



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**Μελέτη Θεμάτων Συστημάτων Συστάσεων και Επέκταση
Συστήματος Συστάσεων Εστιατορίων**

Δημήτριος Σαλταούρας

Επιβλέπουσα : Ιζαμπώ Καραλή, Επίκουρη Καθηγήτρια ΕΚΠΑ

ΑΘΗΝΑ

ΑΠΡΙΛΙΟΣ 2016

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Μελέτη Θεμάτων Συστημάτων Συστάσεων και Επέκταση
Συστήματος Συστάσεων Εστιατορίων

Δημήτριος Α. Σαλταούρας

A.M.: 1115201100114

ΕΠΙΒΛΕΠΟΥΣΑ: Ιζαμπώ Καράλη, Επίκουρη Καθηγήτρια ΕΚΠΑ

ΠΕΡΙΛΗΨΗ

Ο σκοπός της παρούσας πτυχιακής είναι η μελέτη των συστημάτων συστάσεων, των μεθόδων αξιολόγησης τους καθώς και η εφαρμογή της θεωρίας αυτών στην επέκταση του Restaurant Finder. Το Restaurant Finder είναι ένα σύστημα συστάσεων για εστιατόρια του Λονδίνου, που δημιούργησε η φοιτήτρια Δωροθέα Τσιμπίδη.

Πραγματοποιήθηκαν συνολικά τρεις επεκτάσεις. Αρχικά, η μεταφορά της εφαρμογής σε έναν εξυπηρέτη, ώστε να είναι προσβάσιμη από το διαδίκτυο. Το ρόλο του web εξυπηρέτη έπαιξε ο προσωπικός μου υπολογιστής. Έπειτα, σκοπός ήταν η επέκταση των δυνατοτήτων της εφαρμογής προσθέτοντας στη βάση εστιατόρια της Αθήνας και παράλληλα ο έλεγχος της φορητότητας (portability) της, δηλαδή ο έλεγχος για το πόσο εύκολη είναι αυτή η διαδικασία. Τελευταία επέκταση ήταν η τροποποίηση του προγράμματος συλλογής δεδομένων, ώστε να παίρνει ως πληροφορία κριτικές όλων των γλωσσών, αντί μόνο τις κριτικές στα αγγλικά. Στην τροποποίηση αυτή έγινε χρήση του google translate για τη μετάφραση κειμένου.

Κατά τη διαδικασία υλοποίησης καθεμιάς από τις παραπάνω επεκτάσεις χρειάστηκε η μελέτη αντίστοιχων πεδίων. Συνεπώς, για τη μεταφορά της εφαρμογής σε εξυπηρέτη χρειάστηκε η μελέτη των αρχιτεκτονικών εφαρμογών και πιο συγκεκριμένα του client-server μοντέλου, καθώς επίσης και η μελέτη μιας τεχνικής που ονομάζεται Μετάφραση Διεύθυνσης Δικτύου (Network Address Translation, NAT). Επίσης, μελετήθηκε το πεδίο της Επεξεργασίας Φυσικής Γλώσσας.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Συστήματα Συστάσεων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Αρχιτεκτονική Πελάτη-εξυπηρέτη, Μηχανική Μετάφραση, Εστιατόρια Αθήνας

ABSTRACT

The purpose of this project is the study of recommender systems, their evaluation methods and the expansion of Dorothea Tsimpidi's project. The work of Dorothea implements a recommender system restaurants in London.

The expansion of this system was implemented in three stages. At first, the application should be moved to a server, in order to accessible from the web. My personal computer has been used as the server. The next task was the expansion of the application's database by adding Athenian restaurants as well as checking the application for portability, that means checking how easy this task is. Finally, the program that collects data had to be modified in order to collect information from reviews in many other languages apart from English. A java api implementing google translate has been used for text translation.

During the process of each one of the above extensions the need for theoretical study of some corresponding fields popped up. Consequently, theoretical knowledge onto web applications' architectures, specifically the client-server model, should be studied before moving the application to the server. In addition, the study of the web protocol called Network Address Translation (Network Address translation, NAT) as well as the field of Natural Language Processing have been studied.

