ΣΥΛΛΟΓΗ ΚΑΙ ΠΟΙΟΤΙΚΗ ΑΝΑΛΥΣΗ ΣΕΙΣΜΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΕΛΛΑΔΙΚΟ ΧΩΡΟ – ΤΟ ΕΡΓΑΛΕΙΟ SEISMO-SURFER

Γιάννης Θεοδωρίδης[1,2], Γεράσιμος Μαρκέτος[1], Ιωάννης Σ. Καλογεράς[3]

[1] Πανεπιστήμιο Πειραιώς, Τμήμα Πληροφορικής
[2] Ερευνητικό Ακαδημαϊκό Ινστιτούτο Τεχνολογίας Υπολογιστών
[3] Εθνικό Αστεροσκοπείο Αθηνών, Γεωδυναμικό Ινστιτούτο

Περίληψη

Το SEISMO-SURFER είναι ένα εργαλείο για συλλογή και διαχείριση σεισμολογικών δεδομένων, ανάλυσή τους και εξόρυξη γνώσης. Η βάση δεδομένων του εργαλείου ανανεώνεται αυτόματα από απομακρυσμένες πηγές ενώ οι δυνατότητες του εργαλείου περιλαμβάνουν ερωτήσεις με βάση διαφορετικές παραμέτρους, αναλύσεις των δεδομένων για εξαγωγή χρήσιμων πληροφοριών και αναπαράσταση των αποτελεσμάτων σε χάρτες και γραφικές παραστάσεις. Αφού παραθέσουμε στοιχεία για την αρχιτεκτονική του συστήματος και τη λειτουργικότητά του, στην εργασία αυτή παρουσιάζουμε αποτελέσματα μελέτης ανάλυσης των σεισμολογικών δεδομένων του Ελλαδικού χώρου με τεχνικές εξόρυξης γνώσης. Προδιαγράφουμε επίσης τις επεκτάσεις που σχεδιάζουμε για το άμεσο μέλλον, τόσο στη βάση με ενσωμάτωση μακροσεισμικών δεδομένων (μακροσεισμικές εντάσεις, συνοδευόμενες από δημογραφική πληροφορία) όσο και στη λειτουργικότητα με τη μεταφορά του εργαλείου στο Web υπό μορφή *πύλης* (portal).

COLLECTING AND MINING SEISMIC DATA IN GREEK TERRITORY – THE SEISMO-SURFER TOOL

Yannis Theodoridis[1,2], Gerasimos Marketos[1], Ioannis S. Kalogeras[3]

[1] University of Piraeus, Dept. of Informatics
[2] Computer Technology Institute
[3] National Observatory of Athens, Geodynamic Institute

Abstract

SEISMO-SURFER is a tool for collecting, querying and mining seismic data. The database is automatically updated via remote sources while current functionality allows querying on different earthquake parameters, data analysis and mining for extracting useful information, and graphical representation of the results via maps, charts etc. After providing details about system architecture and functionality, in the present work we analyse results of data mining on seismological data of the Greek area. We also plan the prospects that will take place in the near future both related with the integration of macroseismic data (macroseismic intensity and demographic information) and with the functionality of a web based system (portal).

Λέξεις κλειδιά: βάσεις δεδομένων, ανάλυση σεισμολογικών δεδομένων, εξόρυξη γνώσης, οπτικοποίηση
Key words: databases, earthquake data analysis, data mining, visualization.

1. Εισαγωγή

Ο όγκος των σεισμολογικών δεδομένων είναι τεράστιος. Μόνο στον Ελλαδικό χώρο, κάθε χρόνο καταγράφονται κατά μέσο όρο 2000 συμβάντα, ενώ αν αναφερθούμε σε παγκόσμια κλίμακα, η συχνότητα είναι ένας σεισμός $M < 3R$ κάθε δευτερόλεπτο και ένας σεισμός $M \geq 3R$ κάθε 10 λεπτά. Ευτυχώς, ελάχιστοι από αυτούς γίνονται αισθητοί και ένα ακόμη μικρότερο ποσοστό προκαλεί ζημιές σε ανθρώπινες κατασκευές (λόγω μεγέθους, βάθους, απόστασης από κατοικημένες περιοχές κλπ.).

Προφανώς, η μελέτη του φαινομένου και της πληροφορίας που προκύπτει είναι έργο των Σεισμολόγων και, γενικότερα, των ανθρώπων των Γεωφυσικών Επιστημών. Από την άλλη πλευρά, οι άνθρωποι της Πληροφορικής, και ειδικότερα, αυτοί που ασχολούνται με την επιστημονική περιοχή της Διαχείρισης Πληροφορίας & Γνώσης (Information & Knowledge Management), βρίσκουν στο χώρο αυτό έναν πολύτιμο 'θησαυρό' από ακατέργαστα δεδομένα για περαιτέρω επεξεργασία και ανάλυση.

Η περιοχή αυτή της Πληροφορικής έχει να επιδείξει μια σειρά από τεχνικές που βρίσκουν στα σεισμολογικά δεδομένα μια χαρακτηριστική εφαρμογή τους. Τα σεισμολογικά δεδομένα είναι πολυδιάστατα και σε αντίθεση με τα παραδοσιακά αλφαριθμητικά δεδομένα για την αποθήκευση και ανάκτηση τους απαιτούνται ειδικές, περίπλοκες τεχνικές. Η μελέτη αυτών των τεχνικών αποτελεί αντικείμενο του

ερευνητικού πεδίου των βάσεων μη-παραδοσιακών δεδομένων, όπου εμπλέκεται χωρική ή/και χρονική πληροφορία. Η εξόρυξη γνώσης από αυτά τα δεδομένα, ευρύτερα γνωστή με τον όρο Data Mining, έχει στόχο την ανακάλυψη «κρυμμένης γνώσης» μέσα στους μεγάλους αυτούς όγκους πληροφορίας. Ως ενδεικτικά παραδείγματα, αναφέρουμε τα εξής:

