παραδείγματα αυτά αφορούν του πρόβλημα της ταξινόμησης των ειδήσεων του δικτυακού τόπου Flash.

Ας ξεκινήσουμε με ένα παράδειγμα ορθής πρόβλεψης για την κλάση *Οικονομία*. Το στιγμιότυπο (Σχήμα 4.2) προέρχεται από τη διεύθυνση: http://financial.flash.gr/ και κατά συνέπεια θα έπρεπε το σύστημα να το κατατάξει στην κλάση *Οικονομία* όπως και έγινε.

Σε δύο εβδομάδες αναμένεται να ξεκινήσει η διαδικασία του διεθνούς πλειοδοτικού διαγωνισμού για την πώληση του 17% που κατέχει η Εμπορική στην Τράπεζα Αττικής. Εντός των επομένων ημερών, οι σύμβουλοι, η J. P. Morgan και η Τράπεζα Επενδύσεων θα παραδώσουν τον φάκελο, στον οποίο θα θα περιλαμβάνονται οι όροι βάσει των οποίων θα διεξαχθεί ο διαγωνισμός.

Παράλληλα, εξετάζεται και το ενδεχόμενο πώλησης μέρους του ποσοστού που κατέχουν το ΤΣΜΕΔΕ και το Ταμείο Παρακαταθηκών και Δανείων στην τράπεζα. Αυτό συναρτάται με την εκχώρηση του μάνατζμεντ της Αττικής.

Πρακτικά καταβάλλεται προσπάθεια ώστε σε μεσοπρόθεσμη βάση τα ταμεία να αποδεχτούν να πωλήσουν πακέτα μετοχών τους, τα οποία θα ήταν δυνατό να εξαγοράσει ο στρατηγικός επενδυτικής με option.

Σημειώνεται ότι πρόθεση της διοίκησης της Εμπορικής σε ότι αφορά στη μεταβίβαση του πακέτου της Αττικής, είναι να έχουν ολοκληρωθεί οι διαδικασίες το Σεπτέμβριο.

Σχήμα 4.2 Παράδειγμα σωστής πρόβλεψης για την κλάση *Οικονομία*

Είναι εύκολο να εξηγήσουμε γιατί το σύστημα κατέταξε το στιγμιότυπο στην κλάση *Οικονομία*. Οι λέξεις "πώληση", "τράπεζα", "δάνειο", "πακέτο" για παράδειγμα είναι αντιπροσωπευτικές της κλάσης *Οικονομία* και εμφανίζονται σ' αυτή με συχνότητα πολύ πιο μεγάλη από τη συχνότητα εμφάνισής τους στις υπόλοιπες κλάσεις του προβλήματος: *Τεχνολογία*, *Αθλητισμός* και *Αυτοκίνητο*. Συνεπώς, το κείμενο της είδησης έχει μεγαλύτερη πιθανότητα να εμφανιστεί στην κλάση *Οικονομία* και γι' αυτό το λόγο κατατάσσεται από το σύστημα στη συγκεκριμένη κλάση.

Συνεχίζοντας ας δούμε ένα παράδειγμα λανθασμένης πρόβλεψης για την κλάση *Οικονομία*. Το στιγμιότυπο (Σχήμα 4.3) προέρχεται από τη διεύθυνση: http://financial.flash.gr/ και κατά συνέπεια θα έπρεπε το σύστημα να το κατατάξει στην κλάση *Οικονομία* . Το σύστημα όμως το κατέταξε στην κλάση *Τεχνολογία*.

Η συνεργασία με την Interwise δίνει τη δυνατότητα στην Unisystems να προσφέρει υψηλού επιπέδου λύσεις εκπαίδευσης μέσω Ιντερνετ (e-learning) προς κάθε ευρύ σύνολο εργαζομένων, υποστηρίζοντας τις αυξανόμενες διδακτικές ανάγκες των μεγάλων επιχειρήσεων.

Με βάση τη συμφωνία, η Unisystems αναλαμβάνει να παρέχει προς κάθε ενδιαφερόμενο την πλατφόρμα Enterprise Communication Platform (ECP) της Interwise. Η Interwise δραστηριοποιείται στις πλατφόρμες e-learning και εταιρικής επικοινωνίας. Ιδρύθηκε το 1994 και εδρεύει στην Santa Clara της Καλιφόρνιας, ενώ διαθέτει εταιρική παρουσία σε 9 ακόμα σημαντικές αγορές σε όλο τον κόσμο.

Σχήμα 4.3 Παράδειγμα λανθασμένης πρόβλεψης για την κλάση *Οικονομία*

Και πάλι μπορούμε να εξηγήσουμε γιατί το σύστημα κατέταξε το στιγμιότυπο στην κλάση *Τεχνολογία* αντί της κλάσης *Οικονομία*. Μελετώντας το κείμενο της είδησης παρατηρούμε ότι δεν είναι ένα αμιγώς οικονομικό κείμενο αλλά σχετίζεται άμεσα με την τεχνολογία. Πιο συγκεκριμένα, οι λέξεις "Ιντερνετ", "e-learning", "πλατφόρμα" κ.λ.π. είναι αντιπροσωπευτικές της κλάσης *Τεχνολογία* και εμφανίζονται σ' αυτή με συχνότητα πολύ πιο μεγάλη από τη συχνότητα εμφάνισής τους στην κλάση *Οικονομία*. Από την άλλη υπάρχουν και αρκετές λέξεις αντιπροσωπευτικές της κλάσης *Οικονομία* όπως για παράδειγμα οι λέξεις "επιχειρήσεων", "εταιρική παρουσία", "αγορές" κ.λ.π. Απλά στην προκειμένη περίπτωση συμβαίνει η πιθανότητα εμφάνισης της κλάσης *Τεχνολογία* να είναι μεγαλύτερη από την πιθανότητα εμφάνισης της κλάσης *Οικονομία* (να αναφέρουμε εδώ πως η διαφορά είναι μικρή) και για το λόγο αυτό η κλάση *Τεχνολογία* είναι αυτή που προβλέπεται από το σύστημα.

Ας δούμε στη συνέχεια ένα παράδειγμα ορθής πρόβλεψης για την κλάση *Τεχνολογία*. Το στιγμιότυπο (Σχήμα 4.4) προέρχεται από τη διεύθυνση: http://tech.flash.gr/ και κατά συνέπεια θα έπρεπε το σύστημα να το κατατάξει στην κλάση *Τεχνολογία* όπως και έγινε.

Ενας από τους σημαντικότερους τομείς, στον οποίο χρησιμοποιούνται εκτενώς οι ηλεκτρονικοί υπολογιστές, είναι αυτός της μηχανογράφησης. Η ηλεκτρονική μηχανογράφηση συνοψίζεται, σε γενικές γραμμές, στη δημιουργία βάσεων δεδομένων (databases), στις οποίες καταχωρείται όγκος πληροφοριών. Με αυτόν τον τρόπο, η ταξινόμηση και η αναζήτηση συγκεκριμένων αντικειμένων, που άλλοτε απαιτούσε ώρες περιπλάνησης σε σκονισμένες βιβλιοθήκες, μετατρέπεται σε υπόθεση μερικών δευτερολέπτων.

Τα τελευταία χρόνια, οι μεγαλύτερες εταιρείες στο χώρο της Πληροφορικής έχουν προσπαθήσει να καθιερώσουν τη δική τους λύση στην αγορά, προσφέροντας διάφορα προγράμματα διαχείρισης βάσεων δεδομένων. Ανάμεσα τους η Microsoft, η Lotus, η Borland (νυν Inprise) και η Oracle. Ιδίως η τελευταία φημίζεται για την ταχύτητα διαχείρισης βάσεων δεδομένων που προσφέρει το πακέτο της.

Αυτός ο εταιρικός ανταγωνισμός δημιούργησε διάφορα προβλήματα στην κοινότητα των προγραμματιστών. Η κάθε εταιρεία υλοποίησε το δικό της database format, με αποτέλεσμα να απαιτείται από τον προγραμματιστή γνώση τεχνικών λεπτομερειών - των οποίων η τεκμηρίωση, συνήθως, είναι ανύπαρκτη. Σκεφθείτε, για παράδειγμα, μια εφαρμογή μηχανογράφησης που χρησιμοποιεί βάσεις δεδομένων Access (Microsoft). Για κάποιο λόγο, αποφασίζεται από την εταιρεία ότι τα δεδομένα πρέπει να αποθηκεύονται σε Oracle αντί για Access. Αυτό σημαίνει ότι πρέπει, με κάποιον τρόπο, όλη η πληροφορία που τηρείται σε Access, να μετατραπεί, ώστε να αποθηκευθεί σε βάσεις Oracle. Από την άλλη, το πρόγραμμα μηχανογράφησης που αρχικά χρησιμοποιείτο, είναι πλέον άχρηστο, μια και μπορούσε να διαχειριστεί μόνο βάσεις τύπου Access. Οπως καταλαβαίνετε, τα παραπάνω δημιουργούν μεγάλα οικονομικά προβλήματα στη συγκεκριμένη εταιρεία.

Σχήμα 4.4 Παράδειγμα σωστής πρόβλεψης για την κλάση *Τεχνολογία*

Μπορούμε εύκολα να εξηγήσουμε γιατί το σύστημα κατέταξε το στιγμιότυπο στην κλάση *Τεχνολογία*. Οι λέξεις "ηλεκτρονικοί υπολογιστές", "μηχανογράφησης", "βάσεις δεδομένων", "προγραμματιστών" για παράδειγμα είναι αντιπροσωπευτικές της κλάσης *Τεχνολογία* και εμφανίζονται σ' αυτή με συχνότητα πολύ πιο μεγάλη από τη συχνότητα εμφάνισής τους στις υπόλοιπες κλάσεις του προβλήματος: *Οικονομία*, *Αθλητισμός* και *Αυτοκίνητο*. Συνεπώς το κείμενο της είδησης έχει μεγαλύτερη πιθανότητα εμφάνισης στην κλάση *Τεχνολογία* και γι' αυτό το λόγο κατατάσσεται από το σύστημα στη συγκεκριμένη κλάση.

Συνεχίζοντας ας δούμε ένα παράδειγμα λανθασμένης πρόβλεψης για την κλάση *Τεχνολογία*. Το στιγμιότυπο (Σχήμα 4.5) προέρχεται από τη διεύθυνση:

http://tech.flash.gr/ και κατά συνέπεια θα έπρεπε το σύστημα να το κατατάξει στην κλάση *Τεχνολογία* . Το σύστημα όμως το κατέταξε στην κλάση *Οικονομία*.

> Εμπορική συμφωνία συνεργασίας με την εταιρία Πλαίσιο για την προώθηση των κινητών τηλεφώνων και γνήσιων αξεσουάρ Nokia στην Ελληνική Αγορά για το έτος 2003 υπέγραψε η Alpha Copy.
>
> Η συνεργασία θα έχει ως αποτέλεσμα την πλήρη παρουσίαση όλων των κινητών τηλεφώνων και γνήσιων αξεσουάρ Nokia και μέσα από τα καταστήματα που διαθέτει το Πλαίσιο. Τόσο η Alpha Copy όσο και το Πλαίσιο έχουν σκοπό να προσφέρουν αντίστοιχα προϊόντα και υπηρεσίες που θα ικανοποιήσουν πληρέστερα την εξυπηρέτηση των καταναλωτών κινητών τηλεφώνων Nokia.
>
> Με την συνέχιση της συνεργασίας, θα ξεκινήσουν όλες οι απαραίτητες εκπαιδεύσεις των στελεχών με στόχο την καλύτερη και ποιοτικότερη εξυπηρέτηση και ενημέρωση του πελάτη και την αναβάθμιση όλων των προσφερόμενων ψηφιακών υπηρεσιών. Στόχοι και των δύο εταιριών είναι να καλύπτουν άμεσα και ταχύτατα τις ανάγκες σε προϊόντα και υπηρεσίες στο χώρο της κινητής τηλεφωνίας.

Σχήμα 4.5 Παράδειγμα λανθασμένης πρόβλεψης για την κλάση *Τεχνολογία*

Και πάλι μπορούμε να εξηγήσουμε γιατί το σύστημα κατέταξε το στιγμιότυπο στην κλάση *Οικονομία* αντί της κλάσης *Τεχνολογία*. Μελετώντας το κείμενο της είδησης παρατηρούμε ότι δεν είναι ένα αμιγώς τεχνολογικό αλλά σχετίζεται άμεσα και με την οικονομία καθώς αφορά τη συνεργασία δύο εταιρειών. Πιο συγκεκριμένα, οι λέξεις "Εμπορική συμφωνία συνεργασίας", "εταιρία", "στελέχη" κ.λ.π. είναι αντιπροσωπευτικές της κλάσης *Οικονομία* και εμφανίζονται σ' αυτή με συχνότητα πολύ πιο μεγάλη από τη συχνότητα εμφάνισής τους στην κλάση *Τεχνολογία*. Από την άλλη υπάρχουν και αρκετές λέξεις αντιπροσωπευτικές της κλάσης *Τεχνολογία* όπως για παράδειγμα οι λέξεις "κινητών τηλεφώνων", "ψηφιακών υπηρεσιών". Στην προκειμένη περίπτωση όμως, συμβαίνει η πιθανότητα εμφάνισης της κλάσης *Οικονομία* να είναι μεγαλύτερη από την πιθανότητα εμφάνισης της κλάσης *Τεχνολογία* (και πάλι η διαφορά των δύο πιθανοτήτων είναι μικρή) και για το λόγο αυτό η κλάση *Οικονομία* είναι αυτή που προβλέπεται από το σύστημα.

Ας δούμε στη συνέχεια ένα παράδειγμα ορθής πρόβλεψης για την κλάση *Αθλητισμός*. Το στιγμιότυπο (Σχήμα 4.6) προέρχεται από τη διεύθυνση: http://sportnews.flash.gr/ και κατά συνέπεια θα έπρεπε το σύστημα να το κατατάξει στην κλάση *Αθλητισμός* όπως και έγινε.

Παναχαϊκή και Προοδευτική αναδείχτηκαν ισόπαλες με 1-1 στην Πάτρα και το τελικό αποτέλεσμα αφήνει ικανοποιημένους μάλλον τους φιλοξενούμενους. Η ομάδα του Σούλη Παπαδόπουλου παρέμεινε στις πρώτες θέσεις της βαθμολογίας και έδωσε συνέχεια στην πολύ καλή φετινή της πορεία σε αντίθεση με το αχαϊκό συγκρότημα που παρέμεινε στον... πάτο της κατάταξης. Μετά από ένα... άσφαιρο, όσο και απελπιστικά άχρωμο πρώτο ημίχρονο (μοναδική καλή στιγμή στο 38ο λεπτό, όταν από την καρφωτή κεφαλιά του Πάντου, ο Μπαλτιμάς έσωσε την εστία του), η Παναχαϊκή ευτύχησε ν ανοίξει το σκορ στο 60ο λεπτό της συνάντησης. Ο Στεφανής, που είχε μπει ως αλλαγή στο 40 στη θέση του Κωνσταντινίδη, έπιασε δυνατό διαγώνιο σουτ και η μπάλα κατέληξε στο βάθος της εστίας του Κόη (1-0). Οι φιλοξενούμενοι έφεραν το παιχνίδι στα ίσα επτά λεπτά αργότερα όταν ο Ζαϊμι κέρδισε την εσχάτη των ποινών στην ανατροπή από τον Μπαντά, την οποίο κλήθηκε να εκτελέσει ο Μάγγος που με περίσσεια ψυχραιμία δεν δυσκολεύτηκε να παραβιάσει την εστία του Μπαλτιμά. Η Παναχαϊκή με την οποία έκαναν το ντεμπούτο τους οι Αρμένιοι αδελφοί Καραμιάν είχε την ευκαιρία να πάρει τους τρεις βαθμούς στο 84, όμως το βολέ του Μορίνι έξω από την περιοχή, κατέληξε στο αριστερό δοκάρι της εστίας του Κόη.

Ο Τεροβίτσας κιτρίνισε τον Ανδράλα.

ΠΑΝΑΧΑΙΚΗ: Μπαλτιμάς, Λυγνός, Μπαντάς, Γκουζιώτης, Ιωάννου, Στοίνοβιτς, Αρταβάζντ Ιστιόλ, Καραμιάν, Κωνσταντινίδης (40` Στεφανής), Μπονάνι, Αρμάν Καραμιάν ΠΡΟΟΔΕΥΤΙΚΗ: Κόης, Καλλιμάνης, Κουλοχέρης, Πουλόπουλος, Αγγελόπουλος, Σιδηρόπουλος, Αντωνόπουλος, Μάγγος, Πάντος, Τάτσης (88` Προβίδας), Ζαΐμι (88` Ανδράλας).