SUBJECT AREA: Recommender Systems

KEYWORDS: Client-Server Model, Machine Translation, Restaurants of Athens

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ	8
2.ΣΥΣΤΗΜΑΤΑ ΣΥΣΤΑΣΕΩΝ.....	10
2.1 Τύποι Συστημάτων Συστάσεων.....	10
2.1.1 Συνεργατικό Φιλτράρισμα (Collaborative Filtering – CF) ...	11
2.1.2 Φιλτράρισμα βασισμένο στο Περιεχόμενο (Content-Based Filtering – CBF)	12
2.1.3 Φιλτράρισμα βασισμένο στη Γνώση (Knowledge-based Filtering – KBF).....	13
2.1.4 Φιλτράρισμα βασισμένο στα Δημογραφικά Στοιχεία (Demographic Filtering – DF)	14
2.1.5 Υβριδικό φιλτράρισμα (Hybrid filtering)	14
2.2 Αξιολόγηση Συστημάτων Συστάσεων.....	15
3.ΑΡΧΕΣ ΔΙΑΔΙΚΤΥΟΥ.....	19
3.1 Αρχιτεκτονικές Web Εφαρμογών	19
3.2 Μεταφραστές Διευθύνσεων Δικτύου	20
4.ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ.....	23
4.1 Μηχανική Μετάφραση	25
4.1.1 Μετάφραση Βασισμένη Σε Κανόνες (Rule-based Machine Translation)	26
4.1.2 Μετάφραση Βασισμένη Στην Στατιστική (Statistical Machine Translation - SMT)	27
4.1.3 Μετάφραση Βασισμένη Σε Παραδείγματα (Example-based Machine Translation - EBMT)	28
4.1.4 Υβριδικές Μέθοδοι Μετάφρασης.....	29
4.1.5 Μετάφραση Google.....	29
5. ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ RESTAURANT FINDER	31
6. ΕΠΕΚΤΑΣΗ ΤΟΥ RESTAURANT FINDER	34
6.1 Μεταφορά Της εφαρμογής Σε Εξυπηρετή	34
6.2 Προσθήκη Πληροφοριών Για Εστιατόρια Της Αθήνας.....	36
6.3 Αξιοποίηση Κριτικών Από Πολλές Γλώσσες.....	37
7.ΣΥΜΠΕΡΑΣΜΑΤΑ.....	40
ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ.....	41
ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ.....	44
ΑΝΑΦΟΡΕΣ.....	45

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1: Απεικόνιση της λειτουργίας ενός NAT δρομολογητή.....	21
Εικόνα 2: Τα κύρια στάδια της NLP.....	25
Εικόνα 3: Η πυραμίδα του Bernard Vauquois.....	27
Εικόνα 4: Η δομή του Restaurant Finder	33
Εικόνα 5: Στον NAT Address Table η πρώτη εγγραφή ορίζει ότι τα αφικνούμενα πακέτα στη θύρα 8080 του δρομολογητή θα προωθούνται στη θύρα 8080 του υπολογιστή με διεύθυνση υποδικτύου 192.168.1.5	35
Εικόνα 6: Restaurant Finder σε τοπικό υπολογιστή. Δεν είναι δυνατή η προσπέλαση του από το διαδίκτυο	35
Εικόνα 7: Μεταφορά του Restaurant Finder σε εξυπηρετή. Η προσπέλαση του μέσω του διαδικυου είναι εφικτή	35
Εικόνα 8: Η αρχική σελίδα της εφαρμογής με τα αποτελέσματα της αναζήτησης σε αλληνικά εστιατόρια	36
Εικόνα 9: Η σελίδα με τις πληροφορίες του εστιατορίου "Dinner in the Sky".....	37
Εικόνα 10: Αρχικό κείμενο στα ελληνικά	38
Εικόνα 11: Μεταφρασμένο κείμενο από τα ελληνικά στα αγγλικά.....	39

ΠΡΟΛΟΓΟΣ

Η πτυχιακή εργασία εκπονήθηκε στο τμήμα Πληροφορικής και Τηλεπικοινωνιών του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών κατά τη διάρκεια του ακαδημαϊκού έτους 2015-2016. Επιβλέπουσα καθηγήτρια ήταν η κυρία Ιζαμπώ Κανάλη, την οποία ευχαριστώ θερμά για το ενδιαφέρον και την καθοδήγηση της. Επίσης, ευχαριστώ θερμά την φοιτήτρια Δωροθέα Τσιμπίδη, η οποία μου προσέφερε την εφαρμογή που έγινε αντικείμενο της δικής μου πτυχιακής, τα αρχεία της και η οποία με βοήθησε όπου χρειάστηκε.

1. ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια στο διαδίκτυο γίνεται ευρεία χρήση εφαρμογών που ειδικεύονται στην πρόβλεψη των προτιμήσεων, των αναγκών και επιλογών του χρήστη. Για παράδειγμα, ένας ιστότοπος πώλησης βιβλίων μπορεί να χρησιμοποιεί ένα τέτοιο πρόγραμμα, ώστε το σύστημα να προτείνει στον χρήστη βιβλία που είναι πιθανό να του αρέσουν. Αυτού του είδους τα συστήματα ονομάζονται συστήματα συστάσεων (recommender systems).

Τα συστήματα συστάσεων έχουν όλα τον ίδιο σκοπό, την σύσταση αντικειμένων του εκάστοτε ιστοτόπου στον χρήστη, ωστόσο υπάρχουν πολλές διαφορετικές προσεγγίσεις σχετικά με τον τρόπο που το κάνουν αυτό. Ορισμένοι αλγόριθμοι βγάζουν αποτελέσματα βασιζόμενοι σε χρήστες με παρόμοια συμπεριφορά, άλλοι λαμβάνουν υπόψιν μόνο αντικείμενα που έχουν κάποια ομοιότητα με προηγούμενες επιλογές του χρήστη, ενώ άλλοι συνδιάζουν τις δύο προηγούμενες προσεγγίσεις. Γενικά, υπάρχουν αρκετά είδη αλγορίθμων για συστήματα συστάσεων. Το γεγονός αυτό δίνει αξία σε μια άλλη διαδικασία, την αξιολόγηση συστημάτων συστάσεων.