- Η πληροφορία που εξάγεται από την καταγραφή ενός σεισμικού φαινομένου έχει έντονο το χωρικό (επίκεντρο, βάθος) και το χρονικό στοιχείο (στιγμή εκδήλωσης φαινομένου, διάρκεια). Σε συνδυασμό με το γεγονός ότι ο όγκος πληροφορίας που συσσωρεύεται ιστορικά είναι μεγάλος,, το αποτέλεσμα είναι μια τεράστια ιστορική βάση χωροχρονικών δεδομένων.
- Σε μια μεγάλη βάση δεδομένων που περιέχει σεισμολογικά δεδομένα, αξίζει να ερευνήσει κάποιος για «κρυμμένη γνώση», δηλαδή να ανακαλύψει πιθανές συσχετίσεις ή συνειρμούς («η εμφάνιση του Α συνήθως συνεπάγεται την εμφάνιση του Β»), να βρει πρότυπα ή μορφές που επαναλαμβάνονται, όπως, για παράδειγμα, προσεισμικές και μετασεισμικές ακολουθίες, ή ακραία φαινόμενα, και να εμφανίσει ομαδοποιήσεις στη χωρική ή τη χρονική διάσταση. Επίσης, μπορούν να χρησιμοποιηθούν διαφορετικά στρώματα θεματικής πληροφορίας, όπως γεωλογικοί, τεκτονικοί, πληθυσμιακοί χάρτες. Αυτό επιτρέπει τη διερεύνηση πιθανών συσχετίσεων μεταξύ του βαθμού της καταστροφής (έντασης του σεισμού) και της απόστασης από το επίκεντρο ή μεταξύ της έντασης και της γεωλογίας της περιοχής.

Από τα παραπάνω απλά παραδείγματα προκύπτει η χρησιμότητα και λειτουργικότητα του SEISMO-SURFER (Theodoridis 2003), ενός εργαλείου του οποίου η φιλοσοφία είναι να υποστηρίζει διαδικασίες συνεχούς εμπλουτισμού και αναβάθμισης ώστε να λαμβάνονται υπόψη οι νέες εγγραφές, άλλες παράμετροι, διαφορετικές τεχνικές εξορυξης γνώσης και οπτικοποίησης.

Η δομή της εργασίας είναι η εξής: Στην Ενότητα 2 παραθέτουμε στοιχεία για την αρχιτεκτονική του συστήματος, με έμφαση στο σχήμα της βάσης δεδομένων που έχει αναπτυχθεί, και τα βασικά στοιχεία που συνθέτουν τη λειτουργικότητά του. Στην Ενότητα 3 παρουσιάζουμε αποτελέσματα μελέτης ανάλυσης των σεισμολογικών δεδομένων του Ελλαδικού χώρου με δύο τεχνικές εξόρυξης γνώσης, κατηγοριοποίηση (classification) και ομαδοποίηση (clustering). Η Ενότητα 4 περιλαμβάνει μια συγκριτική παρουσίαση σχετικών εργαλείων/συστημάτων που έχουν αναπτυχθεί για διαχείριση γεωλογικών/γεωφυσικών δεδομένων και εξόρυξη γνώσης. Η εργασία ολοκληρώνεται με την Ενότητα 5 στην οποία προδιαγράφουμε επίσης τις επεκτάσεις που σχεδιάζουμε για το άμεσο μέλλον, τόσο στη βάση με ενσωμάτωση μακροσεισμικών δεδομένων (μακροσεισμικές εντάσεις, συνοδευόμενες από δημογραφική πληροφορία) όσο και στη λειτουργικότητα με τη μεταφορά του εργαλείου στο Web υπό μορφή *πύλης* (portal).

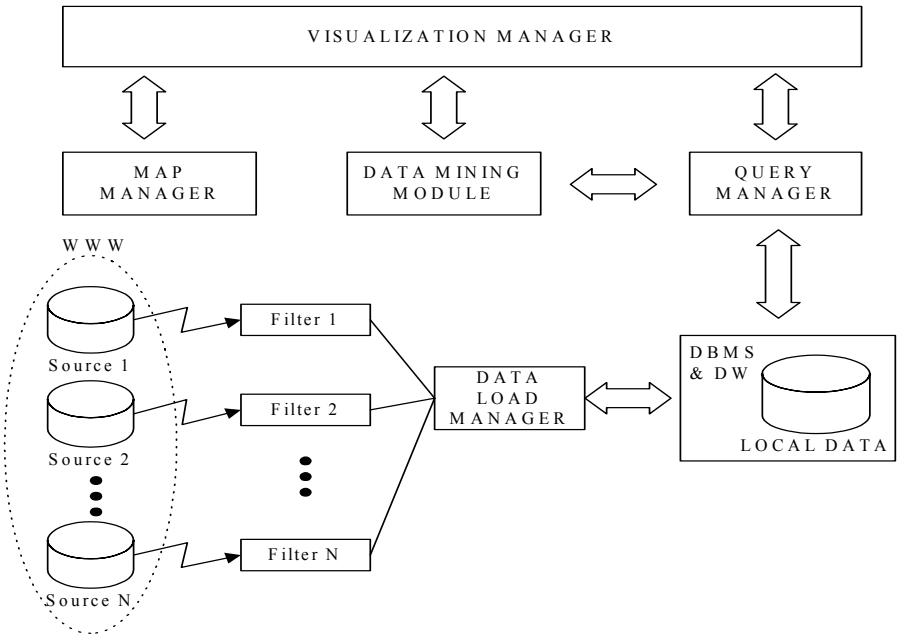
2. Το εργαλείο SEISMO-SURFER

Ο στόχος ανάπτυξης του λογισμικού ήταν να συνδυαστούν πρόσφατες ερευνητικές τάσεις και αποτελέσματα στις περιοχές των βάσεων χωρικών και χωροχρονικών δεδομένων και της εξόρυξης γνώσης από μεγάλες βάσεις δεδομένων, καθώς και καθιερωμένες τεχνικές οπτικοποίησης που χρησιμοποιούνται στα συστήματα GIS, όπως χάρτες κλπ. Επιπρόσθετα, επιθυμία μας ήταν να ενσωματώσουμε πηγές σεισμολογικών δεδομένων, διαθέσιμες στο Internet, που να παρέχουν τουλάχιστον τις τρεις βασικές παραμέτρους ενός σεισμικού φαινομένου: (α) γεωγραφικές συντεταγμένες επικέντρου, (β) χρόνος γένεσης και (γ) μέγεθος στην κλίμακα Richter. Στα πλαίσια αυτά, αναπτύξαμε το SEISMO-SURFER, την αρχιτεκτονική του οποίου καθώς και τη βάση δεδομένων που διαχειρίζεται παρουσιάζουμε στη συνέχεια.