Σχήμα 4.6 Παράδειγμα σωστής πρόβλεψης για την κλάση *Αθλητισμός*

Μπορούμε να εξηγήσουμε γιατί το σύστημα κατέταξε το στιγμιότυπο στην κλάση *Αθλητισμός*. Οι λέξεις "ομάδα", "ημίχρονο", "δοκάρι", "εστία" για παράδειγμα είναι αντιπροσωπευτικές της κλάσης *Αθλητισμός* και εμφανίζονται σ' αυτή με συχνότητα πολύ πιο μεγάλη από τη συχνότητα εμφάνισής τους στις υπόλοιπες κλάσεις: *Οικονομία*, *Τεχνολογία* και *Αυτοκίνητο*. Συνεπώς το κείμενο της είδησης έχει μεγαλύτερη πιθανότητα να εμφανιστεί στην κλάση *Αθλητισμός* και γι' αυτό το λόγο κατατάσσεται από το σύστημα στη συγκεκριμένη κλάση.

Συνεχίζοντας ας δούμε ένα παράδειγμα λανθασμένης πρόβλεψης για την κλάση *Αθλητισμός*. Το στιγμιότυπο (Σχήμα 4.7) προέρχεται από τη διεύθυνση: http://sportnews.flash.gr/ και κατά συνέπεια θα έπρεπε το σύστημα να το κατατάξει στην κλάση *Αθλητισμός*. Το σύστημα όμως το κατέταξε στην κλάση *Οικονομία*.

ΠΑΤΡΑ: Τα χρέη της Παναχαϊκής δεν είναι τελικά στα 850 εκατομμύρια δραχμές που επιμόνως διατυμπάνιζε ο πρόεδρος της ΠΑΕ Αρης Λουκόπουλος, αλλά ούτε καν ένα δις. Είναι πολλά περισσότερα και συγκεκριμένα κάτι περισσότερο από δύο δις σύμφωνα με λεπτομερή έρευνα που έκανε ο πρώην δήμαρχος Πατρέων Ευάγγελος Φλωράτος, ο οποίος αρχικά είχε αποδεχτεί πρόταση που του είχε γίνει να αναλάβει επικεφαλής της Διοικούσας Επιτροπής που θα όριζε το πρωτοδικείο με στόχο να βγάλει τους "κοκκινόμαυρους" από το αδιέξοδο.
Οπως αναφέρει στην ανακοίνωση που εξέδωσε ο κ. Φλωράτος, μόνο στην Εφορεία η Παναχαϊκή έχει οφειλές που αγγίζουν το ένα δις! Επειτα μάλιστα από αυτή την εξέλιξη, όπως τονίζεται στην ίδια ανακοίνωση, ο κ. Φλωράτος αδυνατεί να βρει λύση με αυτά τα δεδομένα, καθώς οι υποψήφιοι να αναλάβουν την ΠΑΕ έκαναν πίσω και απέσυραν το ενδιαφέρον τους, όταν γνωστοποιήθηκε το πραγματικό ύψος των οφειλών.
Αναλυτικά το πλήρες κείμενο της ανακοίνωσης του Ευάγγελου Φλωράτου έχει ως εξής:
"Σύμφωνα με την ισχύουσα νομοθεσία όλες οι ΠΑΕ για να πάρουν άδεια συμμετοχής στο νέο πρωτάθλημα 2003-2004, από την επιτροπή επαγγελματικού Αθλητισμού, θα πρέπει να έχουν εξοφλήσει ή ρυθμίσει τις υποχρεώσεις τους προς το ΙΚΑ και την Εφορία και να έχουν καταβάλλει τις μέχρι τον Ιούνιο υποχρεώσεις τους προς τους ποδοσφαιριστές.

Οι σημερινές οικονομικές υποχρεώσεις της ΠΑΕ είναι οι πιο κάτω:
- Συμβόλαια ποδοσφαιριστών 938.016 ευρώ άμεσα απαιτητά
- ΙΚΑ 100.000 ευρώ υπό ρύθμιση
- Φόροι βεβαιωμένοι 934.550 ευρώ υπό ρύθμιση
- Φόροι μη βεβαιωμένοι 70.000 ευρώ
- Πιστωτές 563.134 ευρώ
- Απρόβλεπτα 394.300 ευρώ
==========
Σύνολο: 3.000.000 ευρώ

Μετά τη δυσμενή αυτή εξέλιξη με λύπη πιστεύουμε ότι είναι σχεδόν αδύνατο η εξεύρεση επενδυτού ο οποίος θα αναλάβει το οικονομικό αυτό βάρος για την εξυγείανση της ιστορικής μας ομάδας.

Σχήμα 4.7 Παράδειγμα λανθασμένης πρόβλεψης για την κλάση *Αθλητισμός*

Και πάλι μπορούμε να εξηγήσουμε γιατί το σύστημα κατέταξε το στιγμιότυπο στην κλάση *Οικονομία* αντί της κλάσης *Αθλητισμός*. Μελετώντας το κείμενο της είδησης παρατηρούμε ότι δεν είναι ένα αμιγώς αθλητικό αλλά σχετίζεται άμεσα και με την οικονομία καθώς αφορά την οικονομική κατάσταση μιας συγκεκριμένης ποδοσφαιρικής ομάδας. Πιο συγκεκριμένα, οι λέξεις "χρέη", "οφειλές", "οικονομικές υποχρεώσεις", "φόροι", "πιστωτές" κ.λ.π. είναι αντιπροσωπευτικές της κλάσης

Οικονομία και εμφανίζονται σ' αυτή με συχνότητα πολύ πιο μεγάλη από τη συχνότητα εμφάνισής τους στην κλάση *Αθλητισμός*. Από την άλλη υπάρχουν και αρκετές λέξεις αντιπροσωπευτικές της κλάσης *Αθλητισμός* όπως για παράδειγμα οι λέξεις "ΠΑΕ", "ποδοσφαιριστών", "ομάδας" κ.λ.π. Στην προκειμένη περίπτωση όμως, συμβαίνει η πιθανότητα εμφάνισης της κλάσης *Οικονομία* να είναι μεγαλύτερη από την πιθανότητα εμφάνισης της κλάσης *Αθλητισμός* (και πάλι η διαφορά των δύο πιθανοτήτων είναι μικρή) και για το λόγο αυτό η κλάση *Οικονομία* είναι αυτή που προβλέπεται από το σύστημα.

Ας δούμε τέλος ένα παράδειγμα ορθής πρόβλεψης για την κλάση *Αυτοκίνητο*. Το στιγμιότυπο (Σχήμα 4.8) προέρχεται από τη διεύθυνση: http://automoto.flash.gr/ και κατά συνέπεια θα έπρεπε το σύστημα να το κατατάξει στην κλάση *Αυτοκίνητο* όπως και έγινε.

Duratec SCiΠιθανότατα από το ερχόμενο φθινόπωρο θα είναι διαθέσιμη στην αγορά, η νέα έκδοση του Ford Mondeo που θα εφοδιάζεται με τον κινητήρα Duratec SCi (Smart Charge injection), που θα έχει απόδοση 130 ίππων. Ο κινητήρας SCi διαθέτει άμεσο ψεκασμό και εκμεταλλεύεται στο έπακρο τα όποια πλεονεκτήματα του φτωχού μείγματος, στις χαμηλές στροφές και σε συνθήκες μικρού φορτίου. Το εργοστάσιο ανακοινώνει μείωση της τιμής της μέσης κατανάλωσης κατά 6-8% σε σχέση με τον κινητήρα των 1.8 λίτρων που τοποθετείται μέχρι σήμερα στο Mondeo (125 ίπποι) και δεν είναι άμεσου ψεκασμού. Αν κάτι τέτοιο επιβεβαιωθεί στην πράξη, η μέση κατανάλωση του Mondeo SCi θα κυμαίνεται μεταξύ 7,2-7,3 λ/100 χλμ.

Duratec SCiO διευθυντής μηχανολογίας βενζινοκινητήρων της Ford Ευρώπης, Rudolf J. Menne, δήλωσε χαρακτηριστικά πως στόχος του ήταν «η οικονομία στην κατανάλωση, στο μεγαλύτερο δυνατό εύρος συνθηκών λειτουργίας του κινητήρα». Με άλλα λόγια, στόχος της Ford ήταν να αυξήσει το περιθώριο στο οποίο ο κινητήρας θα μπορούσε να λειτουργεί με φτωχό μείγμα. Να σημειωθεί πως για καλύτερη απόκριση στις προσταγές του οδηγού, ο κινητήρας αυτός θα συνδυάζεται με το νέο χειροκίνητο κιβώτιο 6 σχέσεων της Ford, ενώ όπως μας ενημέρωσε η ελληνική αντιπροσωπεία, η έκδοση με τον κινητήρα SCi, δεν αντικαθιστά τον κινητήρα των 125 ίππων, αλλά τουλάχιστον αρχικά θα διατίθεται παράλληλα με αυτόν.

Σχήμα 4.8 Παράδειγμα σωστής πρόβλεψης για την κλάση *Αυτοκίνητο*

Μπορούμε εύκολα να εξηγήσουμε γιατί το σύστημα κατέταξε το στιγμιότυπο στην κλάση *Αυτοκίνητο*. Οι λέξεις "κινητήρα", "ίππων", "βενζινοκινητήρων", "οδηγού" για παράδειγμα είναι αντιπροσωπευτικές της κλάσης *Αυτοκίνητο* και εμφανίζονται σ' αυτή με συχνότητα πολύ πιο μεγάλη από τη συχνότητα εμφάνισής τους στις

υπόλοιπες κλάσεις: *Οικονομία*, *Τεχνολογία* και *Αθλητισμός*. Συνεπώς το κείμενο της είδησης έχει μεγαλύτερη πιθανότητα εμφάνισης στην κλάση *Αυτοκίνητο* και γι' αυτό το λόγο κατατάσσεται από το σύστημα στη συγκεκριμένη κλάση.

Συνεχίζοντας ας δούμε ένα παράδειγμα λανθασμένης πρόβλεψης για την κλάση *Αυτοκίνητο*. Το στιγμιότυπο (Σχήμα 4.9) προέρχεται από τη διεύθυνση: http://automoto.flash.gr/ και κατά συνέπεια θα έπρεπε το σύστημα να το κατατάξει στην κλάση *Αυτοκίνητο* . Το σύστημα όμως το κατέταξε στην κλάση *Αθλητισμός*.

Τελικά, όσοι εξέφρασαν επιφυλάξεις σχετικά με την κατάταξη του επεισοδιακού Γκραν Πρι Βραζιλίας, δικαιώθηκαν. Η FIA, στην έκτακτη συνεδρίασή της το πρωί της Παρασκευής 11/4, με σκοπό την αξιολόγηση των νέων στοιχείων που τέθηκαν στη διάθεσή της μετά τον αγώνα της Βραζιλίας, έκρινε πως ο Φιζικέλα είναι ο νικητής του αγώνα σύμφωνα με τους ισχύοντες κανονισμούς και αφαίρεσε τους 10 βαθμούς της νίκης από τον Ραϊκόνεν. Σύμφωνα με τα στοιχεία αυτά «σε αντίθεση με την πληροφόρηση που υπήρχε από τους χρονομετρητές του Γκραν Πρι Βραζιλίας, το αυτοκίνητο με τον αριθμό 11 (σ.σ. Φιζικέλα), είχε αρχίσει τον 56ο γύρο του αγώνα όταν αυτός διακόπηκε». Ως εκ τούτου, η τελική κατάταξη του αγώνα έπρεπε να είναι αυτή που ίσχυε όταν συμπληρώθηκε ο 54ος γύρος και όχι η κατάταξη μετά τη συμπλήρωση του 53ου γύρου.

Να θυμίσουμε πως σύμφωνα με τους κανονισμούς της F1, όταν ένας αγώνας διακοπεί για κάποιο λόγο, η τελική κατάταξη είναι αυτή που είχαν τα αυτοκίνητα δύο γύρους πριν τη διακοπή. Στη συνεδρίαση της Παρασκευής είχαν κληθεί και συμμετείχαν εκπρόσωποι όλων των ομάδων που μπορεί να επηρεάζονταν από τυχόν αλλαγή της κατάταξης. Το ουσιαστικό πάντως είναι πως ο Φιζικέλα, έστω και μέσω μιας συνεδρίασης του ανώτατου οργάνου διεξαγωγής των αγώνων που πραγματοποιήθηκε 5 ημέρες μετά τον αγώνα, κατέκτησε την πρώτη νίκη της καριέρας του. Από την άλλη πλευρά, ο Κίμι Ραϊκόνεν εξακολουθεί να οδηγεί μέχρι στιγμής την κούρσα του πρωταθλήματος μιά και παραμένει στην πρώτη θέση, αλλά με 24 αντί για 26 βαθμούς, καθώς περιορίστηκε στους 8 βαθμούς της 2ης θέσης.

Σχήμα 4.9 Παράδειγμα λανθασμένης πρόβλεψης για την κλάση *Αυτοκίνητο*

Εύκολα πάλι μπορούμε να εξηγήσουμε γιατί το σύστημα κατέταξε το στιγμιότυπο στην κλάση *Αθλητισμός* αντί της κλάσης *Αυτοκίνητο*. Μελετώντας το κείμενο της είδησης παρατηρούμε ότι δεν αναφέρεται αποκλειστικά σε αυτοκίνητα αλλά σχετίζεται άμεσα και με τον αθλητισμό. Πιο συγκεκριμένα, οι λέξεις "αγώνα", "Γκραν Πρι", "κούρσα", "πρωτάθλημα" κ.λ.π. είναι αντιπροσωπευτικές της κλάσης *Αθλητισμός* και εμφανίζονται σ' αυτή με συχνότητα πολύ πιο μεγάλη από τη συχνότητα εμφάνισής τους στην κλάση *Αυτοκίνητο*.

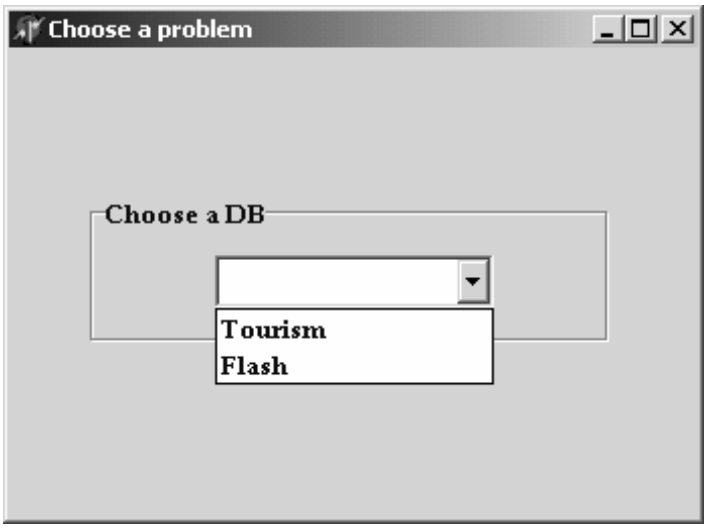
Από την άλλη υπάρχουν και αρκετές λέξεις αντιπροσωπευτικές της κλάσης *Αυτοκίνητο* όπως για παράδειγμα οι λέξεις "αυτοκίνητο", "χρονομετρητές" κ.λ.π. Στην προκειμένη περίπτωση όμως, συμβαίνει η πιθανότητα εμφάνισης της κλάσης *Αθλητισμός* να είναι μεγαλύτερη από την πιθανότητα εμφάνισης της κλάσης *Αυτοκίνητο* (και πάλι η διαφορά των δύο πιθανοτήτων είναι μικρή) και για το λόγο αυτό η κλάση *Αθλητισμός* είναι αυτή που προβλέπεται από το σύστημα.

5. Το περιβάλλον διεπαφής χρήστη *(user interface)*

Στο κεφάλαιο αυτό θα περιγράψουμε τη λειτουργία του συστήματος μέσα από χαρακτηριστικά παραδείγματα τρεξίματος. Πρόκειται στην ουσία για έναν οδηγό χρήσης της εφαρμογής.