Σήμερα, ένας σχεδιαστής συστημάτων συστάσεων βρίσκεται στην ευχάριστη θέση να μπορεί να διαλέξει τον αλγόριθμο του συστήματός του από ένα μεγάλο σύνολο διαφορετικών προσεγγίσεων καθεμιά από τις οποίες έχει τη δική της λογική και τα δικά της πλεονεκτήματα και μειονεκτήματα. Ο σχεδιαστής σε αυτή την περίπτωση χρειάζεται ένα μέτρο σύγκρισης μεταξύ των διαφορετικών αλγορίθμων, έτσι ώστε να αναγνωρίσει ποιός αλγόριθμος παράγει καλύτερα αποτελέσματα στην εφαρμογή του. Επίσης, η αξιολόγηση του συστήματος είναι εξαιρετικά χρήσιμη για τη συντήρηση και βελτίωση του. Ο σχεδιαστής θέλει συνεχώς να βελτιώνει το σύστημά του, οπότε θα προσπαθεί να τροποποιεί τον αλγόριθμο που παράγει συστάσεις. Η διαδικασία αξιολόγησης θα του υποδείξει ποιές τροποποιήσεις βελτιώνουν πραγματικά το σύστημα. Όλα τα ανωτέρω καθιστούν φανερή την αξία της αξιολόγησης συστημάτων συστάσεων.

Στα πλαίσια της παρούσας πτυχιακής εργασίας πραγματοποιήθηκε θεωρητική μελέτη των συστημάτων συστάσεων και των μεθόδων αξιολόγησης τους. Επίσης, μελετήθηκαν το client-server μοντέλο εφαρμογών που βρίσκει εκτεταμένη χρήση στο διαδίκτυο, η ευρέως χρησιμοποιούμενη τεχνική της Μετάφρασης Διευθύνσεων Δικτύου (NAT) για την επικοινωνία τερματικών μέσω του διαδικτύου. Τέλος, έγινε ενασχόληση με την υποπεριοχή του κλάδου της Τεχνητής Νοημοσύνης που ονομάζεται Επεξεργασία Φυσικής Γλώσσας. Όλα τα παραπάνω εφαρμόστηκαν στο σύστημα συστάσεων

Restaurant Finder με σκοπό την επέκτασή του.

Η πτυχιακή αποτελείται από επτά κεφάλαια. Στο επόμενο κεφάλαιο παρουσιάζονται γενικές αρχές των συστημάτων συστάσεων, οι διάφορες μεθοδολογίες, καθώς και οι γενικές αρχές της αξιολόγησης τους. Στο τρίτο κεφάλαιο γίνεται μια παρουσίαση των δύο επικρατούντων αρχιτεκτονικών για διαδικτυακές εφαρμογές, της αρχιτεκτονικής πελάτη-εξυπηρετή (client-server) και της αρχιτεκτονικής ομοτίμων (peer-to-peer), καθώς και η ανάλυση του πρωτοκόλλου Μετάφρασης Διευθύνσεων Δικτύου (Network Address Translation – NAT). Το τέταρτο κεφάλαιο αποτελεί μια παρουσίαση της Επεξεργασίας Φυσικής Γλώσσας. Το πέμπτο κεφάλαιο αποτελεί μια παρουσίαση της δομής του Restaurant Finder [1]. Το έκτο περιγράφει αναλυτικά τις επεκτάσεις που πραγματοποιήθηκαν στο Restaurant Finder. Τέλος, το έβδομο κεφάλαιο παραθέτει τα συμπεράσματα που προέκυψαν από την πτυχιακή εργασία.

2. ΣΥΣΤΗΜΑΤΑ ΣΥΣΤΑΣΕΩΝ

Ο όγκος πληροφοριών που υπάρχει στο διαδίκτυο είναι εξαιρετικά μεγάλος. Κατά συνέπεια, ένας χρήστης που θέλει να αναζητήσει κάτι έρχεται αντιμέτωπος με τεράστια ποσότητα πληροφορίας, μέσα στην οποία είναι δύσκολο να βρει αυτό που ψάχνει και να ξεχωρίσει τις πληροφορίες που του χρειάζονται από τις υπόλοιπες. Στο διαδίκτυο υπάρχει, δηλαδή, το πρόβλημα της υπερπληροφόρησης, το οποίο είναι ένας από τους λόγους για τον οποίο δημιουργήθηκαν τα συστήματα συστάσεων (recommender systems).

Τα συστήματα συστάσεων είναι προσωπικοί για κάθε χρήστη πράκτορες πληροφόρησης (information agents [2]), οι οποίοι προτείνουν στον χρήστη αντικείμενα που πιθανόν να τον ενδιαφέρουν. Για παράδειγμα, σε ένα σύστημα ταινιών το σύστημα συστάσεων θα προτείνει στον χρήστη ταινίες για τις οποίες υπάρχει πιθανότητα να του αρέσουν. Γνωστές εφαρμογές που αξιοποιούν συστήματα συστάσεων είναι το Netflix [3] για ταινίες, τα κοινωνικά δίκτυα Facebook [4], Twitter [5] για να προτείνουν νέους φίλους και ομάδες, το GroupLens [6] για άρθρα.