2.1 Η αρχιτεκτονική και η λειτουργικότητα του συστήματος

Η αρχιτεκτονική του SEISMO-SURFER απεικονίζεται στο Σχήμα 1. Συγκεκριμένα, ένα σύνολο φίλτρων αναλαμβάνουν τον καθαρισμό και την ομογενοποίηση των δεδομένων που προέρχονται από τις εξωτερικές πηγές και η μονάδα φόρτωσης δεδομένων (data load manager) αναλαμβάνει την αποθήκευσή τους στην τοπική βάση δεδομένων. Η μονάδα διαχείρισης ερωτήσεων (query manager) υποστηρίζει τη διαδικασία σχηματισμού ερωτήσεων από το χρήστη με γραφικό τρόπο. Πάνω στα δεδομένα που επιλέγονται με την εκτέλεση των ερωτήσεων είναι δυνατό να εφαρμόζονται τεχνικές εξόρυξης γνώσης, λειτουργία που υποστηρίζεται από τη μονάδα εξόρυξης (data mining module). Τα αποτελέσματα των ερωτήσεων και των τεχνικών εξόρυξης παρουσιάζονται, μέσω του διαχειριστή οπτικοποίησης (visualization manager), σε χάρτες και γραφικές παραστάσεις.

Οι λειτουργίες που ήδη παρέχει στην τρέχουσα έκδοση (v.2, Ιούνιος 2003) ή θα παρέχει στο άμεσο μέλλον το εργαλείο SEISMO-SURFER ταξινομούνται σε τέσσερις κύριες κατηγορίες:



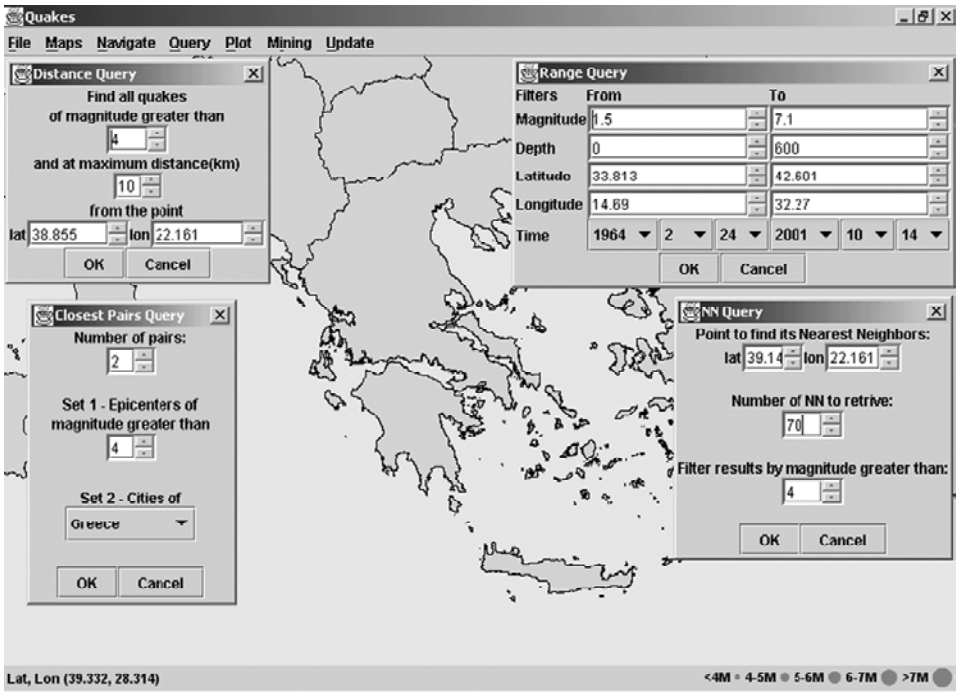
Σχήμα 1. Η αρχιτεκτονική του SEISMO-SURFER

- *Αξιοποίηση εξωτερικών πηγών πληροφορίας*. Με αυτόματο τρόπο η τοπική βάση δεδομένων μπορεί να ενημερώνεται από τις εξωτερικές πηγές πληροφορίας. Έως τώρα έχουν ενσωματωθεί στο σύστημα δύο πηγές σεισμολογικών δεδομένων: η πρώτη συντηρείται από το Γεωδυναμικό Ινστιτούτο και καλύπτει τον Ελλαδικό χώρο[1] (GI-NOA) ενώ η δεύτερη έχει παγκόσμια κάλυψη από την United States Geological Survey (NEIC-USGS).
- *Υποβολή ερωτήσεων που αφορούν μέγεθος, χώρο (επίκεντρο) και χρόνο*. Το Seismo-Surfer υποστηρίζει ερωτήσεις περιοχής (range query), απόστασης (distance query), πλησιέστερου γείτονα (nearest-neighbor query) και εγγυτέρων ζευγών (closest-pairs query), όσον αφορά στην χωρική πληροφορία (επίκεντρο δόνησης), με επιπλέον περιορισμούς ως προς μέγεθος, χώρο και χρόνο. Για παράδειγμα, αναζήτηση των σεισμικών επικέντρων σε απόσταση μέχρι 50 km από την πόλη της Αθήνας τα τελευταία δέκα έτη, εύρεση του κοντινότερου επίκεντρου με βάση ένα σημείο πάνω στο χάρτη, εύρεση των σεισμών άνω των 5R που συνέβησαν σε κοντινή απόσταση από πυκνοκατοικημένες περιοχές, κοκ. Η εισαγωγή των ερωτήσεων γίνεται πλήρως γραφικά (βλ. Σχήμα 2) και τα αποτελέσματα μπορούν να παρουσιαστούν σε πίνακες, γραφικές παραστάσεις και χάρτες.
- *Απλές λειτουργίες εξόρυξης γνώσης από δεδομένα*, όπως ομαδοποίηση επικέντρων με ενσωμάτωση της δημοφιλούς τεχνικής k-means clustering (MacQueen 1967) και, μελλοντικά, δημιουργίας σεισμικών προφίλ κάθε περιοχής και χρονικής περιόδου, ανακάλυψης περιοχών με παρόμοια σεισμική συμπεριφορά, εντοπισμού επαναλαμβανόμενων φαινομένων και συσχέτισης των σεισμολογικών παραμέτρων με πληροφορίες όπως ζημιές, πληθυσμός της περιοχής, κλπ.
- *Εντοπισμός φαινομένων*. Η εξόρυξη γνώσης μπορεί επίσης να χρησιμοποιηθεί για τον αυτόματο εντοπισμό σημασιολογικών στοιχείων από αποθηκευμένα δεδομένα, όπως για παράδειγμα, ο χαρακτηρισμός του κυρίως σεισμού και πιθανών ισχυρών μετασεισμών σε μια δεδομένη σεισμική ακολουθία, με τεχνικές εύρεσης προτύπων (pattern finding).