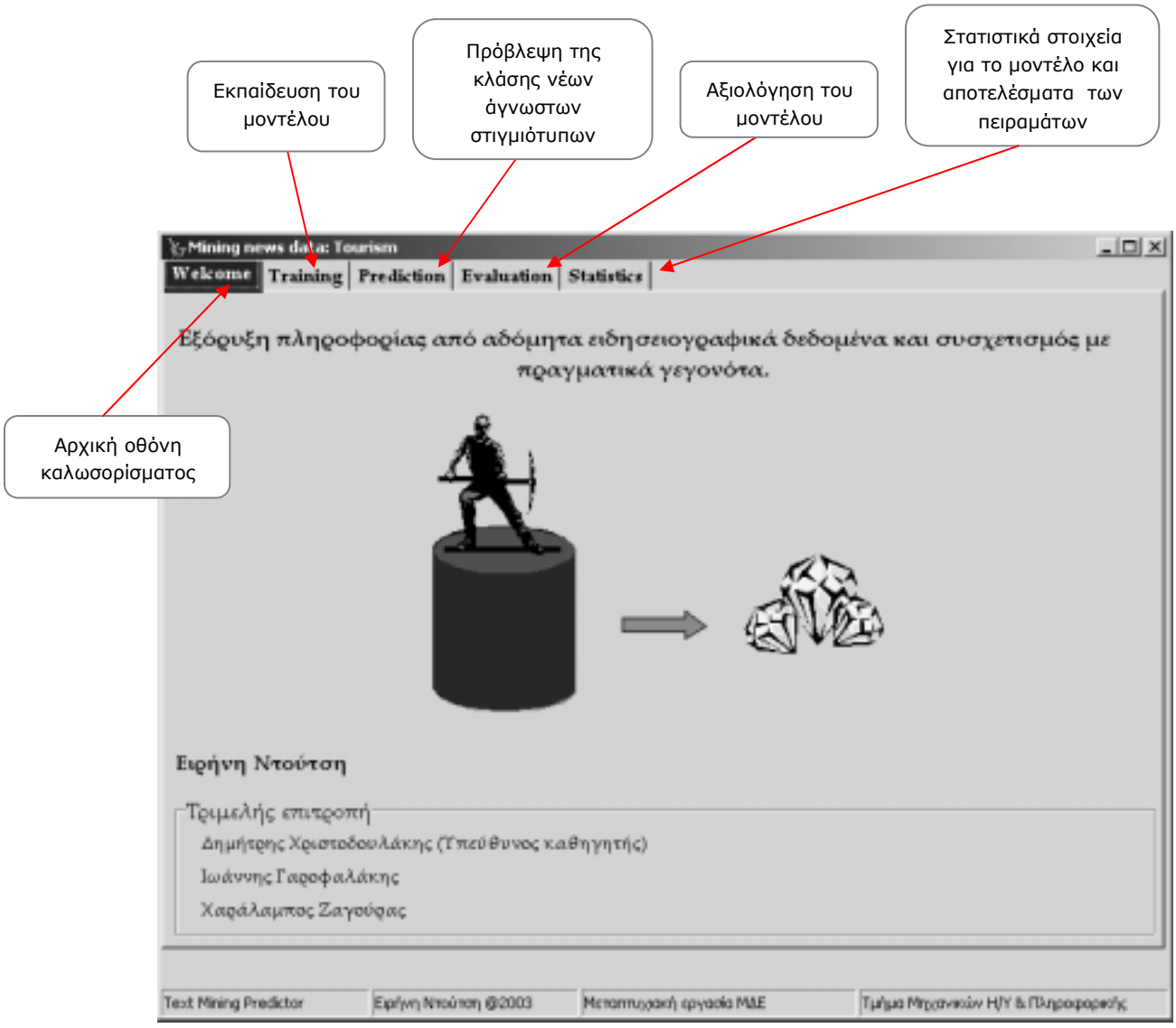
5.1 Εκκίνηση εφαρμογής

Με την εκκίνηση της εφαρμογής ο χρήστης καλείται να επιλέξει ένα από τα δύο προβλήματα: το πρόβλημα της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού (Σχήμα 5.1 - επιλογή "Tourism") ή το πρόβλημα της ταξινόμησης των ειδήσεων ενός δικτυακού τόπου (Σχήμα 5.1 - επιλογή "Flash"). Το βήμα αυτό είναι απαραίτητο καθώς για κάθε πρόβλημα υπάρχει διαφορετική βάση δεδομένων, με την ίδια βέβαια δομή όπως την περιγράψαμε στο Κεφάλαιο 1 του Β μέρους.



Σχήμα 5.1 Επιλογή του προβλήματος

Αμέσως μετά την επιλογή του προβλήματος, εμφανίζεται το κυρίως περιβάλλον της εφαρμογής (Σχήμα 5.2) .



Σχήμα 5.2 Το κυρίως περιβάλλον της εφαρμογής

Οι επιλογές του χρήστη για τη συνέχεια είναι τέσσερις, οι ακόλουθες:

- **Επιλογή "*Training*":** η επιλογή αυτή αφορά την εκπαίδευση του μοντέλου μέσω στιγμιότυπων του συνόλου εκπαίδευσης.
- **Επιλογή "*Prediction*":** η επιλογή αυτή αφορά την πρόβλεψη της κλάσης νέων άγνωστων στιγμιότυπων του προβλήματος.
- **Επιλογή "*Evaluation*":** η επιλογή αυτή αφορά την αξιολόγηση της ικανότητας πρόβλεψης του μοντέλου μέσω στιγμιότυπων του συνόλου ελέγχου.
- **Επιλογή "*Statistics*":** η επιλογή αυτή αφορά την εμφάνιση ενδιαφερόντων στατιστικών στοιχείων σχετικά με το μοντέλο ή τα αποτελέσματα των διαφόρων

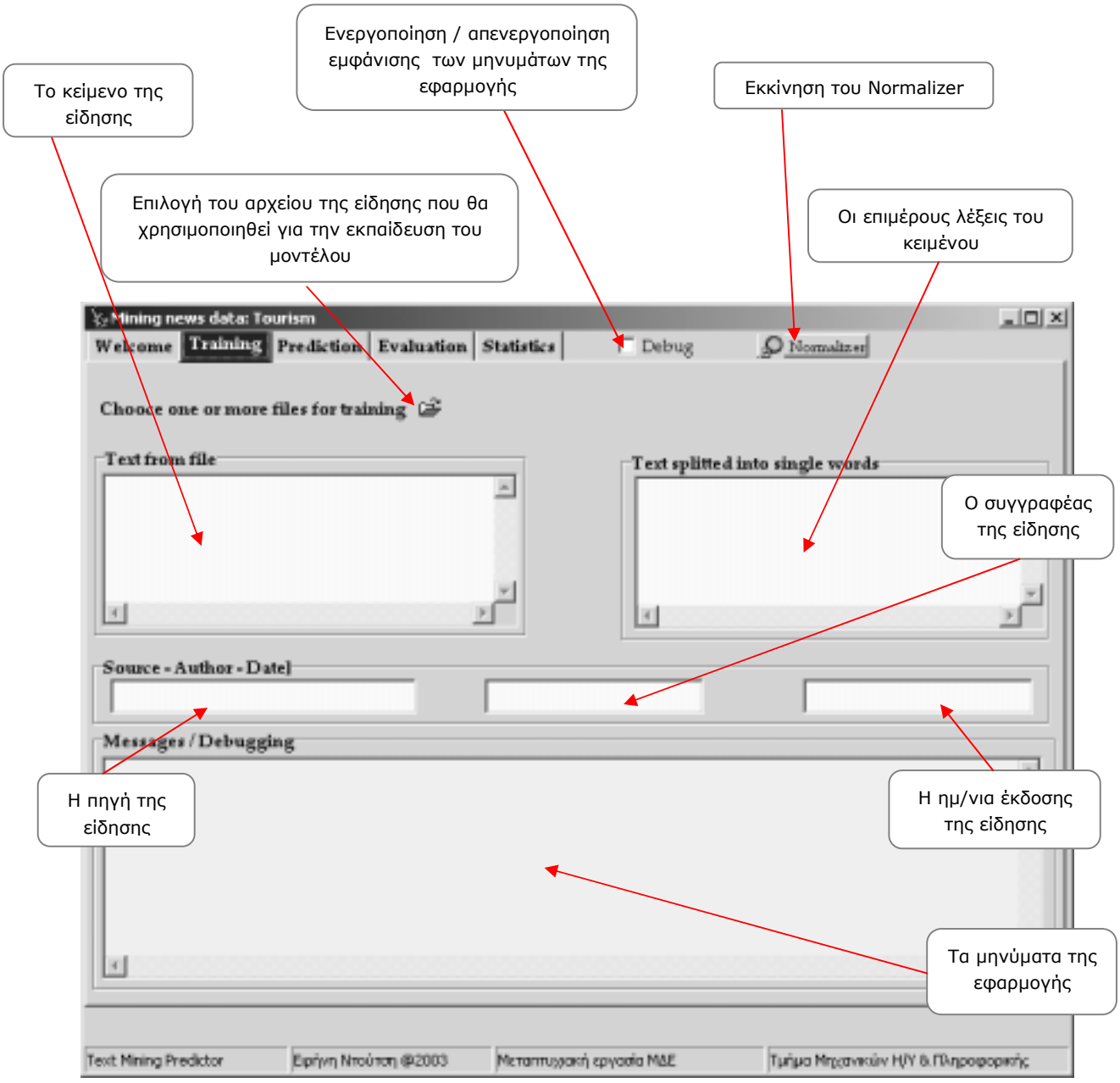
πειραμάτων του χρήστη.

Στη συνέχεια παρουσιάζουμε κάθε επιλογή ξεχωριστά.

5.2 Εκπαίδευση του μοντέλου

Η εκπαίδευση του μοντέλου περιλαμβάνει τη δημιουργία του λεξικού του προς επίλυση προβλήματος το οποίο αποτελείται από τις διακριτές λέξεις των στιγμιότυπων του συνόλου εκπαίδευσης και τη συχνότητα εμφάνισης κάθε λέξης στις διάφορες κλάσεις του προβλήματος. Η εκπαίδευση γίνεται μέσω των στιγμιότυπων του συνόλου εκπαίδευσης. Στη συνέχεια θα δούμε πως πραγματοποιείται η εκπαίδευση μέσω της εφαρμογής.

Από το περιβάλλον έναρξης της εφαρμογής επιλέγουμε την επιλογή "Training". Θα παρουσιαστεί μια οθόνη ανάλογη με αυτή του Σχήματος 5.3.

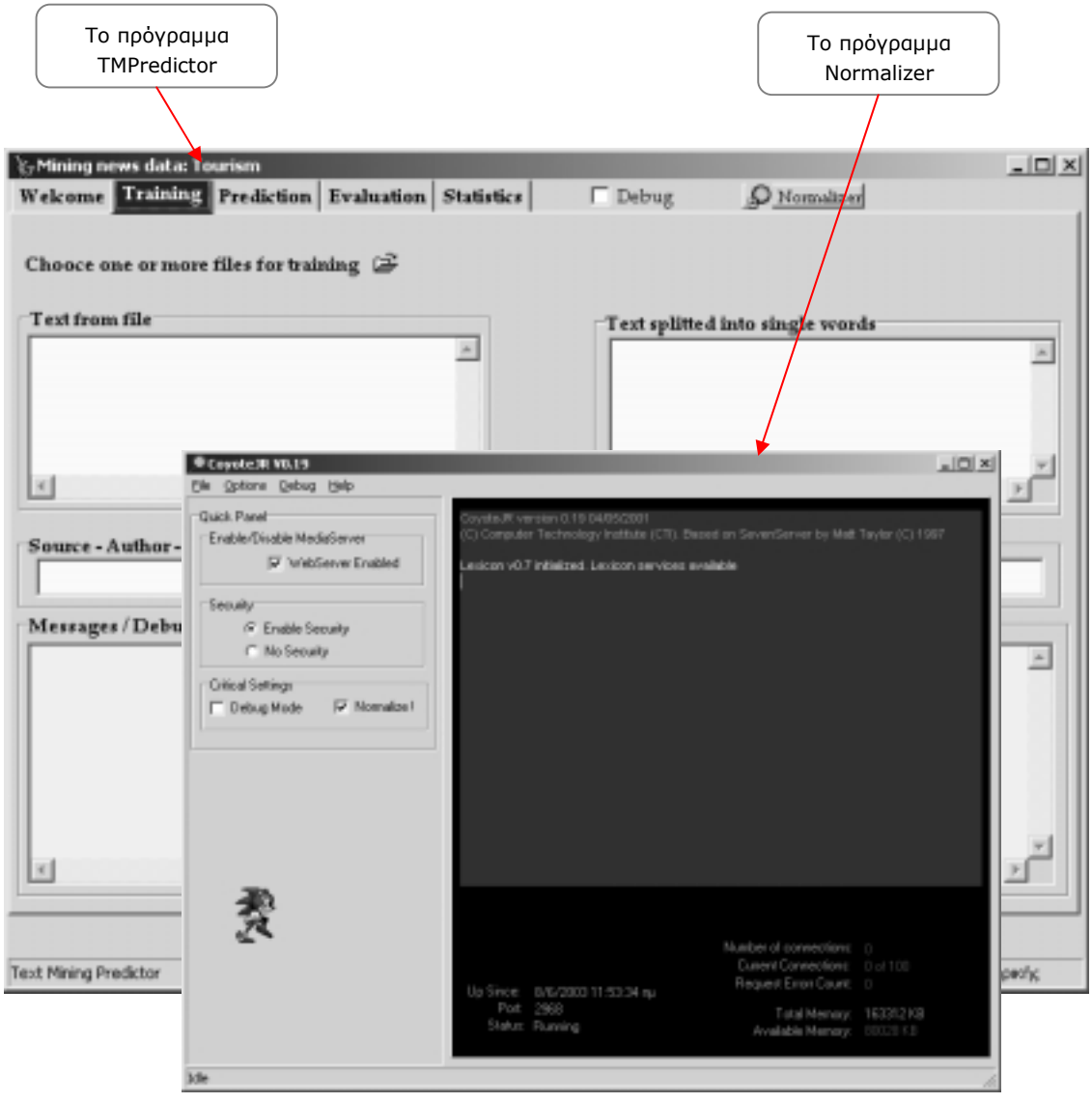


Σχήμα 5.3 Το περιβάλλον εκπαίδευσης του μοντέλου

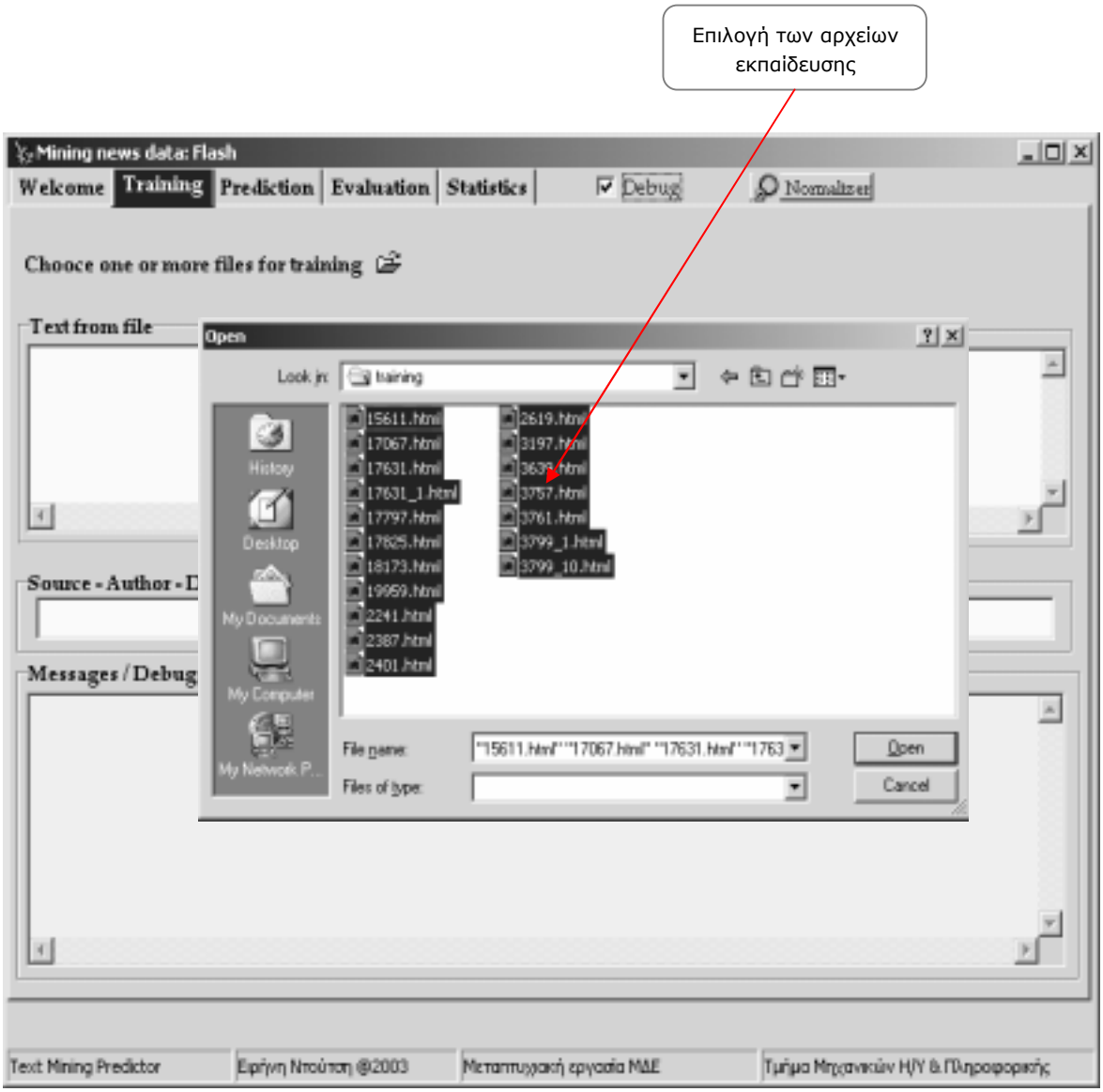
Η επιλογή "Debug" προσφέρει στο χρήστη τη δυνατότητα να βλέπει τη σταδιακή εκτέλεση του προγράμματος μέσα από τα μηνύματα της εφαρμογής.