Παρακάτω παρουσιάζονται διαφορετικές προσεγγίσεις αλγορίθμων για συστήματα συστάσεων.

2.1 Τύποι Συστημάτων Συστάσεων [1] [7]

Ο διαχωρισμός των συστημάτων συστάσεων αφορά τον τρόπο με τον οποίο αυτά παράγουν αποτελέσματα, δηλαδή τους αλγορίθμους που χρησιμοποιεί το σύστημα ώστε να αποφασίσει ποιές είναι οι καλύτερες συστάσεις για τον χρήστη. Οι αλγόριθμοι που χρησιμοποιούνται για την παραγωγή συστάσεων ονομάζονται αλγόριθμοι φιλτραρίσματος και κατηγοριοποιούνται ως εξής:

- Συνεργατικό Φιλτράρισμα (Collaborative Filtering - CF)
- Φιλτράρισμα βασισμένο στο Περιεχόμενο (Content-Based Filtering)
- Φιλτράρισμα βασισμένο στη Γνώση (Knowledge-based Filtering - KB)
- Φιλτράρισμα βασισμένο στα Δημογραφικά Στοιχεία (Demographic Filtering - DF)
- Υβριδικό φιλτράρισμα (Hybrid filtering)

Αξιόλογες πηγές είναι τα άρθρα των J. Bobadilla, F. Ortega, A. Hernando, A. Gutiérrez [8], των Linyuan Lua, Matus Medob, Chi Ho Yeungb, Yi-Cheng Zhanga, Zi-Ke Zhanga, Tao Zhou [9], το άρθρο “A Review and Classification of Recommender Systems Research” από τους Deuk Hee Park, Hyea

Kyeong Kim, Il Young Choi, Jae Kyeong Kim [10], καθώς και το “Introduction to Recommender Systems Handbook” από τους Francesco Ricci, Lior Rokach και Bracha Shapira [11]. Επίσης, το βιβλίο των Prem Melville και Vikas Sindhwani [12].

2.1.1 Συνεργατικό Φιλτράρισμα (Collaborative Filtering – CF) [1] [13] [14]

Οι αλγόριθμοι συνεργατικού φιλτραρίσματος βασίζονται στην υπόθεση ότι οι χρήστες που έχουν εμφανίσει παρόμοια συμπεριφορά και έχουν κάνει παρόμοιες επιλογές θα ξαναέχουν παρόμοια συμπεριφορά ή επιλογές στο μέλλον. Οι αλγόριθμοι αυτοί κάνουν συλλογή και ανάλυση δεδομένων σχετικά με τη δραστηριότητα του χρήστη και βασίζουν τις συστάσεις τους στις επιλογές άλλων χρηστών με παρόμοια συμπεριφορά. Η συλλογή δεδομένων μπορεί να γίνει είτε ρητά (explicitly) είτε σιωπηρά (implicitly).

Η συλλογή δεδομένων ρητά μπορεί να γίνει με έναν από τους παρακάτω τρόπους:

- Προτρέποντας τον χρήστη να βαθμολογήσει ένα αντικείμενο.
- Προτρέποντας τον χρήστη να ταξινομήσει κατά σειρά προτίμησης ένα σύνολο από αντικείμενα.
- Παρουσιάζοντας στον χρήστη δύο αντικείμενα και ρωτώντας τον ποιο προτιμάει.
- Προτρέποντας τον χρήστη να δημιουργήσει ένα σύνολο από αντικείμενα που του αρέσουν.

Η σιωπηρή συλλογή δεδομένων γίνεται με τους παρακάτω τρόπους:

- Παρατηρώντας τα αντικείμενα που επισκέπτεται χρήστης όταν είναι συνδεδεμένος στον ιστότοπο.
- Αναλύοντας πόσες φορές έχει επισκεφτεί ο χρήστης ένα συγκεκριμένο αντικείμενο.
- Καταγράφοντας τα αντικείμενα που αγόρασε ο χρήστης.
- Παρατηρώντας τον κοινωνικό περίγυρο του χρήστη ώστε να βρεθούν παρόμοιες συμπεριφορές.

Οι αλγόριθμοι συνεργατικού φιλτραρίσματος παράγουν αρκετά καλές συστάσεις όταν έχουν μεγάλο όγκο πληροφορίας για τους χρήστες και μάλιστα όσο περισσότερη πληροφορία έχουν τόσο πιο ποιοτικές θα είναι οι συστάσεις τους. Ωστόσο, έχουν τα παρακάτω μειονεκτήματα:

- Το πρόβλημα της αργής εκκίνησης (cold start) ουσιαστικά σημαίνει ότι οι συνεργατικοί αλγόριθμοι δεν έχουν αρκετή πληροφορία στην αρχή της λειτουργίας ενός συστήματος, οπότε δεν μπορούν να παράξουν

αξιόπιστα αποτελέσματα.