Για να είναι πιο φιλική η οπτική παρακολούθηση της σεισμικής δραστηριότητας ο χρήστης μπορεί να επιλέξει μεταξύ τριών ψηφιακών χαρτών (του παγκόσμιου, της Μεσογείου, της Ελλάδας) ή να δηλώσει τα όρια γεωγραφικού μήκους/πλάτους που ενδιαφέρεται να εστιάσει. Παρέχεται επίσης η βασική δυνατότητα περιήγησης στο χάρτη (zoom in/out, μετακίνηση).

Μια ακόμη ενδιαφέρουσα δυνατότητα του εργαλείου είναι η ικανότητα οπτικοποίησης των επικέντρων ανά χρονική περίοδο. Για το σκοπό αυτό υπάρχει ένας χάρακας (slider) με τον οποίο ο χρήστης μπορεί να καθορίσει τις ακραίες και ενδιάμεσες τιμές και καθώς μετακινεί τον δρομέα να 'πλοηγείται' εικονικά στο χρόνο λαμβάνοντας στο χάρτη της οθόνης τα αντίστοιχα επίκεντρα.

[1] Με πλήρη, σύμφωνα με τα όσα προαναφέραμε για τις τρεις βασικές παραμέτρους (επίκεντρο, χρονική στιγμή, μέγεθος), στοιχεία για όλους τους σεισμούς που έχουν καταγραφεί από το 1964 έως σήμερα.

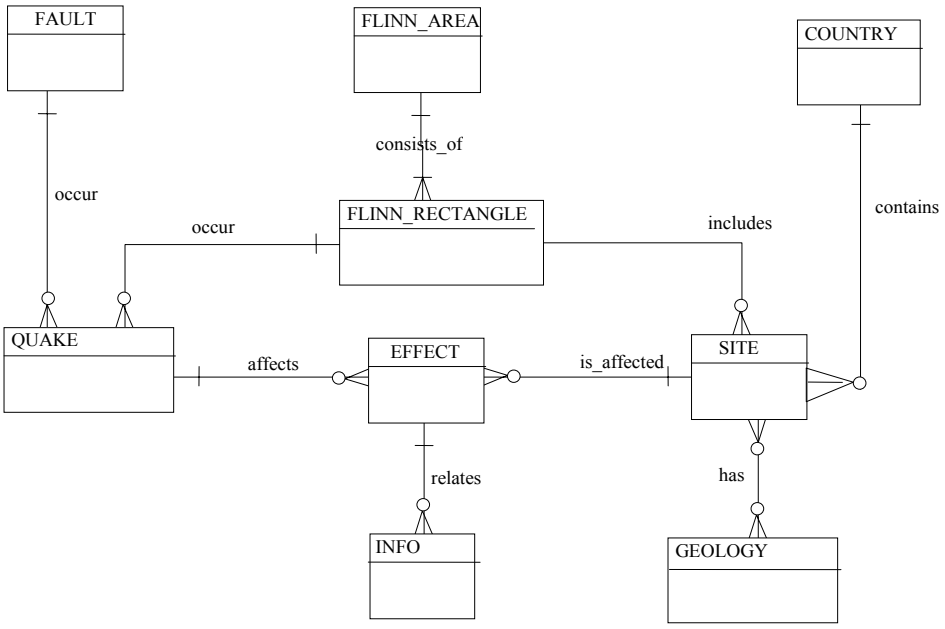


Σχήμα 2.Επιλογές του SEISMO-SURFER για χωροχρονικές ερωτήσεις

2.2 Η βάση σεισμολογικών δεδομένων

Το διάγραμμα του Σχήματος 3 παρουσιάζει τις οντότητες και τις μεταξύ τους συσχετίσεις τους που συνθέτουν τη βάση δεδομένων του SEISMO-SURFER.

Η ελάχιστη πληροφορία που επιθυμεί να γνωρίζει κάποιος είναι ο χρόνος γένεσης του σεισμού, το γεωγραφικό μήκος/πλάτος και βάθος του επικέντρου (πίνακας QUAKE). Αυτό όμως δεν βοηθά ιδιαίτερα στην ερμηνεία των αποτελεσμάτων. Για το σκοπό αυτό χρησιμοποιήσαμε το γεωγραφικό διαχωρισμό του Ελλαδικού χώρου κατά Flinn, που έχει υιοθετήσει το Γεωδυναμικό Ινστιτούτο όταν αναφέρεται σε γεωγραφικές περιοχές[2] (πίνακας FLINN_AREA). Επειδή οι περιοχές κατά Flinn καλύπτουν πολυγωνικές περιοχές στο χάρτη, καθεμία αποτελείται από πολλά ορθογώνια τα οποία καταγράφονται στον πίνακα FLINN_RECTANGLE.



Σχήμα 3. Η βάση δεδομένων του SEISMO-SURFER

[2] Το Γεωδυναμικό Ινστιτούτο βασίζεται στον ευρύτερο διαχωρισμό σε παγκόσμιο επίπεδο κατά Flinn and Engdahl πάνω στον οποίο για πρακτικούς λόγους έχει γίνει περαιτέρω διαχωρισμός της περιοχής "Greece".