Τα βήματα που πρέπει να ακολουθήσει ο χρήστης προκειμένου να εκπαιδεύσει το μοντέλο είναι τα ακόλουθα:

1. Εκκίνηση το προγράμματος κανονικοποίησης Normalizer (Σχήμα 5.4) – όπως έχουμε ήδη αναφέρει είναι απαραίτητο να τρέχει ο Normalizer κατά τη διάρκεια της εκπαίδευσης.
2. Επιλογή των αρχείων ειδήσεων που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου (Σχήμα 5.5). Ο χρήστης έχει τη δυνατότητα να επιλέξει ένα ή περισσότερα αρχεία κάθε φορά.
3. Ενεργοποίηση της δυνατότητας εμφάνισης των μηνυμάτων της εφαρμογής (επιλογή Debug) για την παρακολούθηση των λεπτομερειών του τρεξίματος. Το βήμα αυτό είναι προαιρετικό.



Σχήμα 5.4 Εκκίνηση Normalizer.

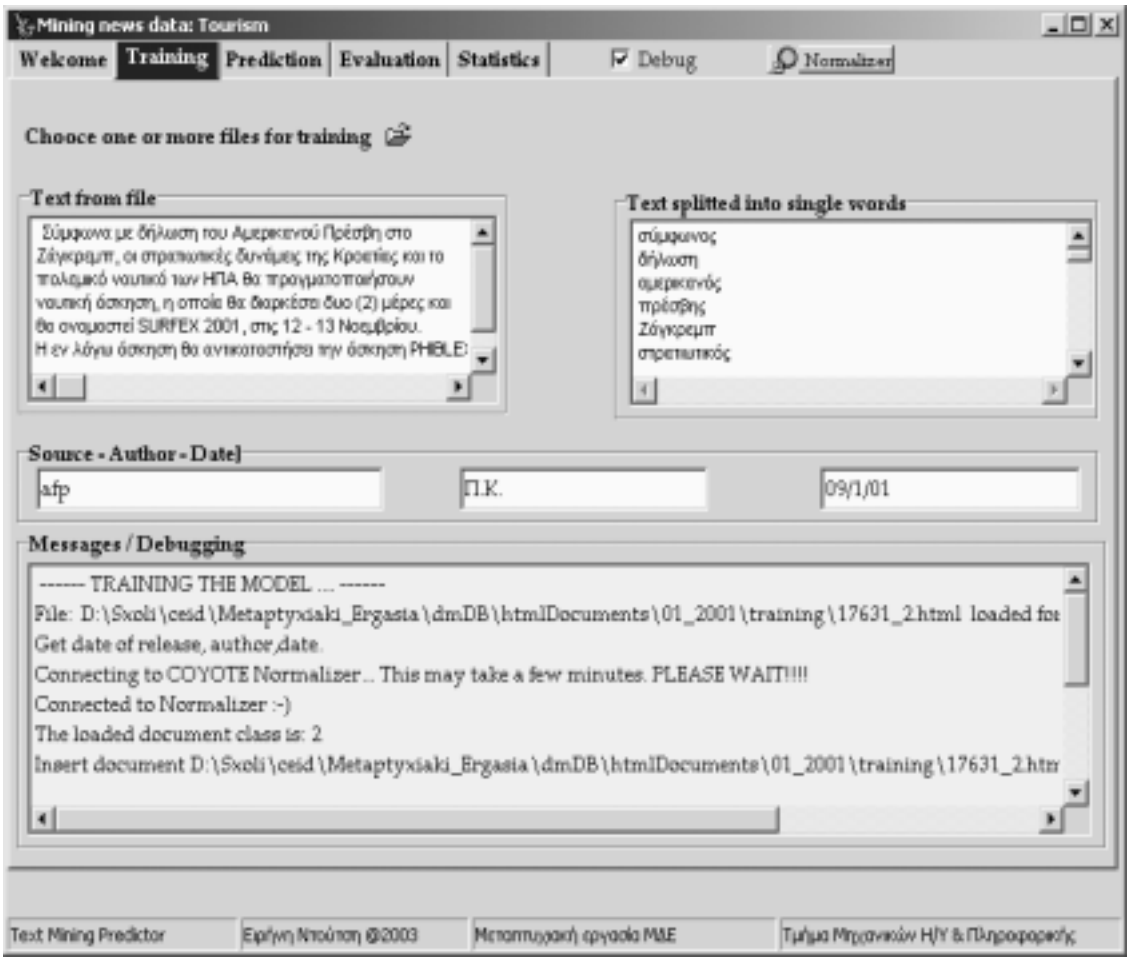


Σχήμα 5.5 Επιλογή των αρχείων των ειδήσεων που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου.

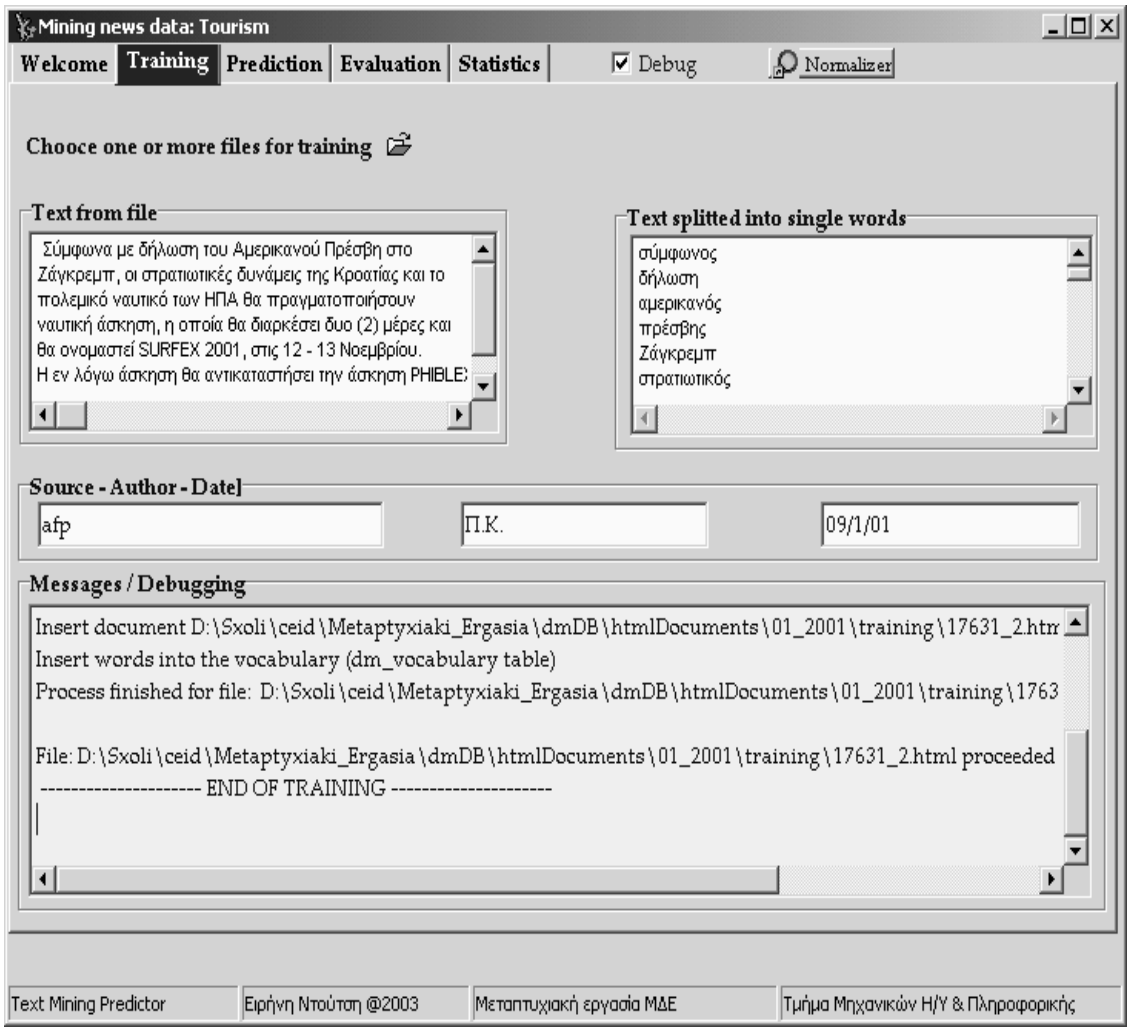
Αφού λοιπόν, επιλεγούν τα αρχεία εκπαίδευσης ξεκινάει η διαδικασία της εκπαίδευσης του μοντέλου. Αν είναι ενεργοποιημένη η επιλογή εμφάνισης των μηνυμάτων της εφαρμογής (Debug) θα εμφανίζονται μηνύματα σχετικά με τη σταδιακή εκτέλεση της εφαρμογής στην περιοχή των μηνυμάτων της εφαρμογής στο κάτω μέρος της οθόνης (Σχήμα 5.6). Σε κάθε περίπτωση, μόλις η εκτέλεση ολοκληρωθεί θα εμφανιστεί ένα μήνυμα τέλους στην περιοχή των μηνυμάτων της εφαρμογής.

Το μήνυμα τέλους μπορεί να είναι είτε μήνυμα επιτυχίας (Σχήμα 5.7), είτε μήνυμα αποτυχίας αν π.χ. συνέβη κάποιο λάθος κατά την εκτέλεση ή αν το

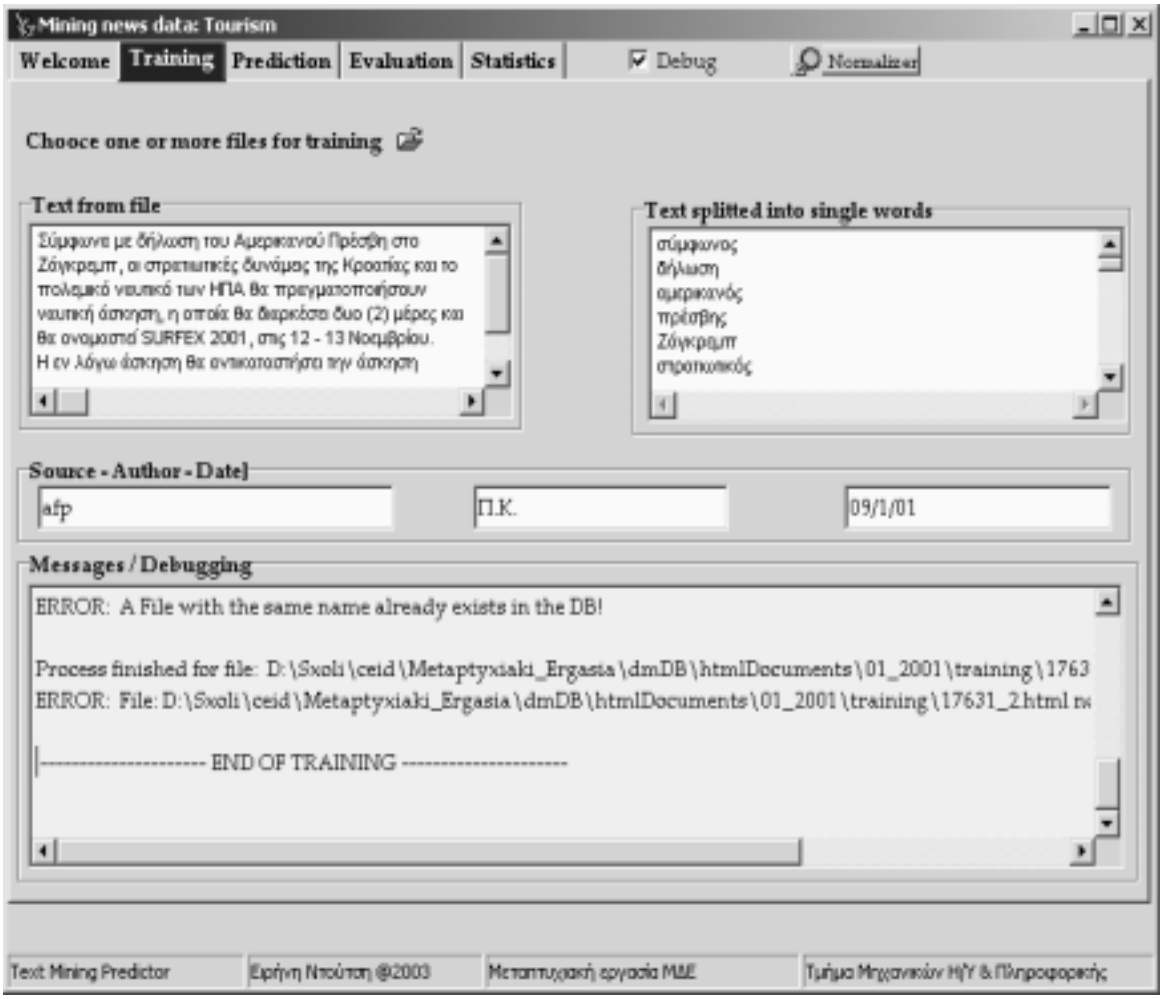
πρόγραμμα δεν μπόρεσε να συνδεθεί με τον Normalizer ή το αρχείο εκπαίδευσης υπήρχε ήδη στη βάση δεδομένων της εφαρμογής κ.λ.π. (Σχήμα 5.8). Σε κάθε περίπτωση το μήνυμα τέλους που εμφανίζεται στην περιοχή των μηνυμάτων της εφαρμογής θα ενημερώνει κατάλληλα το χρήστη.



Σχήμα 5.6 Εκτέλεση της εφαρμογής για την εκπαίδευση του μοντέλου.



Σχήμα 5.7 Επιτυχημένη ολοκλήρωση της εκτέλεσης της εφαρμογής για την εκπαίδευση του μοντέλου.

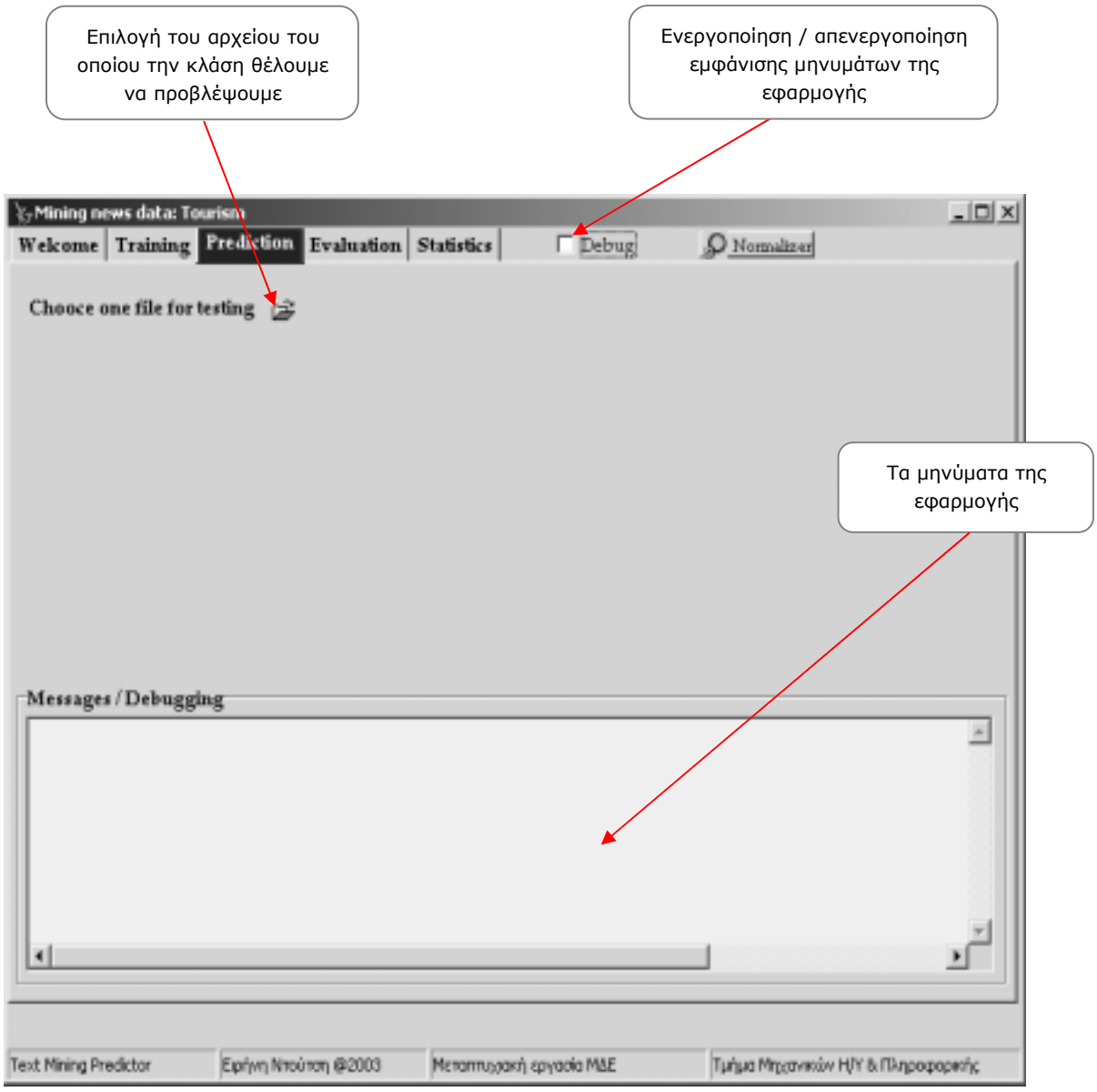


Σχήμα 5.8 Αποτυχημένη ολοκλήρωση της εκτέλεσης της εφαρμογής για την εκπαίδευση του μοντέλου (το αρχείο εκπαίδευσης υπήρχε ήδη στη βάση δεδομένων).