- Η κλιμάκωση (scalability) είναι ένα ακόμη πρόβλημα. Ο τεράστιος όγκος δεδομένων που απαιτούν αυτού του είδους οι αλγόριθμοι για να λειτουργήσουν αποτελεσματικά απαιτεί συνήθως μεγάλη υπολογιστική ισχύ.
- Ένα ακόμη πρόβλημα είναι η αραιότητα (sparsity) [15]. Ακόμη και οι πιο ενεργοί χρήστες ενός συστήματος συνήθως βαθμολογούν μόνο ένα μικρό υποσύνολο από τα αντικείμενα που υπάρχουν στο σύστημα. Κατά συνέπεια, οι πληροφορίες που θα συγκεντρωθούν πιθανότατα να μην είναι αρκετές, ώστε ο αλγόριθμος να λειτουργήσει αποτελεσματικά.
- Μια μερίδα χρηστών των οποίων οι προτιμήσεις είναι πολύ διαφορετικές από αυτές των υπολοίπων, δεν επωφελούν τους αλγορίθμους συνεργατικού φιλτραρίσματος (πρόβλημα του γκρίζου προβάτου [16]).

Όσον αφορά κάποια γνωστά συστήματα συστάσεων, το amazon.com χρησιμοποιεί αλγορίθμους συνεργατικού φιλτραρίσματος για να προτείνει προϊόντα στους χρήστες. Επίσης, τα κοινωνικά δίκτυα Facebook [4], LinkedIn [17] χρησιμοποιούν συνεργατικό φιλτράρισμα για να προτείνουν στους χρήστες νέους φίλους ή ομάδες.

2.1.2 Φιλτράρισμα βασισμένο στο Περιεχόμενο (Content-Based Filtering – CBF) [1] [18] [14]

Το φιλτράρισμα με βάση το περιεχόμενο είναι μία προσέγγιση η οποία έχει χρησιμοποιηθεί εκτεταμένα. Οι έννοιες του προφίλ χρήστη και του αντικειμένου παίζουν κυρίαρχο ρόλο σε αυτούς τους αλγορίθμους. Για κάθε χρήστη διατηρείται ένα προφίλ με τις επιλογές που έχει κάνει - αντικείμενα για τα οποία εκδήλωσε ενδιαφέρον - και ο αλγόριθμος προτείνει εκείνα τα αντικείμενα τα οποία μοιάζουν με τις προηγούμενες επιλογές του χρήστη.

Ένα πολύ σημαντικό ζήτημα στους content-based αλγορίθμους είναι κατά πόσον ο αλγόριθμος έχει τη δυνατότητα να “μάθει” μέσα από τις επιλογές του χρήστη, έτσι ώστε να είναι ικανός να του προτείνει αντικείμενα διαφορετικού τύπου από αυτά για τα οποία ένας χρήστης έχει ήδη εκδηλώσει ενδιαφέρον. Για παράδειγμα, ένα πραγματικό σύστημα συστάσεων για ταινίες και μουσική θα είχε πολύ μεγάλη αξία αν δεδομένων κάποιων ταινιών που αρέσουν στον χρήστη θα μπορούσε να προτείνει και μουσική, αντί μόνο ταινίες.

Τα πλεονεκτήματα του φιλτραρίσματος με βάση το περιεχόμενο είναι τα ακόλουθα:

- Οι αλγόριθμοι που παράγουν τις συστάσεις εξετάζουν αποκλειστικά τις προτιμήσεις του χρήστη και αγνοούν εντελώς την συμπεριφορά των υπόλοιπων χρηστών. Ως εκ τούτου, η παραγωγή συστάσεων γίνεται σχετικά εύκολη υποθεση.
- Οι αλγόριθμοι αυτοί μπορούν πολύ εύκολα να δικαιολογήσουν τις συστάσεις τους.
- Οι content-based αλγόριθμοι μπορούν να προτείνουν αντικείμενα για τα οποία δεν έχει γίνει καμία κριτική ή βαθμολόγηση.

Σημαντικό μειονέκτημα των content-based αλγορίθμων είναι η αδυναμία τους να παράγουν συστάσεις όταν δεν έχουν καταγραφεί οι προτιμήσεις χρηστών στη βάση καθώς και το πρόβλημα γκρίζου προβάτου [16]. Αυτό συνήθως συμβαίνει στο ξεκίνημα της λειτουργίας του συστήματος (cold start) .

Μερικά συστήματα συστάσεων που βασίζονται στο περιεχόμενο είναι τα Rotten Tomatoes [19], Internet Movie Database [20], Jinni [21], Rovi Corporation [22], Jama [23], pandora radio [24]. Μια εξαιρετική πηγή για το φιλτράρισμα με βάση το περιεχόμενο αποτελεί το paper των Robin van Meteren και Maarten van Someren [18].

2.1.3 Φιλτράρισμα βασισμένο στη Γνώση (Knowledge-based Filtering – KBF) [25]

Το φιλτράρισμα με βάση τη γνώση αποτελεί μία ακόμα διαφορετική προσέγγιση των συστημάτων συστάσεων. Τέτοιου είδους αλγόριθμοι χρησιμοποιούνται συνήθως όταν δεν είναι δυνατό να χρησιμοποιηθούν rating-based αλγόριθμοι, δηλαδή collaborative και content-based. Οι knowledge-based αλγόριθμοι εξαγουν συμπεράσματα από τη βάση δεδομένων τους λαμβάνοντας υπόψιν τους περιορισμούς που θέτει ο χρήστης.