Στον πίνακα SITE αποθηκεύονται τα γεωγραφικά όρια καθώς και διοικητικές/δημογραφικές πληροφορίες των ΟΤΑ, όπως διαμορφώθηκαν με το Ν 2539/97 "Καποδίστρια", με πρόβλεψη για ενσωμάτωση τοποθεσιών και άλλων κρατών πέραν της Ελλάδας (πίνακας COUNTRY).

Ένα πρόσθετο στοιχείο που επιθυμούμε να γνωρίζουμε για κάθε σεισμική δόνηση είναι πάνω σε ποιο ρήγμα εμφανίστηκε. Για το σκοπό αυτό, στον πίνακα FAULT καταγράφονται τα ρήγματα που μας ενδιαφέρουν. Ο πίνακας EFFECT καταγράφει τις μακροσεισμικές παρατηρήσεις, εκείνα δηλαδή τα στοιχεία που αφορούν το ζευγάρι ένταση - τοποθεσία. Επίσης για κάθε τοποθεσία μας ενδιαφέρουν και τα γεωλογικά στοιχεία της ώστε να μπορούμε να βλέπουμε τι επίδραση έχουν αυτά στα σεισμικά φαινόμενα (πίνακας GEOLOGY). Τέλος, ο πίνακας INFO χρησιμοποιείται για να αποθηκεύονται φωτογραφίες, περιγραφές και αναφορές σχετικές με τα αποτελέσματα που προκάλεσε μια σεισμική δόνηση σε μια περιοχή.

Η βάση δεδομένων του SEISMO-SURFER περιέχει σήμερα πάνω από 33.000 εγγραφές σεισμικών δονήσεων στον πίνακα QUAKE (πλήρης βάση για $M \geq 4R$ από το 1964 έως σήμερα, για $M \geq 6R$ από το 1900 έως σήμερα), πάνω από 10.000 εγγραφές μακροσεισμικών παρατηρήσεων στον πίνακα EFFECT (πρώτο δείγμα από 30 περίπου ισχυρούς σεισμούς του ελληνικού χώρου) και 6.133 εγγραφές ΟΤΑ στον πίνακα SITE, με το περιεχόμενο της βάσης να εμπλουτίζεται συνεχώς.

3. Εξόρυξη γνώσης από σεισμολογικά δεδομένα

Οι τεχνολογίες εξόρυξης γνώσης έχουν ωριμάσει αρκετά ώστε να αποτελούν πλέον συστατικό αρκετών συστημάτων βάσεων δεδομένων. Για παράδειγμα στα συστήματα Oracle 9i και Microsoft SQL Server 2000, η εξόρυξη γνώσης έχει ενσωματωθεί στο βασικό 'πυρήνα' του λογισμικού. Τα παραπάνω συστήματα παρέχουν τόσο εργαλεία εξόρυξης γνώσης από κάποια βάση δεδομένων όσο και από κάποια αποθήκη δεδομένων που επίσης μπορεί εύκολα να αναπτυχθεί. Επιπλέον, αυτά τα εργαλεία παρέχουν τη δυνατότητα ανάπτυξης εφαρμογών που να χρησιμοποιούν τεχνικές εξόρυξης γνώσης.

Για την ανάλυση των δεδομένων του SEISMO-SURFER, επιλέξαμε να εστιάσουμε σε σχετικά μεγάλους σεισμούς, συγκεκριμένα σε μεγέθη $M \geq 4.5R$ για το χρονικό διάστημα 1964-2002 (το πλήθος των εγγραφών με αυτά τα χαρακτηριστικά ανέρχεται στις 6.730). Στη συνέχεια εφαρμόσαμε δύο δημοφιλείς τεχνικές εξόρυξης γνώσης, την κατηγοριοποίηση με δέντρα αποφάσεων και την ομαδοποίηση.

3.1 Κατηγοριοποίηση

Η διαδικασία της κατηγοριοποίησης βασίστικε πάνω σε δέντρα αποφάσεων, όπως έχουν ενσωματωθεί στο εργαλείο Analysis Services του Microsoft SQL Server 2000[3]. Ο αλγόριθμος αυτός χρησιμοποιείται κυρίως για πρόβλεψη και αποτελεί μια από τις βασικές τεχνικές της ανάλυσης αποφάσεων. Χρησιμοποιείται δηλαδή για ανάλυση προβλημάτων υπό το καθεστώς αβεβαιότητας.

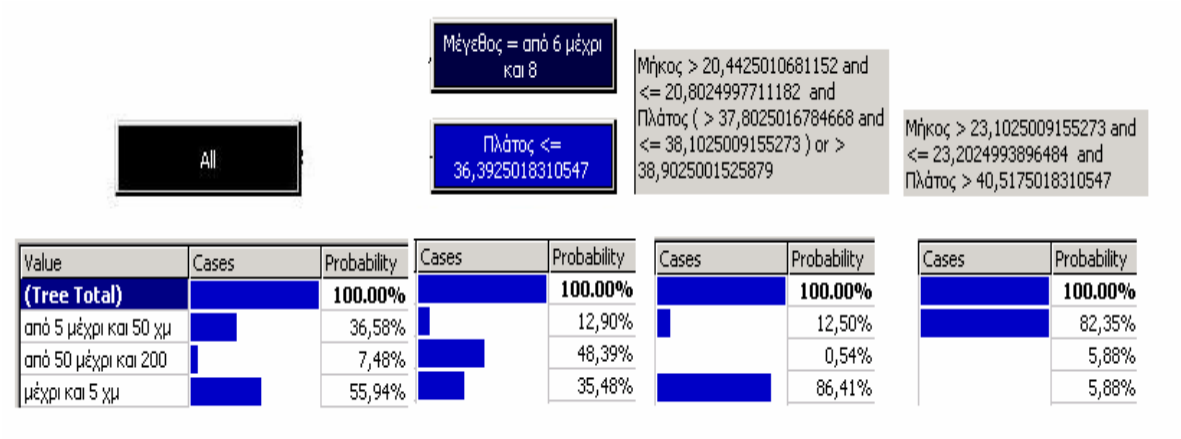
Η εφαρμογή αυτού του αλγορίθμου απαιτεί να γνωρίζουμε ποιο είναι το σύνολο δεδομένων και τι προσπαθούμε να προβλέψουμε. Το τελικό αποτέλεσμα θα είναι μια σειρά από 'αν' που καταλήγουν σε κάποια συμπεράσματα (συνήθως μορφής Boolean). Δηλαδή αν συμβαίνει το Α και το Β και το Γ τότε κατά x% συμπεραίνουμε το Δ, κατά y% το Ε κοκ. Η ιδέα αυτή μπορεί να εφαρμοστεί και στους σεισμούς. Το σύνολο δεδομένων μπορεί να είναι όλα τα χαρακτηριστικά του σεισμού: μέγεθος, βάθος, γεωγραφική περιοχή, χρονική στιγμή. Αυτό που προσπαθούμε να 'συμπεράνουμε' είναι η πιθανή τιμή κάποιου από τα παραπάνω χαρακτηριστικά (το σύνολο πρόβλεψης). Αφού εκπαιδεύσουμε το δέντρο αποφάσεων, μπορούμε να δούμε ένα από τα στοιχεία του συνόλου πρόβλεψης που είχαμε ορίσει πριν. Όλα αυτά συνιστούν ένα μοντέλο εξόρυξης (mining model).