5.3 Πρόβλεψη της κλάσης νέων άγνωστων στιγμιότυπων του προβλήματος

Όπως έχουμε ήδη αναφέρει η πρόβλεψη της κλάσης νέων άγνωστων στιγμιότυπων του προβλήματος γίνεται μέσω των Naïve Bayes ταξινομητών. Η πρόβλεψη αφορά αρχεία ειδήσεων για τα οποία δεν ξέρουμε σε ποια κλάση ανήκουν. Στη συνέχεια θα δούμε πως πραγματοποιείται η πρόβλεψη μέσω της εφαρμογής.

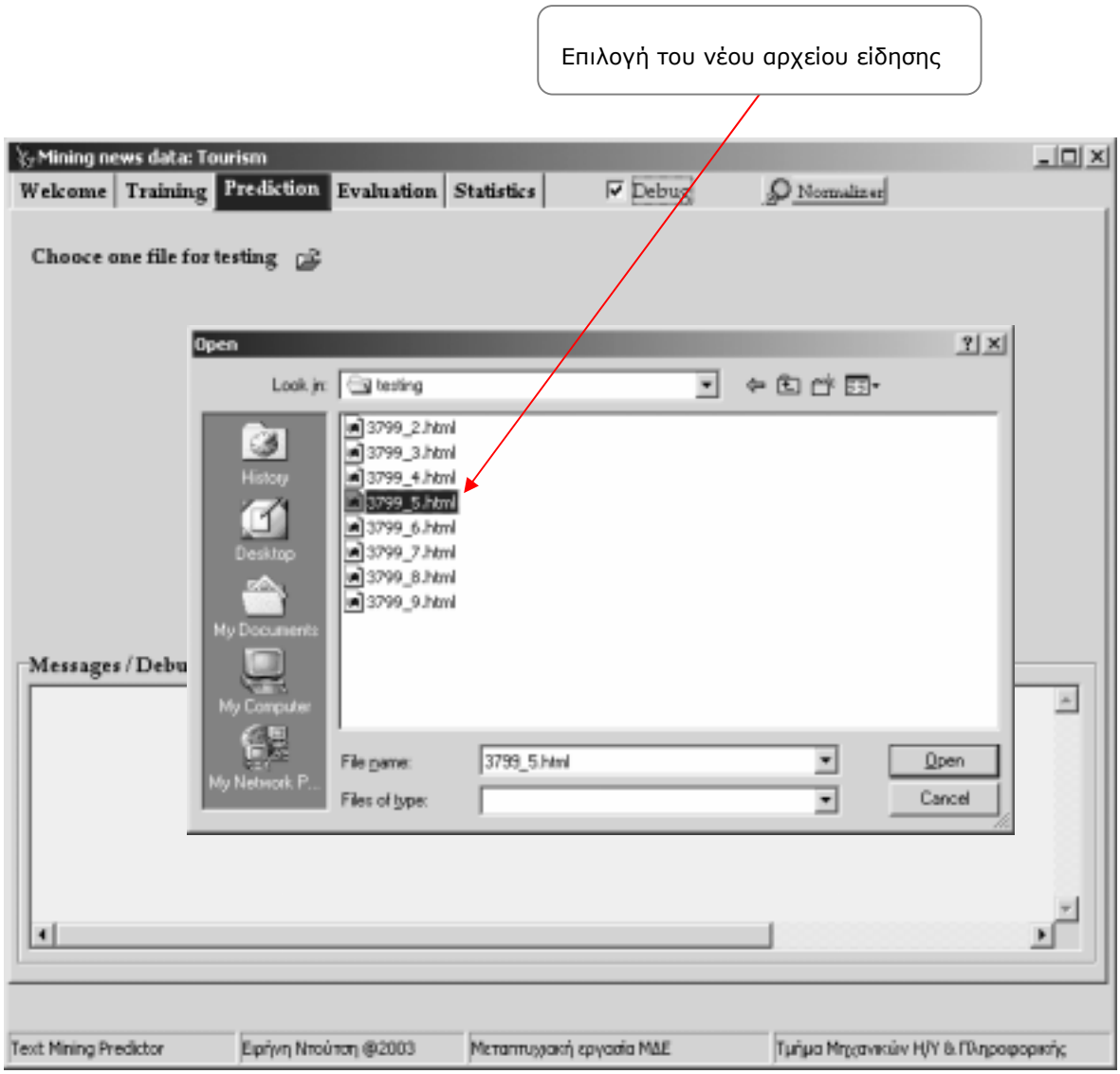
Από το περιβάλλον έναρξης της εφαρμογής επιλέγουμε την επιλογή "Prediction". Θα παρουσιαστεί μια οθόνη ανάλογη με αυτή του Σχήματος 5.9.



Σχήμα 5.9 Το περιβάλλον πρόβλεψης της κλάσης ενός νέου άγνωστου στιγμιότυπου

Τα βήματα που πρέπει να ακολουθήσει ο χρήστης προκειμένου να προβλέψει την κλάση ενός νέου άγνωστου στιγμιότυπου του προβλήματος είναι τα ακόλουθα:

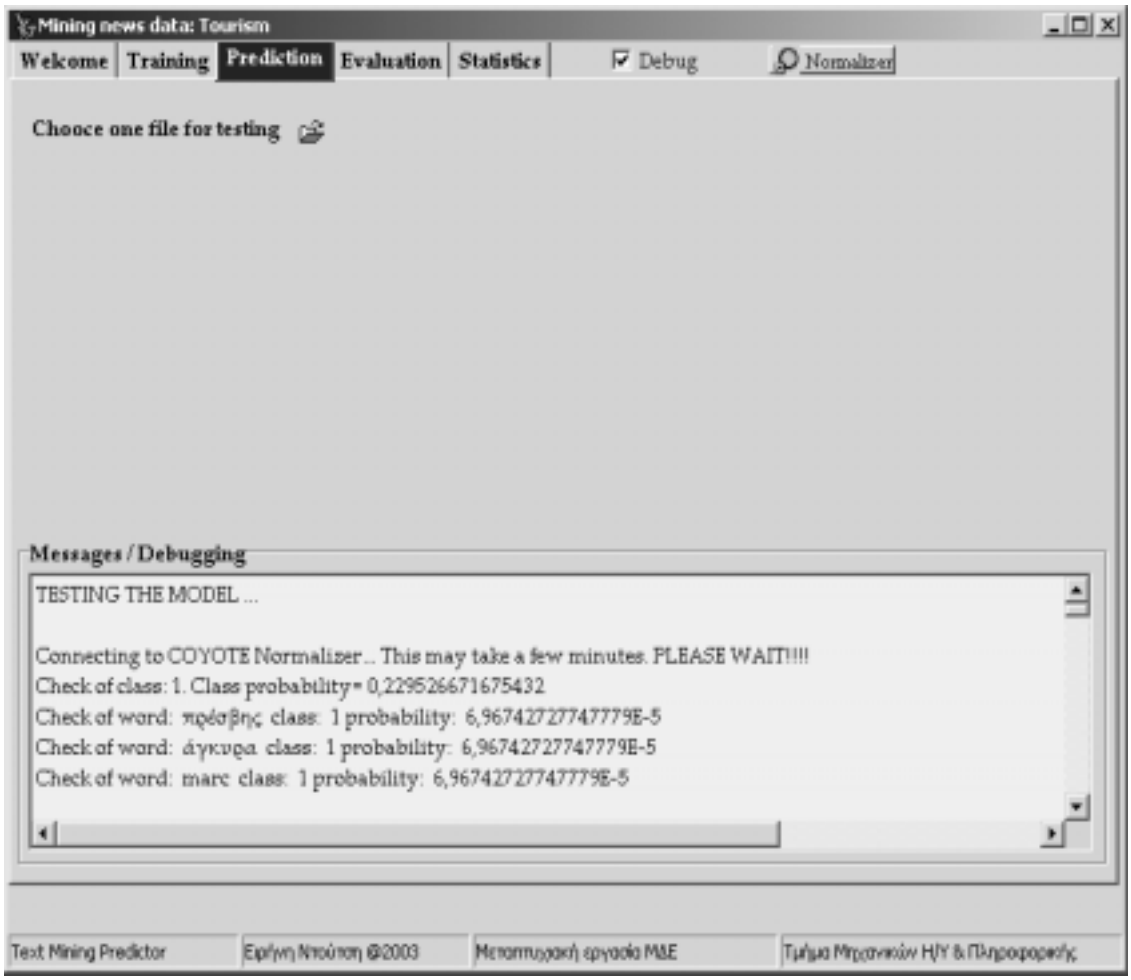
1. Εκκίνηση του προγράμματος κανονικοποίησης Normalizer (Σχήμα 5.4) – όπως έχουμε ήδη αναφέρει είναι απαραίτητο να τρέχει ο Normalizer κατά τη διάρκεια της πρόβλεψης.
2. Επιλογή του νέου αρχείου για το οποίο θέλουμε να προβλέψουμε την κλάση του (Σχήμα 5.10). Κάθε φορά η επιλογή αφορά ένα μόνο αρχείο.
3. Ενεργοποίηση της δυνατότητας εμφάνισης των μηνυμάτων της εφαρμογής (επιλογή Debug) για την παρακολούθηση των λεπτομερειών του τρεξίματος. Το βήμα αυτό είναι προαιρετικό.



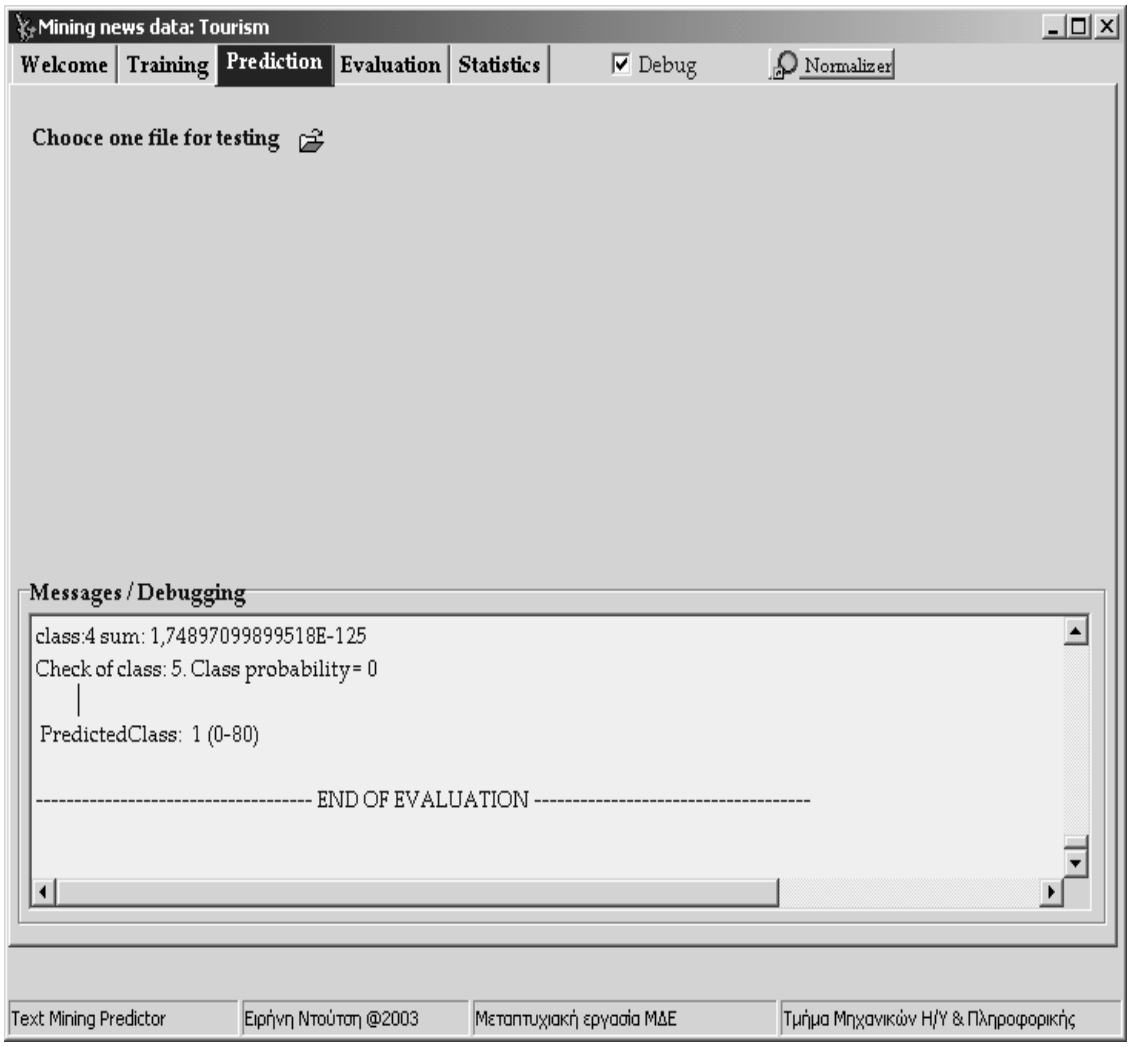
Σχήμα 5.10 Επιλογή του νέου αρχείου είδησης του οποίου την κλάση θέλουμε να προβλέψουμε.

Μετά την επιλογή του νέου άγνωστου αρχείου είδησης ξεκινάει η διαδικασία πρόβλεψης της κλάσης του. Αν είναι ενεργοποιημένη η επιλογή εμφάνισης των μηνυμάτων της εφαρμογής, θα εμφανιστούν στην περιοχή των μηνυμάτων της εφαρμογής μηνύματα που αφορούν τη σταδιακή εκτέλεση της εφαρμογής (Σχήμα 5.11). Μόλις η εκτέλεση ολοκληρωθεί θα εμφανιστεί ένα μήνυμα τέλους στην περιοχή των μηνυμάτων της εφαρμογής που βρίσκεται στο κάτω μέρος της οθόνης.

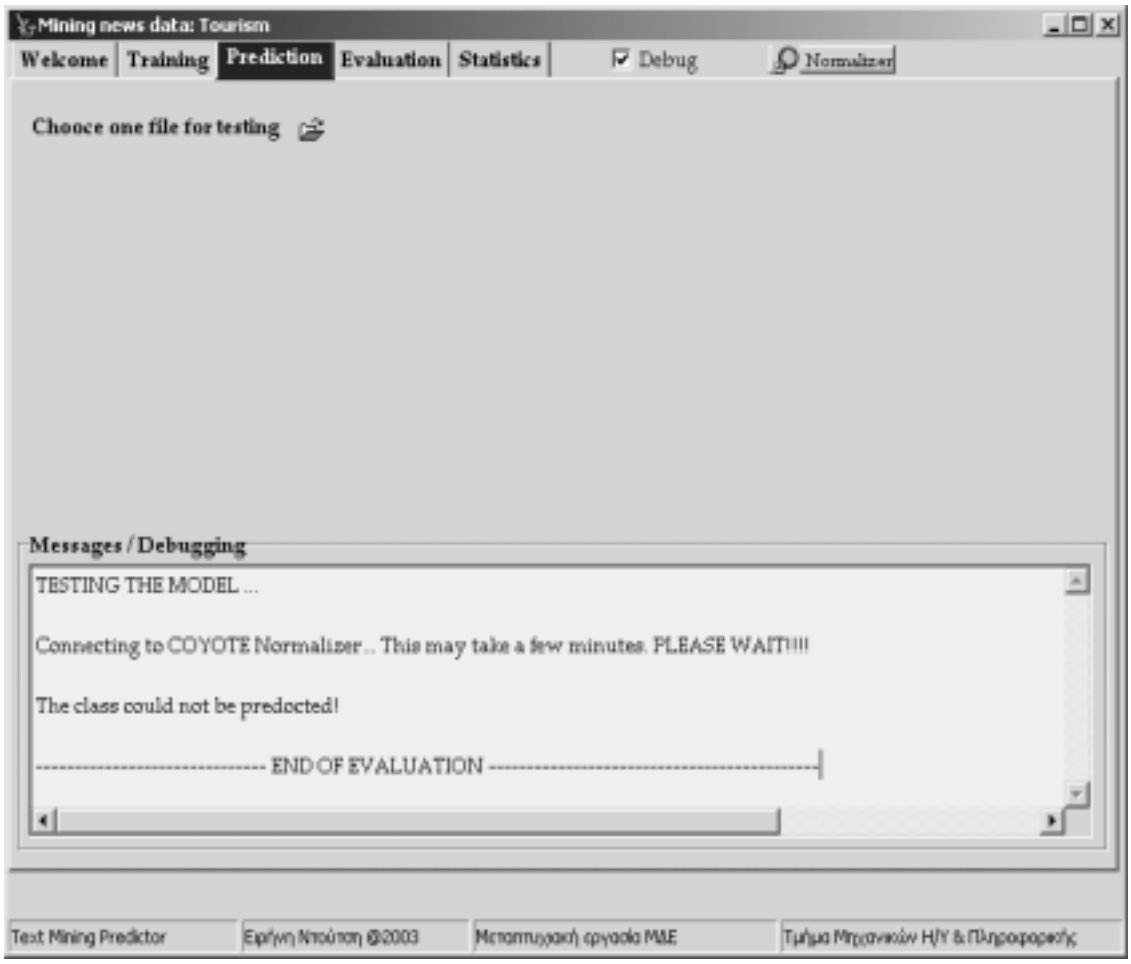
Το μήνυμα τέλους μπορεί να είναι είτε μήνυμα επιτυχίας (Σχήμα 5.12) σε περίπτωση που το σύστημα μπόρεσε να προβλέψει μία από τις κλάσεις του προβλήματος, είτε μήνυμα αποτυχίας (Σχήμα 5.13) σε περίπτωση που το σύστημα δεν μπόρεσε να προβλέψει καμία από τις κλάσεις του προβλήματος.