Οι αλγόριθμοι φιλτραρίσματος με βάση τη γνώση έχουν τα εξής πλεονεκτήματα:

- Δεν υποφέρουν από το πρόβλημα αργής εκκίνησης (cold start).
- Η απόδοσή τους δεν εξαρτάται από τον αριθμό των χρηστών ούτε από βαθμολογίες.
- Μπορούν εύκολα να δικαιολογήσουν τις συστάσεις τους.

Ωστόσο, αυτού του είδους οι αλγόριθμοι συνοδεύονται από σημαντικά μειονεκτήματα. Συγκεκριμένα:

- Πάσχουν από το πρόβλημα απόκτησης γνώσης, καθώς δεν παρακολουθούν τη συμπεριφορά των χρηστών, αλλά πρέπει κάποιος

άνθρωπος (πιθανώς προγραμματιστής) να εισάγει τα δεδομένα-γνώση στο σύστημα.

- Υπάρχει η ανάγκη τα δεδομένα να ανανεώνονται συνεχώς, ώστε το σύστημα να μπορεί να παράγει έγκυρες συστάσεις.

Η διατριβή του R. Burke επάνω στο θέμα είναι πολύ καλή [25].

2.1.4 Φιλτράρισμα βασισμένο στα Δημογραφικά Στοιχεία (Demographic Filtering – DF) [1] [14]

Η συγκεκριμένη προσέγγιση βασίζεται στην υπόθεση ότι άτομα με ίδια ή παρόμοια δημογραφικά χαρακτηριστικά - φύλο, ηλικία, επάγγελμα κλπ - θα έχουν και ίδιες προτιμήσεις. Οι συστάσεις γίνονται με βάση τα χαρακτηριστικά αυτά, τα οποία συνήθως κάθε χρήστης τα εισάγει στο σύστημα κατά τη δημιουργία λογαριασμού. Τα χαρακτηριστικά κάθε χρήστη συγκεντρώνονται στο προφίλ του. Το φιλτράρισμα με βάση δημογραφικά στοιχεία μπορεί να αποτελέσει καλή λύση στο πρόβλημα της αργής εκκίνησης.

Το κύριο πελόνεκτημα του demographic-filtering είναι ότι δεν εξαρτάται από τη συμπεριφορά των χρηστών και από βαθμολογίες, οπότε η εφαρμογή τους δεν κινδυνεύει από ανεπάρκεια δεδομένων και αργή εκκίνηση.

Τα μειονεκτήματα του είναι:

- Προκύπτουν ζητήματα ιδιωτικότητας, λόγω των προσωπικών στοιχείων που εισάγονται στο σύστημα.
- Τα δημογραφικά στοιχεία πολλές φορές δεν αρκούν για να γίνουν αποτελεσματικές συστάσεις.
- Υποφέρει από το πρόβλημα του γκρίζου προβάτου [16].

2.1.5 Υβριδικό φιλτράρισμα (Hybrid filtering) [26]

Πρόκειται για κάποιου είδους συνδυασμό των παραπάνω μεθόδων. Συνήθως οι υβριδικοί αλγόριθμοι είναι πιο αποτελεσματικοί από τους απλούς, καθώς οι τεχνικές που αναφέρθηκαν προηγουμένως συνδυάζονται με τρόπο ώστε η μία να καλύπτει τις αδυναμίες της άλλης. Υπάρχουν πολλοί τρόποι με τους οποίους μπορεί να γίνει υβριδοποίηση (hybridization). Κάποιοι από αυτούς είναι :

- **Σταθμισμένη Υβριδοποίηση (Weighted Filtering):** Τα αποτελέσματα που παράγουν διαφορετικές τεχνικές φιλτραρίσματος συνδυάζονται μεταξύ τους, με διαφορετικό βάρος η κάθε μια.
- **Επιλεκτική Υβριδοποίηση (Switching Hybridization):** Το σύστημα

επιλέγει κάθε φορά ανάμεσα σε ένα σύνολο από διαφορετικές τεχνικές φιλτραρίσματος για το ποιά θα χρησιμοποιήσει.

- **Μεικτή Υβριδοποίηση (Mixed Hybridization):** Τα αποτελέσματα διαφορετικών τεχνικών συναθροίζονται.
- **Συνδυασμός Χαρακτηριστικών (Feature Combination) :** Χαρακτηριστικά διαφορετικών τεχνικών φιλτραρίσματος συνδυάζονται, ώστε να δημιουργήσουν έναν νέο αλγόριθμο φιλτραρίσματος.
- **Αύξηση χαρακτηριστικών (Feature Augmentation):** Μια τεχνική συστάσεων χρησιμοποιείται προκειμένου να υπολογίσει ένα χαρακτηριστικό ή μια σειρά από χαρακτηριστικά τα οποία εφαρμόζονται ως είσοδο σε επόμενη τεχνική.
- **Αλληλουχία υβριδοποίησης (Cascade Hybridization):** Χρησιμοποιούνται πολλές διαφορετικές τεχνικές φιλτραρίσματος τηρώντας μια σειρά προτεραιότητας. Όταν μια τεχνική δεν βγάζει ξεκάθαρο αποτέλεσμα τότε χρησιμοποιείται η αμέσως επόμενη στην ιεραρχία τεχνική επάνω στο αποτέλεσμα της προηγούμενης.
- **Υβριδοποίηση μεταεπιπέδου (Meta-level hybridization):** Μια τεχνική σύστασης εφαρμόζεται και παράγει κάποιο μοντέλο το οποίο χρησιμοποιείται ως είσοδο σε κάποια άλλη τεχνική.