Ένα ενδεικτικό αποτέλεσμα κατηγοριοποίησης είναι το συμπέρασμα για το βάθος του επικέντρου, σε εξάρτηση με τις υπόλοιπες παραμέτρους ή, με άλλα λόγια, η απάντηση στην ερώτηση: «Αν γνωρίζουμε το μέγεθος, το μήκος και το πλάτος μπορούμε να συμπεράνουμε κάτι σχετικά με το βάθος;» (Σχήμα 4). Πράγματι, παρατηρούμε την πολύ μεγάλη αύξηση που σημειώνεται στην πιθανότητα (48.39%) να πραγματοποιηθεί σεισμός σε βάθος από 50 μέχρι και 200 km στην περίπτωση που γνωρίζουμε ότι θα είναι μεγέθους $M \geq 6R$ και θα σημειωθεί σε περιοχή με γεωγραφικό πλάτος ≤ 36.39, δηλαδή ο χώρος του Ν. Αιγαίου και της Κρήτης (από 7.48% που είναι για όλο τον Ελλαδικό χώρο και όλα τα μεγέθη).

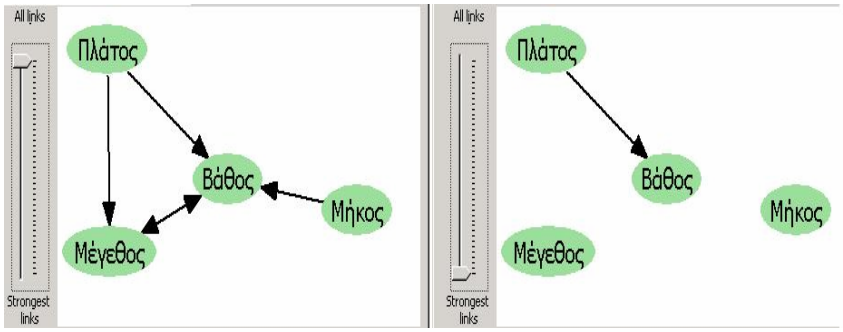
Συμπληρωματικό του παραπάνω δέντρου αποφάσεων είναι ότι συμπέρασμα για το βάθος μπορεί να εξαχθεί μόνο σε σχέση με το γεωγραφικό πλάτος και όχι με τα άλλα χαρακτηριστικά (γεωγραφικό μήκος,

[3] Η επιλογή του λογισμικού SQL Server έναντι του αντίστοιχου της Oracle, ειδικά για την κατηγοριοποίηση, στηρίχθηκε στο γεγονός ότι το πρώτο διαθέτει ενσωματωμένο περιβάλλον εκπαίδευσης του μοντέλου και παρουσίασης των αποτελεσμάτων σε αντίθεση με το δεύτερο που προϋποθέτει την συγγραφή κατάλληλου κώδικα προγράμματος για την προσπέλαση των αλγορίθμων που διαθέτει, κάτι που αποτελεί μέρος των μελλοντικών στόχων του συστήματος.

μέγεθος). Αυτό απεικονίζεται παραστατικά με τα δίκτυα εξάρτησεων των χαρακτηριστικών πάνω στα οποία έχει εκπαιδευτεί το μοντέλο: στο Σχήμα 5, αριστερά απεικονίζεται το δίκτυο εξαρτήσεων όλων των χαρακτηριστικών (με 'στόχο' το βάθος) ενώ δεξιά εμφανίζονται μόνο οι ισχυροί δεσμοί (ένας, στο παράδειγμά μας).



Σχήμα 4. Αποτελέσματα κατηγοριοποίησης με προβλεπόμενο στοιχείο το βάθος επικέντρου