Σχήμα 5.11 Εκτέλεση της εφαρμογής για την πρόβλεψη της κλάσης ενός νέου άγνωστου στιγμιότυπου του προβλήματος.



Σχήμα 5.12 Επιτυχημένη ολοκλήρωση της εκτέλεσης της εφαρμογής για την πρόβλεψη της κλάσης ενός νέου στιγμιότυπου

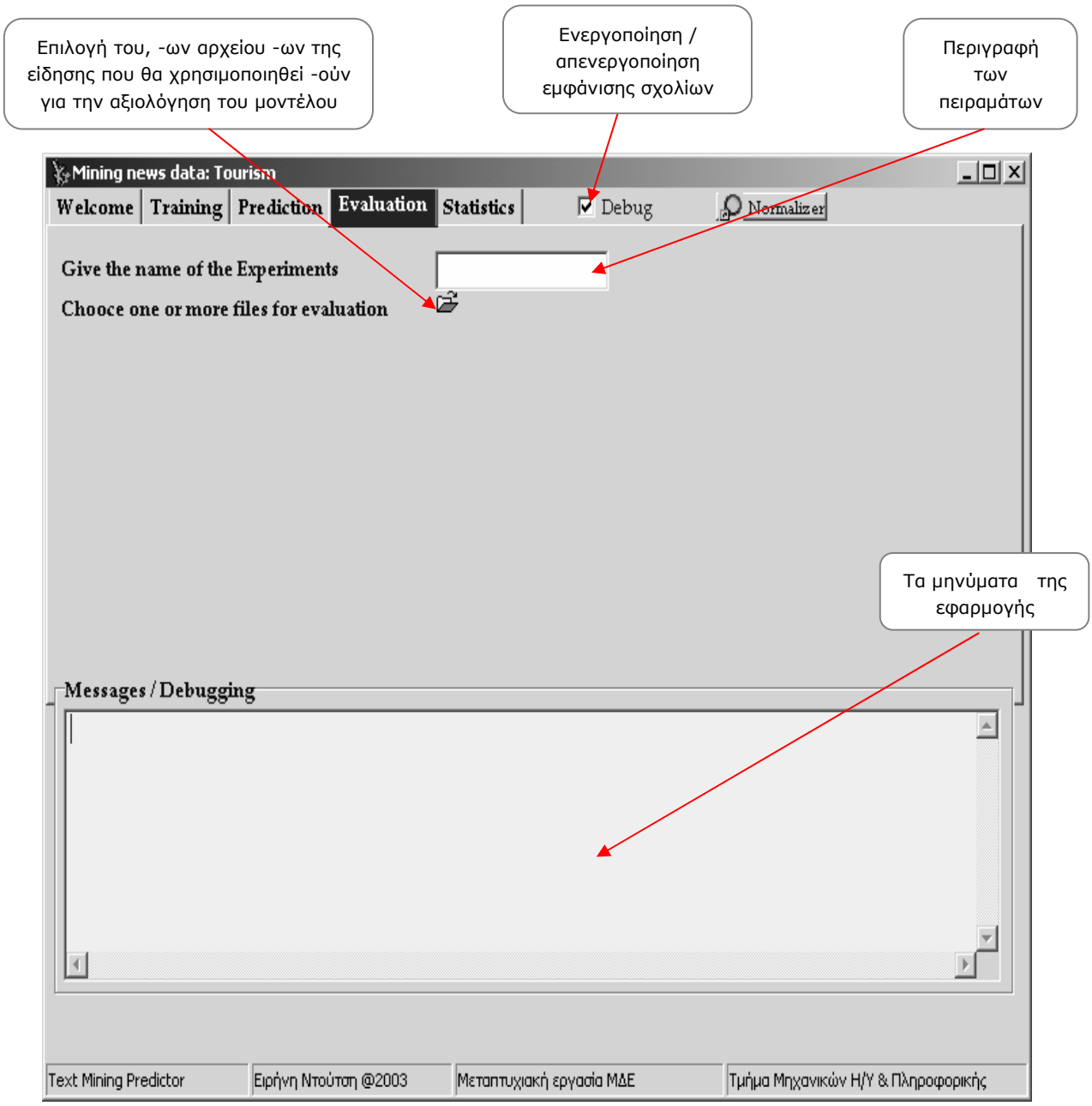


Σχήμα 5.13 Αποτυχημένη ολοκλήρωση της εκτέλεσης της εφαρμογής για την πρόβλεψη της κλάσης ενός νέου στιγμιότυπου (η κλάση δεν μπόρεσε να προβλεφθεί)

5.4 Αξιολόγηση του μοντέλου

Η αξιολόγηση του μοντέλου αφορά την αξιοπιστία με την οποία το μοντέλο πρόβλεψης προβλέπει την κλάση νέων άγνωστων στιγμιότυπων του προβλήματος. Η αξιολόγηση στηρίζεται στα στιγμιότυπα του συνόλου ελέγχου του προβλήματος και ισούται με το ποσοστό των σωστά προβλεπόμενων στιγμιότυπων του προβλήματος στο σύνολο των προς πρόβλεψη στιγμιότυπων του προβλήματος.

Από το περιβάλλον έναρξης της εφαρμογής επιλέγουμε την επιλογή "Evaluation". Θα παρουσιαστεί μια οθόνη ανάλογη με αυτή του Σχήματος 5.15.

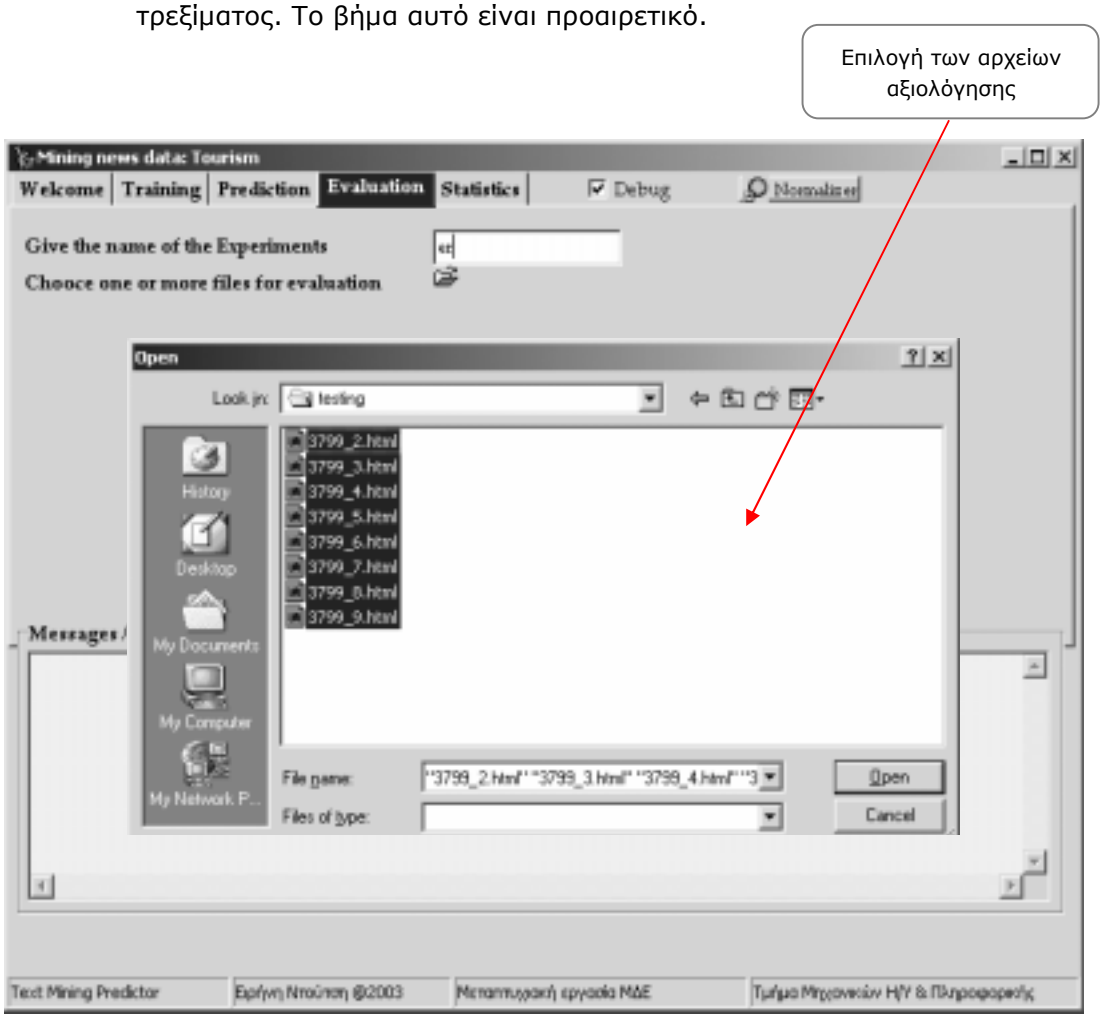


Σχήμα 5.14 Το περιβάλλον αξιολόγησης του μοντέλου

Τα βήματα που πρέπει να ακολουθήσει ο χρήστης προκειμένου να αξιολογήσει την απόδοση του μοντέλου είναι τα ακόλουθα:

1. Εκκίνηση του προγράμματος κανονικοποίησης Normalizer (Σχήμα 5.4) – όπως έχουμε ήδη αναφέρει είναι απαραίτητο να τρέχει ο Normalizer κατά τη διάρκεια της αξιολόγησης.

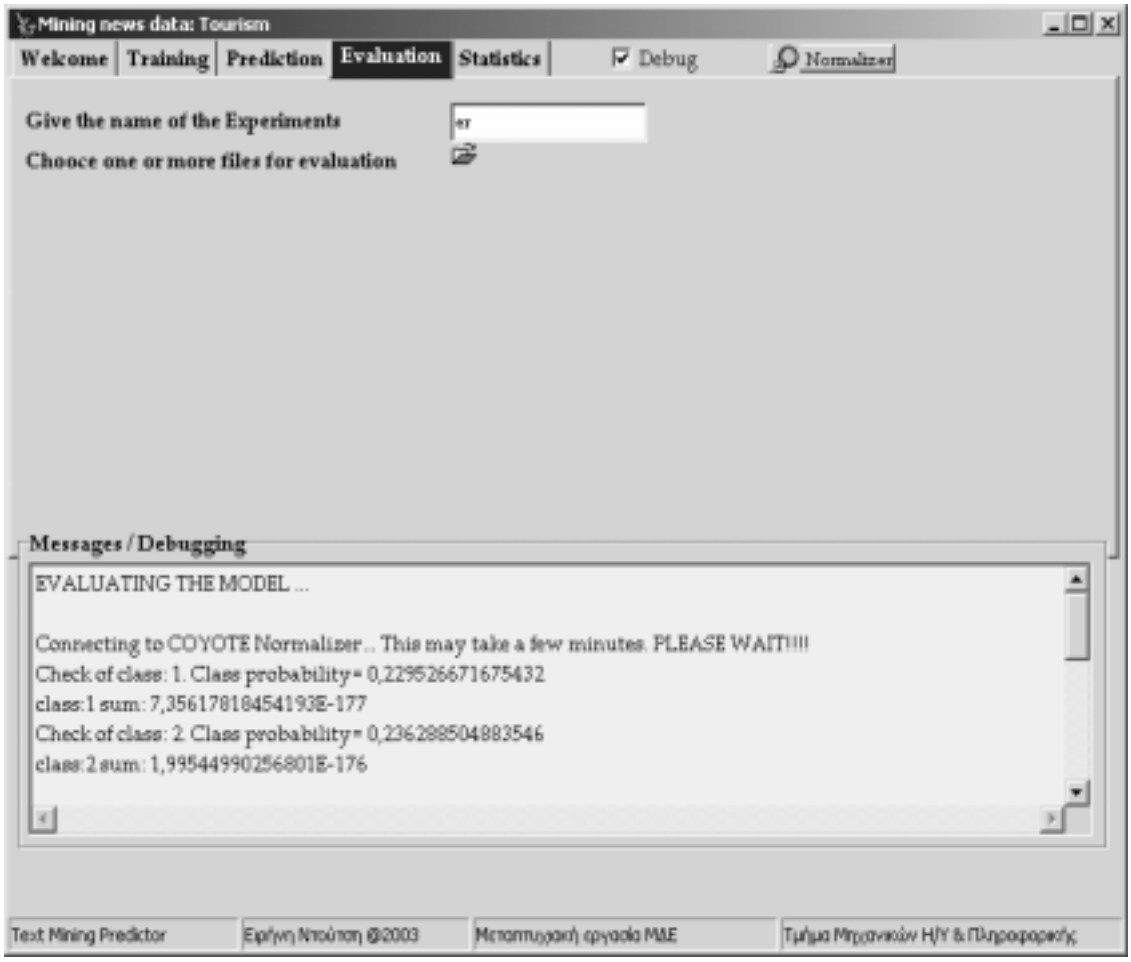
2. Λεκτική περιγραφή των πειραμάτων (συμπλήρωση του αντίστοιχου πεδίου της εφαρμογής) προκειμένου να είναι εύκολη η ανάκτησή τους στη συνέχεια.
3. Επιλογή των αρχείων ειδήσεων που θα χρησιμοποιηθούν την αξιολόγηση (Σχήμα 5.15). Η επιλογή αφορά κάθε φορά ένα ή περισσότερα αρχεία.
4. Ενεργοποίηση της δυνατότητας εμφάνισης των μηνυμάτων της εφαρμογής (επιλογή Debug) για την παρακολούθηση των λεπτομερειών του τρεξίματος. Το βήμα αυτό είναι προαιρετικό.



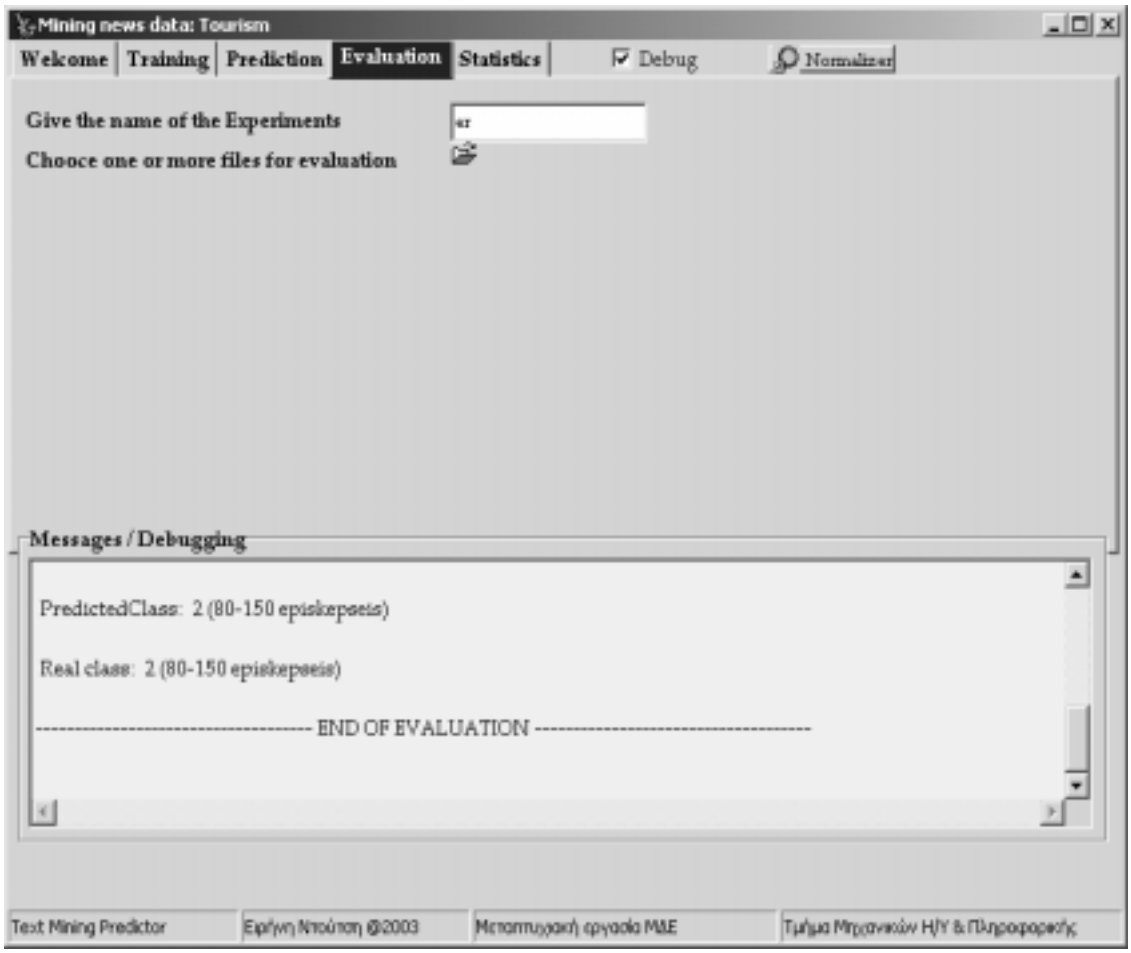
Σχήμα 5.15 Επιλογή των αρχείων ειδήσεων που θα χρησιμοποιηθούν για την αξιολόγηση του μοντέλου.