Το πολύ γνωστό recommendation-system Netflix χρησιμοποιεί υβριδικό φιλτράρισμα. Συνδυάζει συνεργατικό φιλτράρισμα αναλύοντας τις συνήθειες “παρόμοιων” χρηστών και συγχρόνως χρησιμοποιεί φιλτράρισμα ως προς το περιεχόμενο για να προτείνει στον χρήστη ταινίες παρόμοιες με αυτές που ο χρήστης είδε και βαθμολόγησε καλά [27].

2.2 Αξιολόγηση Συστημάτων Συστάσεων [28]

Την τελευταία δεκαετία έχει αφιερωθεί μεγάλη ποσότητα έρευνας στην ανάπτυξη νέων μεθοδολογιών για συστήματα συστάσεων. Πλέον, ένας σχεδιαστής που θέλει να προσθέσει σύστημα συστάσεων στην εφαρμογή του θα πρέπει να διαλέξει ανάμεσα σε πάρα πολλούς αλγορίθμους και να επιλέξει τον πιο ταιριαστό. Για το σκοπό αυτό είναι χρήσιμες κάποιες μέθοδοι αξιολόγησης συστημάτων συστάσεων, ώστε να γίνεται σαφές ποιός αλγόριθμος παράγει ποιοτικότερα αποτελέσματα.

Υπάρχουν πολλά χαρακτηριστικά ως προς τα οποία μπορεί να γίνει σύγκριση αλγορίθμων για συστήματα συστάσεων. Το πιο σημαντικό από αυτά είναι η ακρίβεια των αποτελεσμάτων, δηλαδή πόσο κοντινές είναι οι συστάσεις σε σχέση με τις πραγματικές επιλογές των χρηστών.

Η πιο γνωστή μετρική ακρίβειας αποτελεσμάτων είναι η Τετραγωνική Ρίζα Μέσου Τετραγώνου του Σφάλματος (Root Square Mean Error – RSME). Το σύστημα παράγει τις προβλεπόμενες βαθμολογίες (έστω σύνολο $\hat{r}_{ui}$) για ένα δεδομένο σύνολο χρηστών-αντικειμένων $\mathcal{T}$ και για το οποίο είναι γνωστές οι πραγματικές βαθμολογίες (έστω το σύνολο r_{ui}). Η τιμή RMSE υπολογίζεται από τον τύπο

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} (\hat{r}_{ui} - r_{ui})^2}$$

Μια εναλλακτική της RMSE είναι η Μέση Απόλυτη Τιμή Σφάλματος (Mean Absolute Error – MAE), η οποία δίνεται από τον τύπο

$$\text{MAE} = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(u,i) \in \mathcal{T}} |\hat{r}_{ui} - r_{ui}|}$$

Σε σχέση με την MAE, η RMSE δίνει έμφαση στα μεγάλα λάθη. Αυτό σημαίνει ότι αν σε ένα σύνολο από τέσσερα αντικείμενα έχουμε έναν αλγόριθμο που κάνει σφάλμα μεγέθους 2 στα τρία από τα τέσσερα και μηδενικό σφάλμα στο τέταρτο και έναν άλλον αλγόριθμο ο οποίος κάνει σφάλμα μεγέθους 3 μόνο στο ένα αντικείμενο και μηδενικό στα τρία υπόλοιπα, τότε η RMSE προτιμάει τον πρώτο αλγόριθμο και η MAE τον δεύτερο. Η επιλογή του κατάλληλου αλγορίθμου έχει να κάνει με το τι είδους αλγόριθμο επιθυμεί ο σχεδιαστής και το πόσο ποινικοποιεί τα σφάλματα.

Υπάρχουν παραλλαγές των δύο προαναφερθέντων μετρικών. Η κανονικοποιημένη RMSE (Normalized RMSE-NRMSE) και η κανονικοποιημένη MAE (Normalized MAE-NMAE) είναι κανονικοποιήσεις των RMSE και MAE αντίστοιχα και παράγουν ακριβώς τα ίδια αποτελέσματα. Η μέση RMSE (Average RMSE – ARMSE) και η μέση MAE (Average MAE – AMAE) είναι κατάλληλες για σύνολα αντικειμένων τα οποία δεν έχουν ισορροπημένη κατανομή.