Σχήμα 5. Δίκτυα εξαρτήσεων με προβλεπόμενο στοιχείο το βάθος επικέντρου

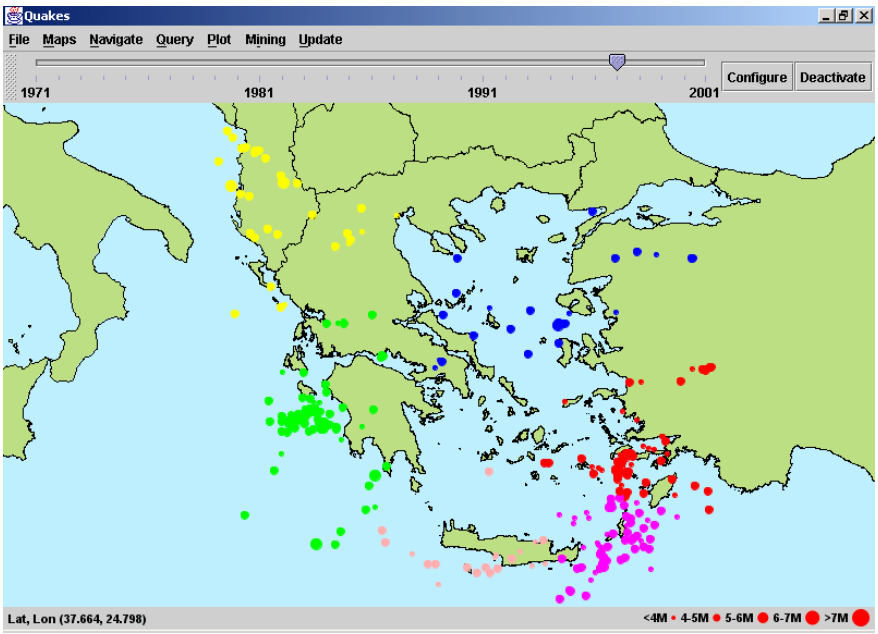
3.2 Ομαδοποίηση

Η μέθοδος της ομαδοποίησης (clustering) αποτελεί και αυτή μία από τις πιο διαδεδομένες και χρησιμότερους μεθόδους για εξόρυξη γνώσης από δεδομένα. Η ομαδοποίηση γίνεται με βάση κάποια χαρακτηριστικά που χαρακτηρίζουν όλα τα δεδομένα πάνω στα οποία θα εφαρμοστεί. Στόχος του κάθε αλγορίθμου ομαδοποίησης είναι να βρει το πώς σχετίζονται τα δεδομένα μεταξύ τους έχοντας ως βάση τα χαρακτηριστικά αυτά. Αυτό έχει ως συνέπεια να βγάζουμε πιο εύκολα κανόνες σχετικά με την συμπεριφορά των εγγραφών που ανήκουν σε μια συγκεκριμένη ομάδα, μελετώντας πια ομάδες και όχι τις εγγραφές ατομικά. Έτσι, αποφεύγοντας να εξετάσουμε κάποιες μεμονωμένες και τυχαίες εγγραφές της βάσης για να βγάλουμε συμπεράσματα, οδηγούμαστε σε πιο αξιόπιστα αποτελέσματα, για τον πολύ απλό λόγο ότι αν κάποιος κανόνας είναι έγκυρος για ένα από τα στοιχεία υπάρχει μεγάλη πιθανότητα να είναι έγκυρος και για τα άλλα στοιχεία που ανήκουν στην ίδια ομάδα και επομένως θα έχουν παρόμοια χαρακτηριστικά και θα συμπεριφέρονται ανάλογα.

Μια ομαδοποίηση που ουσιαστικά απαιτείται να γίνει σε σεισμολογικά δεδομένα είναι αυτή που γίνεται με βάση το επίκεντρο των σεισμών (γεωγραφικό μήκος/πλάτος). Από την εφαρμογή του αλγορίθμου ομαδοποίησης k-means στα δεδομένα προέκυψαν έξι ομάδες: Ν.Δ. Ελλάδα (Ιόνιο, Δ. Στερεά), Κρήτη, Ν.Α. Αιγαίο, Κ. Αιγαίο, Β. Αιγαίο, Β.Δ. Ελλάδα (Ήπειρος – Δ. Μακεδονία). Οπτικοποίηση του αποτελέσματος επιτυγχάνουμε μέσω του εργαλείου SEISMO-SURFER (Σχήμα 6), όπου αν παρατηρήσουμε τις ομάδες που προκύπτουν, θα δούμε ότι οι τρεις πρώτες οριοθετούν τμήμα του Ελληνικού τόξου (Ιόνιο – Κρήτη – Δωδεκάνησα).

4. Σχετικές εργασίες – Σύγκριση

Στη συνέχεια παραθέτουμε σχετικές με το SEISMO-SURFER εργασίες, μια συγκριτική παρουσίαση των οποίων υπάρχει στον Πίνακα 1.



Σχήμα 6. Ομαδοποίηση επικέντρων στον Ελλαδικό χώρο

Οι Han κ.α. (1997) έχουν αναπτύξει το Geo-Miner, ένα σύστημα για εξόρυξη γνώσης από χωρικά δεδομένα. Όπως και στο SEISMO-SURFER, στόχος του Geo-Miner είναι η εξαγωγή 'κρυμμένης' γνώσης, η ανακάλυψη χωρικών σχέσεων και προτύπων.

Ένα άλλο σύστημα που ασχολείται επίσης με γεωγραφικά δεδομένα είναι το Common-GIS (Kretschmer and Roccatagliata, 2000), το οποίο υποστηρίζει τους χρήστες στην οπτικοποίηση και την ανάλυση στατιστικών δεδομένων που σχετίζονται με χωρικά αντικείμενα. Στόχος είναι η ανάπτυξη ενός συστήματος που θα επιτρέπει στους χρήστες να εποπτεύουν και να αναλύουν θεματικά δεδομένα με γεωγραφική αναφορά.

Οι Andrienko and Andrienko (1999) προτείνουν ένα ολοκληρωμένο περιβάλλον (Descartes / Kepler) για ανάλυση χωρικών δεδομένων που υποστηρίζεται από τεχνικές εξόρυξης γνώσης και οπτικοποίησης. Στόχος είναι η ολοκλήρωση παραδοσιακών εργαλείων εξόρυξης γνώσης με αυτόματη χαρτογραφική οπτικοποίηση και εργαλεία για αλληλεπιδραστικό χειρισμό των γραφικών παρουσιάσεων, ώστε να μπορεί ο χρήστης να βλέπει με τη μορφή χαρτών τα δεδομένα που αποτελούν την πηγή αλλά και το αποτέλεσμα της εξόρυξης.

Τέλος, το GEODE (Geo-Data Explorer) είναι μία εφαρμογή που έχει αναπτυχθεί από την USGS για την παροχή δεδομένων με γεωγραφική αναφορά στους χρήστες (USGS). Στόχος είναι η δημιουργία ενός portal που θα παρέχει δεδομένα πραγματικού χρόνου και ανάλυση μέσω του Internet χωρίς την ανάγκη εξειδικευμένου λογισμικού ή εκπαίδευσης.