Μόλις λοιπόν, επιλέξουμε τα αρχεία αξιολόγησης ξεκινάει η διαδικασία της αξιολόγησης του μοντέλου. Αν η επιλογή εμφάνισης των μηνυμάτων της εφαρμογής είναι ενεργοποιημένη, ο χρήστης μπορεί να βλέπει τη σταδιακή εκτέλεση της εφαρμογής (Σχήμα 5.16). Μόλις η εκτέλεση ολοκληρωθεί θα εμφανιστεί ένα μήνυμα τέλους στην περιοχή των μηνυμάτων της εφαρμογής (κάτω μέρος της οθόνης) το οποίο ενημερώνει το χρήστη για την πραγματική και την προβλεπόμενη κλάση κάθε αρχείου αξιολόγησης που επέλεξε.

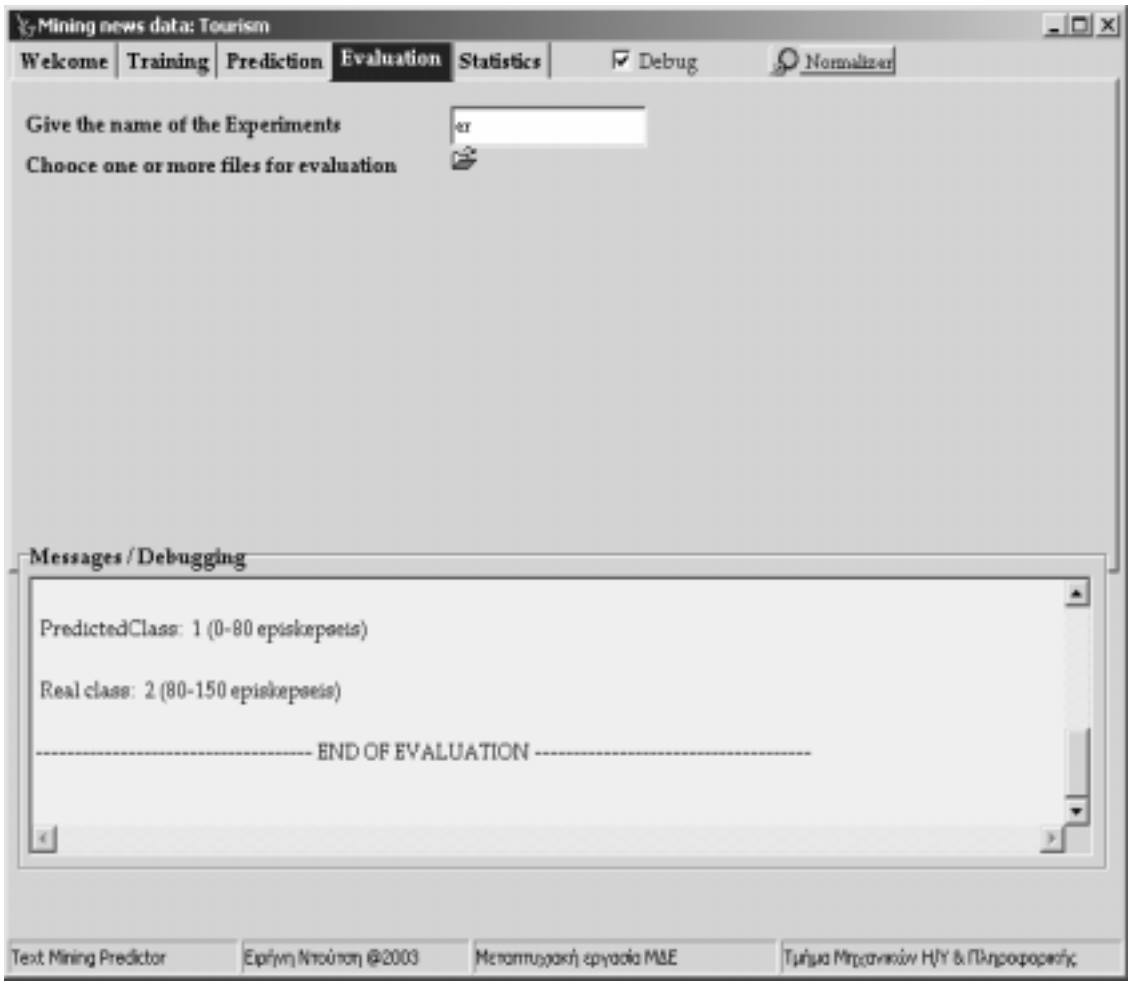
Υπάρχουν δύο πιθανές περιπτώσεις για κάθε αρχείο αξιολόγησης. Η πρώτη περίπτωση είναι να έχει γίνει σωστή πρόβλεψη, δηλαδή η προβλεπόμενη κλάση να συμφωνεί με την πραγματική κλάση του στιγμιότυπου (Σχήμα 5.17). Η δεύτερη περίπτωση είναι να έχει γίνει λανθασμένη πρόβλεψη, δηλαδή η προβλεπόμενη κλάση να μην συμφωνεί με την πραγματική κλάση του στιγμιότυπου (Σχήμα 5.18).



Σχήμα 5.16 Εκτέλεση της εφαρμογής για την αξιολόγηση του μοντέλου.



Σχήμα 5.17 Η προβλεπόμενη κλάση συμφωνεί με την πραγματική κλάση του στιγμιότυπου του προβλήματος (περίπτωση ορθής πρόβλεψης)

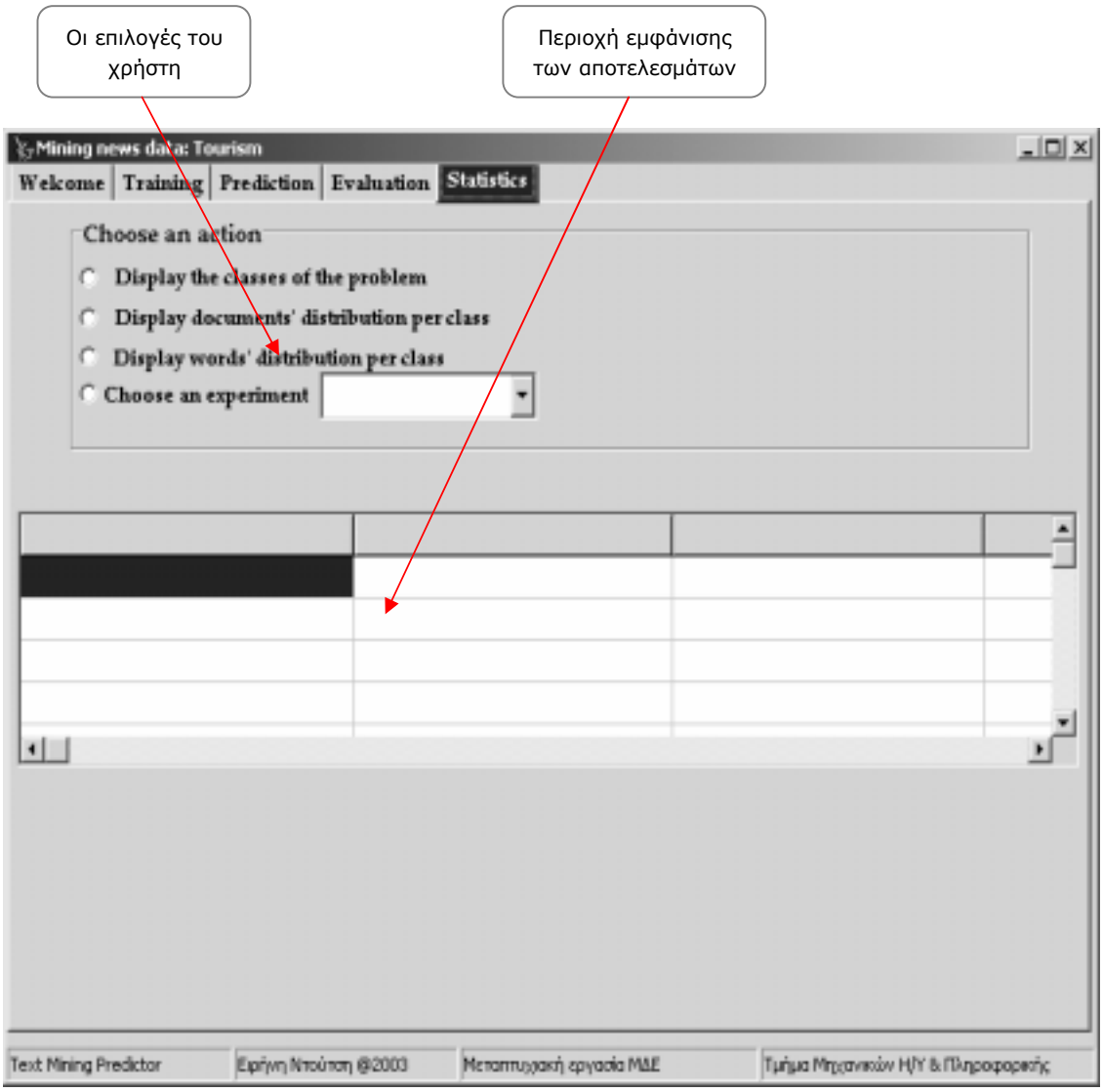


Σχήμα 5.18 Η προβλεπόμενη κλάση δεν συμφωνεί με την πραγματική κλάση του στιγμιότυπου του προβλήματος (περίπτωση λανθασμένης πρόβλεψης)

5.5 Στατιστικά

Το τμήμα αυτό αφορά την παρουσίαση διαφόρων στατιστικών πληροφοριών σχετικά με το μοντέλο και τα πειράματα που τρέξαμε.

Από το περιβάλλον έναρξης της εφαρμογής επιλέγουμε την επιλογή "Statistics". Θα παρουσιαστεί μια οθόνη ανάλογη με αυτή του Σχήματος 5.19.

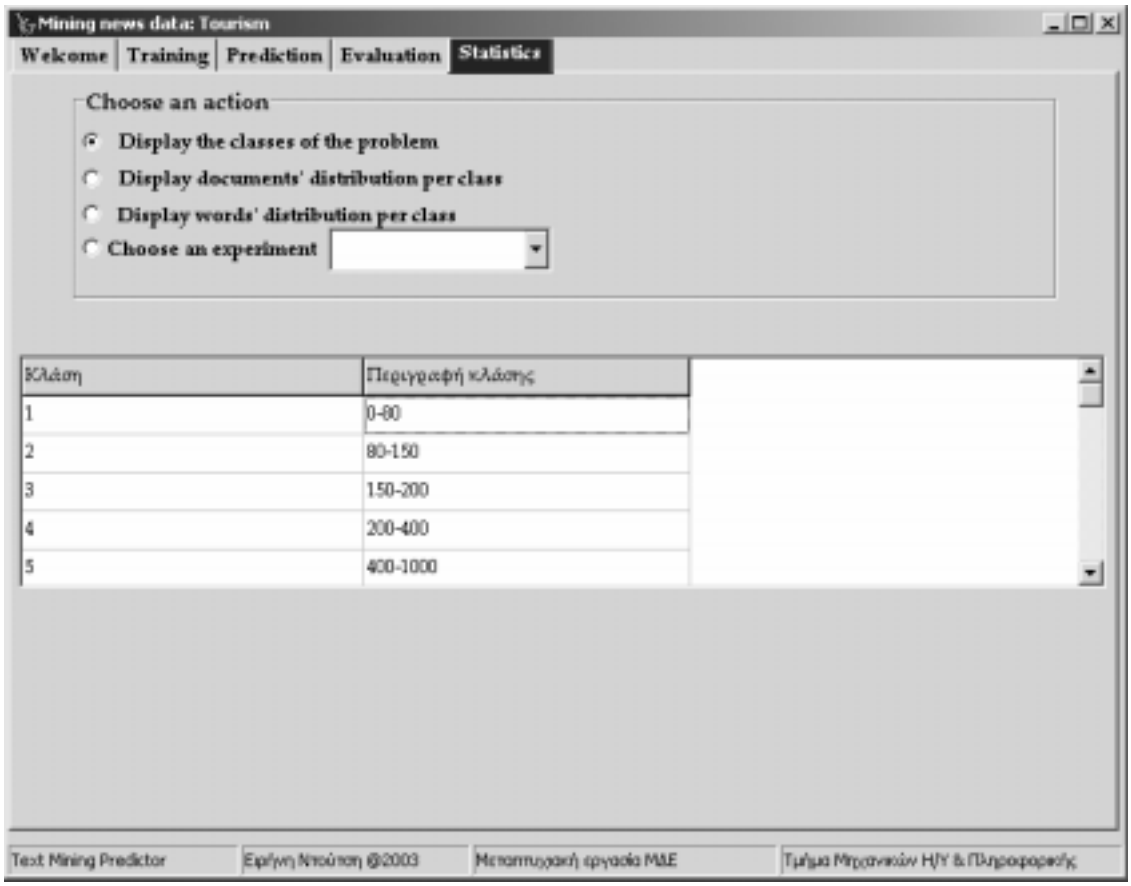


Σχήμα 5.19 Το περιβάλλον αξιολόγησης του μοντέλου

Για τη συνέχεια οι επιλογές του χρήστη είναι οι ακόλουθες:

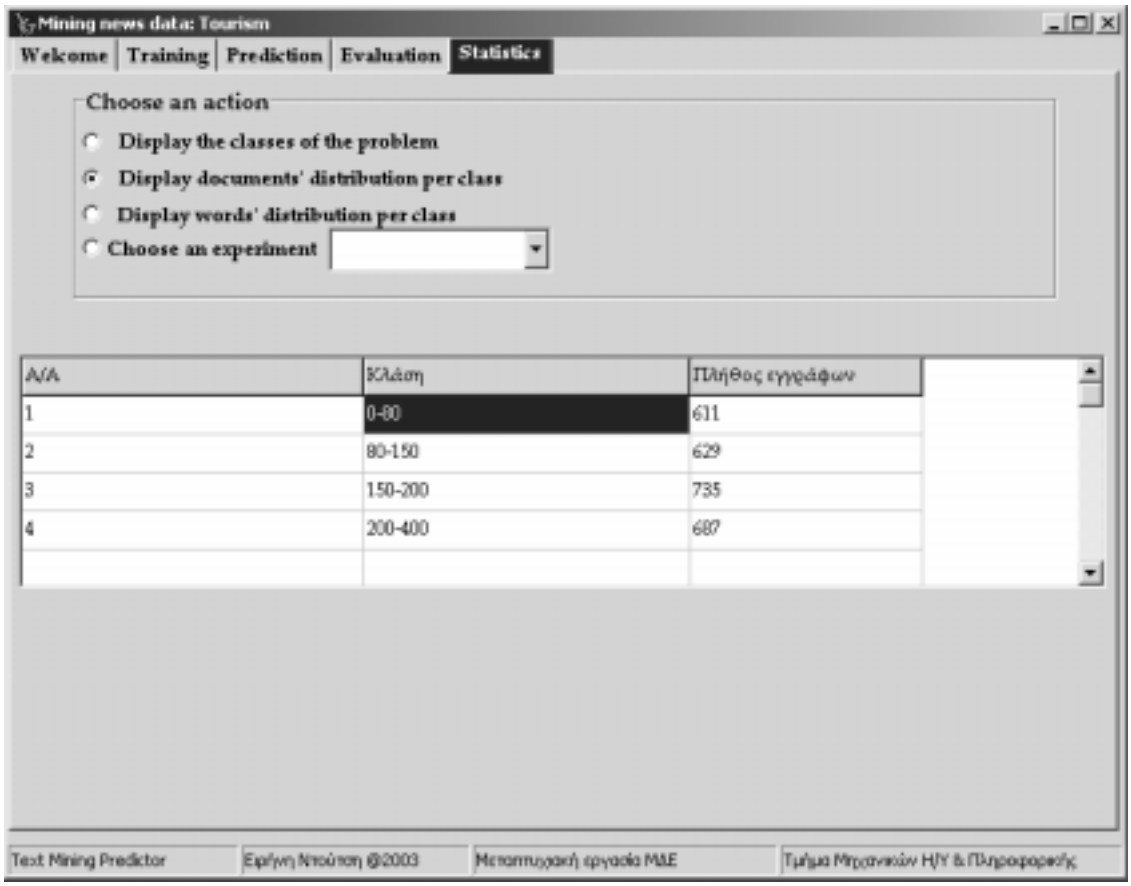
- **Επιλογή "Display the classes of the problem":** μέσω της επιλογής αυτής ο χρήστης μπορεί να ενημερωθεί για τις διάφορες κλάσεις του προβλήματος (τον κωδικό και την περιγραφή τους). Για παράδειγμα, οι κλάσεις του προβλήματος