Πίνακας 1. Συγκριτικά χαρακτηριστικά εργαλείων σχετικών με το SEISMO-SURFER

	Geo-Miner	**Common-GIS**	**Descartes / Kepler**	**GEODE**	**Seismo-Surfer**
Web περιβάλλον	Όχι	Ναι	Όχι	Ναι	Όχι
Δυναμική πληροφορία					
- Ανάκτηση δεδομένων μέσω Internet	Όχι	Ναι	Ναι	Ναι	Ναι
- Αποθήκευση/Φόρτωση Χαρτών	Όχι	Όχι	Ναι	Ναι	Ναι
Τεχνικές Εξόρυξης Γνώσης					
- Ομαδοποίηση	Ναι	Όχι	Ναι	Όχι	Ναι
- Κατηγοριοποίηση	Ναι	Όχι	Ναι	Όχι	Όχι
- Κανόνες συσχέτισης	Ναι	Όχι	Ναι	Όχι	Όχι
Τεχνικές Οπτικοποίησης					
- Χάρτες	Ναι	Ναι	Ναι	Ναι	Ναι
- Γραφικές παραστάσεις	Ναι	Ναι	Ναι	Όχι	Ναι
Τρόποι Υποβολής Ερωτημάτων					
- Γραφική διεπαφή	Ναι	Ναι	Ναι	Ναι	Ναι
- Γλώσσα επερωτήσεων	Ναι	Όχι	Όχι	Όχι	Όχι

5. Συμπεράσματα - Επόμενα βήματα

Στην παρούσα εργασία περιγράψαμε την αρχιτεκτονική και τη λειτουργικότητα του εργαλείου SEISMO-SURFER, το οποίο αναπτύσσουμε για τη συλλογή, διαχείριση και ανάλυση των σεισμολογικών δεδομένων. Παραθέσαμε επίσης ενδεικτικά αποτελέσματα αυτής της ανάλυσης με χρήση τεχνικών εξόρυξης γνώσης. Επειδή η συλλογή σεισμολογικών δεδομένων – και ιδιαίτερα μακροσεισμικών δεδομένων – συνεχίζεται, η ανάλυσή τους με διάφορους αλγορίθμους βρίσκεται σε προκαταρκτικό στάδιο και αποτελεί το βασικό σκέλος της τρέχουσας ερευνητικής μας δραστηριότητας. Στόχος αυτής της δραστηριότητας είναι αφενός η τεκμηρίωση συμπερασμάτων της Σεισμολογίας για συσχετίσεις και αλληλεξαρτήσεις των χαρακτηριστικών μιας σεισμικής δόνησης και με τεχνικές της Πληροφορικής, αφετέρου η δυείσδυση σε μεγαλύτερο βάθος για την καλύτερη κατανόηση αυτού του φαινομένου.

Με γνώμονα την εκμετάλλευση της εμπειρίας και της πληροφορίας που συλλέγει διαρκώς το Γεωδυναμικό Ινστιτούτο αλλά και τη συγκριτική θέση του εργαλείου σε σχέση με άλλα συστήματα (βλ. Πίνακα 1), τα επόμενα βήματα εμπλουτισμού του κινούνται σε 4 κατευθύνσεις:

- *Επέκταση της βάσης δεδομένων*: τόσο με νέες εγγραφές όσο (κυρίως) με νέα χαρακτηριστικά, με πιο ενδιαφέροντα τις μακροσεισμικές παρατηρήσεις για κάθε περιοχή, τη γεωλογία και τα ρήγματα του Ελλαδικού χώρου (πλήρωση πινάκων EFFECT, GEOLOGY, FAULT της βάσης δεδομένων του Σχήματος 3).
- *Ενσωμάτωση περισσότερων τεχνικών εξόρυξης γνώσης:* Όπως έχουμε προαναφέρει, αυτή τη στιγμή στο εργαλείο έχει ενσωματωθεί μόνο ο αλγόριθμος ομαδοποίησης K-means. Ο στόχος είναι να ενσωματωθούν προγραμματιστικά και οι υπόλοιποι αλγόριθμοι που υποστηρίζονται από την Oracle: κατηγοριοποίησης (Naïve Bayes και Adaptive Bayes Network για πρόβλεψη), ομαδοποίησης (O-cluster ομαδοποίηση βασισμένη σε πλέγμα), σημαντικότητας χαρακτηριστικού (για την αξιολόγηση της χρησιμότητας κάθε μη στοχευόμενου χαρακτηριστικού για την εξόρυξη γνώσης), εύρεσης κανόνων συσχέτισης (A-priori).
- *Επιπλέον λειτουργικότητα:* Σκοπεύουμε να προσθέσουμε περισσότερα γραφήματα, να εμπλουτίσουμε τα ερωτήματα με περισσότερες δυνατότητες παραμετροποίησης από πλευράς χρήστη, να βελτιώσουμε τον τρόπο πλοήγησης στον ψηφιακό χάρτη και να ενσωματώσουμε νέες εξωτερικές πηγές άντλησης δεδομένων.
- *Web έκδοση:* Μεταφέρουμε το SEISMO-SURFER από desktop σε web έκδοση ώστε να προσβάσιμο από όλους τους χρήστες του Internet και να αποτελεί μια πύλη πληροφόρησης (portal) σχετικά με τη σεισμική δραστηριότητα στην Ελλάδα και ευρύτερα.

Βιβλιογραφία

Andrienko, G., Andrienko N., 1999: *Knowledge-based visualization to support spatial data mining*. Proceedings of the 3rd Symposium on Intelligent Data Analysis, Amsterdam, the Netherlands, 1999.

GI-NOA: *Earthquake Catalog*. Available at http://www.gein.noa.gr/services/cat.html (accessed: 26 May 2004).

Han, J., Koperski K., Stefanovic N., 1997: *GeoMiner: a system prototype for spatial data mining*. Proceedings of ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, 1997.

Kretschmer, U., Roccatagliata E., 2000: *CommonGIS: a European project for an easy access to geo-data*. Proceedings of the 2nd European GIS Education Seminar, EUGISES, Budapest, Hungary, 2000.

MacQueen J., 1967: *Some methods for classification and analysis of multivariate observations*. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 1967.

NEIC-USGS: *Earthquake Search*. Available at http://neic.usgs.gov/neis/epic/epic_global.html (accessed: 26 May 2004).

Theodoridis Y. 2003: *SEISMO-SURFER: A prototype for collecting, querying and mining seismic data*. In Manolopoulos et al. (eds.) Advances in Informatics – Post Proceedings of the 8th Panhellenic Conference on Informatics, LNCS #2563, Springer-Verlag, Berlin.

USGS: *USGS GEO-DATA Explorer*. Available at http://geode.usgs.gov/ (accessed: 26 May 2004).