της τουριστικής κίνησης φαίνονται στο ακόλουθο σχήμα (Σχήμα 5.15)



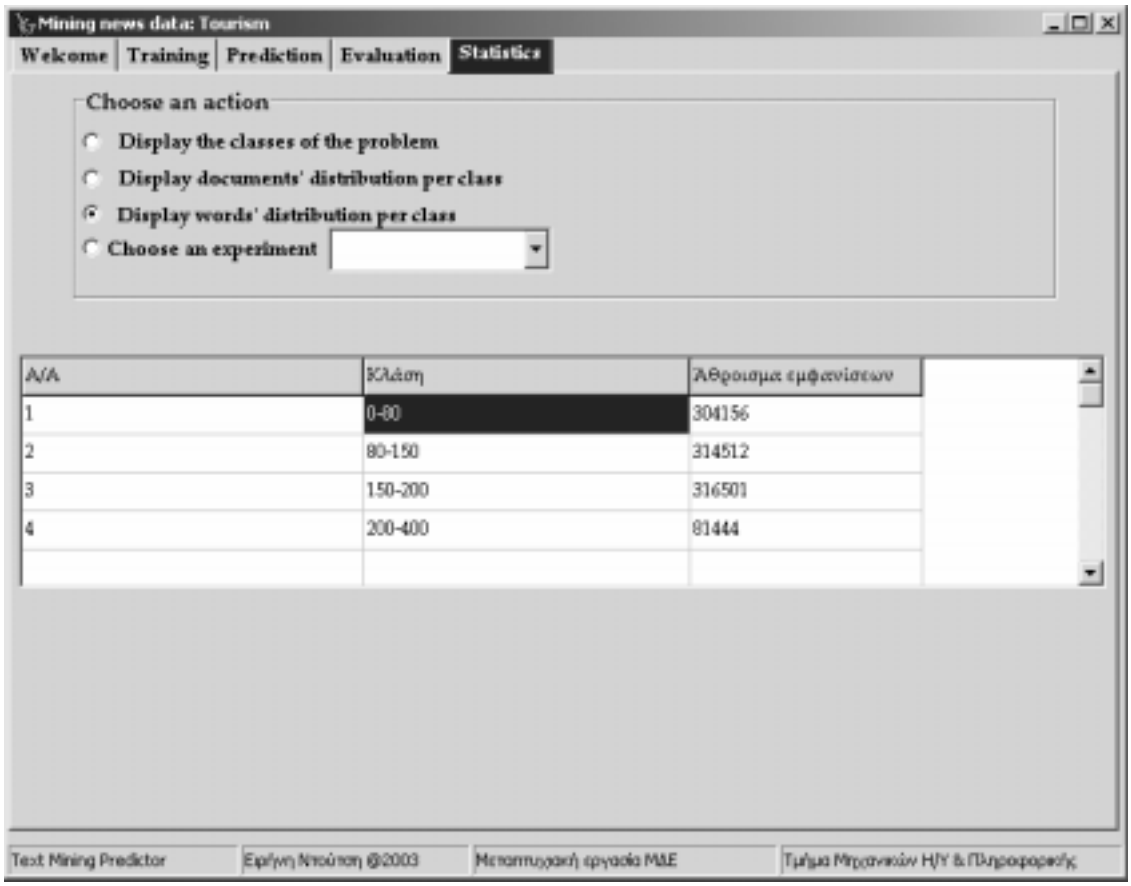
Σχήμα 5.20 Εμφάνιση των κλάσεων του προβλήματος της τουριστικής κίνησης (η δεύτερη στήλη αναφέρεται σε χιλιάδες επισκέψεων).

- **Επιλογή "Display documents distribution per class":** μέσω της επιλογής αυτής ο χρήστης μπορεί να ενημερωθεί για την κατανομή των αρχείων εκπαίδευσης στις διάφορες κλάσεις του προβλήματος (δηλαδή πόσα αρχεία υπάρχουν σε κάθε κλάση). Για παράδειγμα, η κατανομή των αρχείων εκπαίδευσης για το πρόβλημα της τουριστικής κίνησης φαίνεται στο ακόλουθο σχήμα (Σχήμα 5.16)



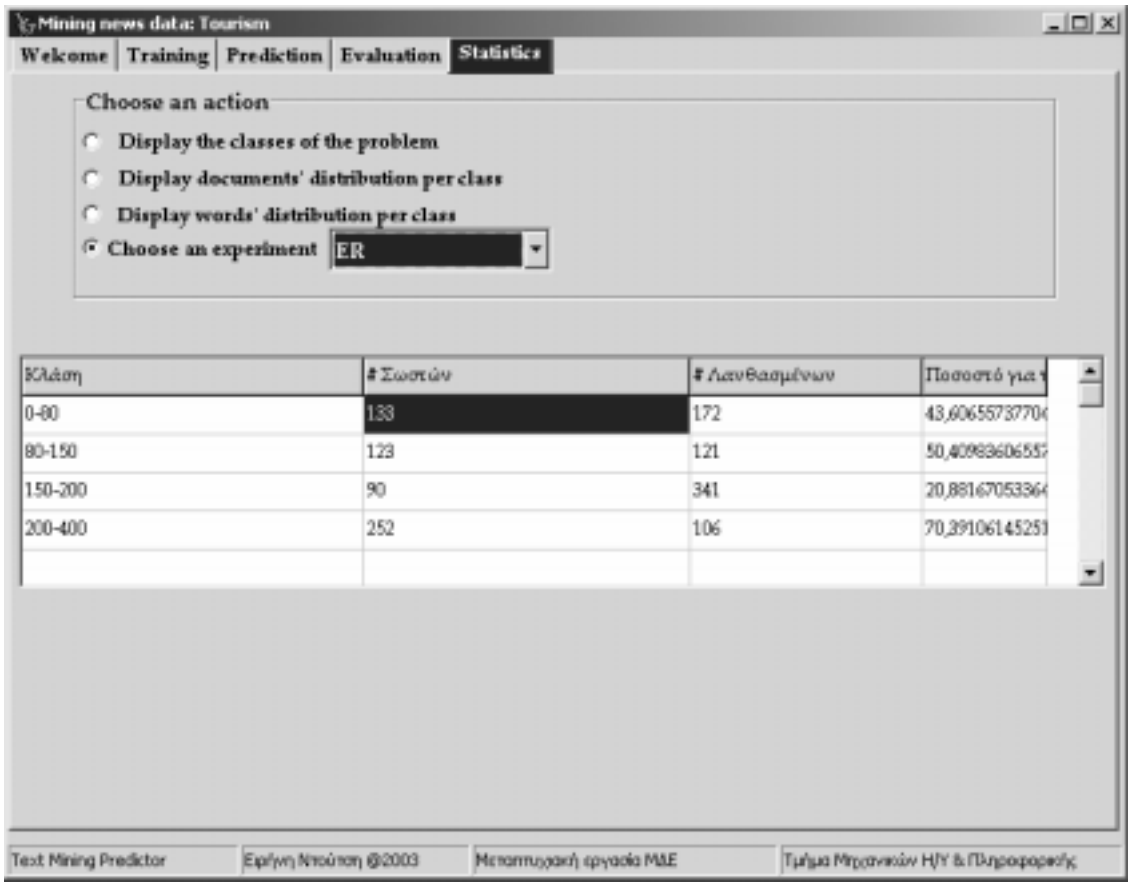
Σχήμα 5.21 Η κατανομή των αρχείων εκπαίδευσης στις διάφορες κλάσεις του προβλήματος της τουριστικής κίνησης (η δεύτερη στήλη αναφέρεται σε χιλιάδες επισκέψεων).

- **Επιλογή "Display words distribution per class":** μέσω της επιλογής αυτής ο χρήστης μπορεί να ενημερωθεί για την κατανομή των σημαντικών λέξεων των στιγμιότυπων εκπαίδευσης του προβλήματος στις διάφορες κλάσεις του προβλήματος (δηλαδή πόσες λέξεις υπάρχουν σε κάθε κλάση). Για παράδειγμα, η κατανομή των λέξεων των στιγμιότυπων εκπαίδευσης για το πρόβλημα της τουριστικής κίνησης φαίνεται στο ακόλουθο σχήμα (Σχήμα 5.17)

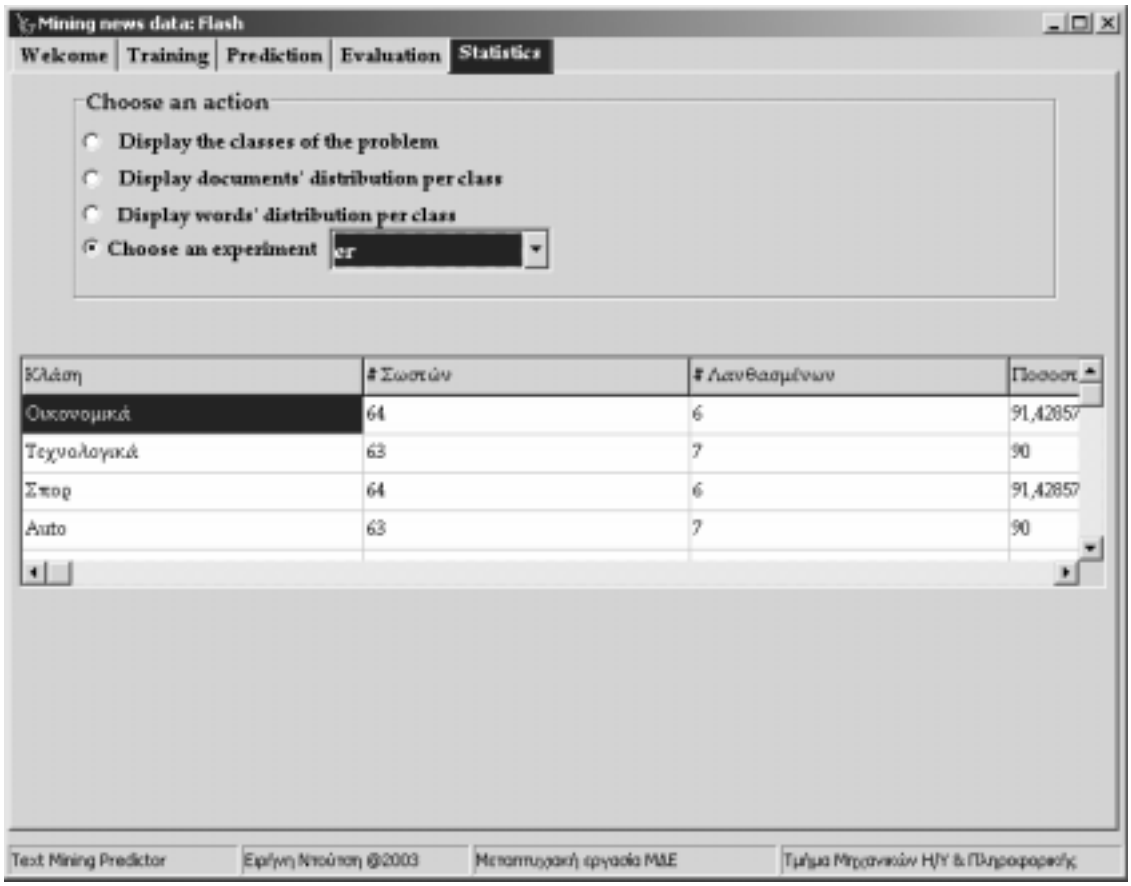


Σχήμα 5.22 Η κατανομή των λέξεων των στιγμιότυπων εκπαίδευσης στις διάφορες κλάσεις του προβλήματος της τουριστικής κίνησης (η δεύτερη στήλη αναφέρεται σε χιλιάδες επισκέψεων).

- **Επιλογή "Choose an experiment":** μέσω της επιλογής αυτής ο χρήστης μπορεί να ενημερωθεί για τα αποτελέσματα των πειραμάτων του (επιλέγοντας κάθε φορά το πείραμα που τον ενδιαφέρει). Τα αποτελέσματα περιλαμβάνουν το πλήθος των σωστά και λάθος ταξινομημένων αρχείων για κάθε κλάση του προβλήματος. Για παράδειγμα, τα αποτελέσματα ενός πειράματος για το πρόβλημα της τουριστικής κίνησης φαίνονται στο Σχήμα 5.23, ενώ τα αποτελέσματα για το πρόβλημα της ταξινόμησης των ειδήσεων του δικτυακού τόπου Flash φαίνονται στο Σχήμα 5.24.



Σχήμα 5.23 Τα αποτελέσματα του πειράματος "ER" (η λεκτική περιγραφή που δόθηκε από το χρήστη) για το πρόβλημα της τουριστικής κίνησης.



Σχήμα 5.24 Τα αποτελέσματα του πειράματος "er" για το πρόβλημα της ταξινόμησης των ειδήσεων του δικτυακού τόπου Flash.

Επίλογος

Η εξόρυξη γνώσης στις βάσεις δεδομένων (*KDD)* αποτελεί ένα ταχέως αναπτυσσόμενο επιστημονικό πεδίο που αποσκοπεί στην ανακάλυψη κρίσιμων "κρυμμένων" πληροφοριών στις βάσεις δεδομένων. Οι εφαρμογές της αυξάνονται συνεχώς λόγω της έκρηξης της πληροφορίας που παρατηρείται τα τελευταία χρόνια και της ανάγκης για εξόρυξη χρήσιμων πληροφοριών από την πληθώρα των διαθέσιμων δεδομένων. Ένας σημαντικός κλάδος της εξόρυξης γνώσης είναι η εξόρυξη γνώσης από κείμενα που αφορά ημι-δομημένα ή αδόμητα δεδομένα όπως για παράδειγμα κείμενα, ιστοσελίδες κ.α.

Ο πρώτος στόχος της παρούσας μεταπτυχιακής εργασίας ήταν η διερεύνηση της έννοιας της εξόρυξης γνώσης και της εξόρυξης γνώσης από κείμενα. Από τη μελέτη της σχετικής βιβλιογραφίας διαφαίνεται καθαρά η αξία της εξόρυξης γνώσης και η επίδραση της σε πολλούς σημαντικούς τομείς της ανθρώπινης δραστηριότητας. Φαίνεται μάλιστα πως καθώς ο όγκος των δεδομένων θα αυξάνεται, λόγω της κυριαρχίας της τεχνολογίας σε όλους τους τομείς της σύγχρονης ζωής, η ανάγκη εξόρυξης γνώσης θα γίνεται πιο έντονη και ουσιαστική.

Ο δεύτερος στόχος μας ήταν η σχεδίαση και η υλοποίηση ενός συστήματος εξόρυξης γνώσης από κείμενα καθώς επίσης και η εφαρμογή και η αξιολόγησή του σε περιπτώσεις πραγματικών προβλημάτων. Ένα πρωτότυπο σύστημα, το TMPredictor, υλοποιήθηκε και αξιολογήθηκε η απόδοσή του για το πρόβλημα της πρόβλεψης της τουριστικής κίνησης στην Ελλάδα από τουρίστες του εξωτερικού και για το πρόβλημα της ταξινόμησης των ειδήσεων του δικτυακού τόπου Flash.

Τα αποτελέσματα ήταν πολύ ενθαρρυντικά καθώς φάνηκε πως ακόμα και στην περίπτωση αδόμητων δεδομένων η εξόρυξη γνώσης έχει ικανοποιητική απόδοση. Αποδείχθηκε επίσης πως η ποιότητα του συνόλου των στιγμιότυπων εκπαίδευσης (δηλαδή το κατά πόσο τα στιγμιότυπα αυτά είναι αντιπροσωπευτικά του προβλήματος) και ο σαφής καθορισμός των κλάσεων του προς αντιμετώπιση προβλήματος παίζει σημαντικό ρόλο στην απόδοση της διαδικασίας εξόρυξης γνώσης.

Βιβλιογραφία

[1] Mitchell Tom, *Machine Learning*, MIT Press and the McGraw-Hill Companies, 1997.

[2] Han Jiawei and Kamber Micheline, *Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 2001*

[3] Joachims, T. *A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization*, Computer Science Technical Report CMU-CS-96-118, Carnegie Mellon University.

[4] Παπαγγελής Αθανάσιος, *Πρότυπη on-line κοινότητα ψηφοφοριών, τεχνικές και αλγόριθμοι*, Μεταπτυχιακή Εργασία, Τμήμα Μηχανικών Η/Υ & Πληροφορικής, 2001

[5] R. Feldman et al (1998),*Text mining at the term level*, In Proc. of the 2nd European Symposium on Princi-ples of Data Mining and Knowledge Discovery (PKDD'98), Nantes, France