Αν και η ακρίβεια των αποτελεσμάτων είναι ο πιο σημαντικός παράγοντας είναι κοινά αποδεκτό ότι από μόνη της δεν αρκεί. Υπάρχουν πολλοί ακόμη παράγοντες που θα πρέπει να λαμβάνονται υπόψιν κατά τη σχεδίαση ενός συστήματος συστάσεων. Μερικοί από αυτούς είναι :

- **Κάλυψη (Coverage)** – Η κάλυψη μπορεί να μετρηθεί ως προς τα αντικείμενα που προτείνονται, αλλά και ως προς τους χρήστες για τους οποίους μπορούν να παραχθούν αποτελέσματα. Στην πρώτη περίπτωση όσο περισσότερα είναι τα αντικείμενα τα οποία το σύστημα έχει τη δυνατότητα να προτείνει τόσο μεγαλύτερη είναι η κάλυψη αντικειμένων. Στη δεύτερη περίπτωση είναι επιθυμητό να το σύστημα να μπορεί να παράξει ακριβείς συστάσεις για μεγάλο εύρος χρηστών. Ένας παράγοντας που περιορίζει το εύρος χρηστών θα μπορούσε να είναι ο αριθμός των κριτικών που έχει κάνει ο χρήστης, δηλαδή να μην μπορούν να παραχθούν συστάσεις αν ο χρήστης δεν έχει πραγματοποιήσει έναν συγκεκριμένο αριθμό από κριτικές. Συνήθως, πρέπει να γίνει ένας συμβιβασμός ανάμεσα στο εύρος χρηστών και στην ακρίβεια των συστάσεων.
- **Αυτοπεποίθηση (Confidence)** – Ο παράγοντας αυτός αφορά το πόσο σίγουρο είναι το σύστημα για τις επιλογές που κάνει. Όταν ο χρήστης βλέπει ότι το σύστημα προτείνει ένα αντικείμενο με πολύ μεγάλη αυτοπεποίθηση και ένα άλλο με μικρή αυτοπεποίθηση τότε μάλλον θα επιλέξει το πρώτο. Η πιο κοινή μετρική για την αυτοπεποίθηση του συστήματος είναι η πιθανότητα μιά σύσταση να είναι σωστή.
- **Εμπιστοσύνη (Trust)** – Ο παράγοντας αυτός υποδεικνύει την εμπιστοσύνη που έχει ο χρήστης για το σύστημα. Όταν ένα σύστημα παράγει συστάσεις οι οποίες αρέσουν στον χρήστη αυτός αποκτά εμπιστοσύνη προς το σύστημα.
- **Καινοτομία (Novelty)** – Η καινοτομία αφορά την παραγωγή συστάσεων για αντικείμενα που ο χρήστης δεν γνωρίζει.
- **Serendipity** – Ο όρος serendipity αφορά το πόσο έκπληξη προκαλούν οι επιτυχημένες συστάσεις στον χρήστη. Για παράδειγμα, αν ένας χρήστης έχει βαθμολογήσει πολύ καλά μια ταινία με έναν συγκεκριμένο ηθοποιό και το σύστημα του προτείνει μια άλλη ταινία του ίδιου ηθοποιού τότε ικανοποιείται η καινοτομία (Novelty) αλλά όχι το serendipity, καθώς η σύσταση είναι γενικά αναμενόμενη. Για συστάσεις με υψηλή serendipity υπάρχει ο κίνδυνος να είναι ανακριβείς.
- **Ποικιλία (Diversity)** – Είναι η δυνατότητα του συστήματος να προτείνει αντικείμενα διαφορετικού τύπου. Γενικά όσο πιο ευρύ είναι το φάσμα των αντικειμένων που προτείνονται τόσο πιο ευχαριστημένος είναι ο χρήστης.
- **Χρησιμότητα (Utility)** – Η χρησιμότητα σχετίζεται με την παραγωγή συστάσεων που είναι χρήσιμες για τον χρήστη ή για το σύστημα. Σε ένα σύστημα συστάσεων ταινιών όπου οι χρήστες βαθμολογούν τις συστάσεις, μπορεί να θεωρηθεί η βαθμολογία του χρήστη ως μέτρο για την χρησιμότητα της σύστασης.

- **Ρίσκο (Risk)** – Όλες οι συστάσεις εμπεριέχουν ρίσκο. Το ρίσκο των συστάσεων μπορεί να θεωρηθεί ότι είναι το εύρος της αναμενόμενης χρησιμότητας (utility variance). Μερικά συστήματα μπορεί να επιθυμούν χαμηλό ρίσκο στις συστάσεις τους, ενώ μερικά άλλα υψηλό.
- **Ευρωστία (Robustness)** – Αφορά τη δυνατότητα του συστήματος να αναγνωρίζει πότε οι χρήστες προσπαθούν να το “ξεγελάσουν” κάνοντας ψευδείς κριτικές, ώστε να αλλοιώσουν τα αποτελέσματα των συστάσεων.
- **Ιδιωτικότητα (Privacy)** – Συνήθως οι χρήστες δίνουν τα προσωπικά τους στοιχεία και τις προτιμήσεις τους στο σύστημα με την προοπτική να πάρουν ποιοτικές συστάσεις. Η ιδιωτικότητα έχει να κάνει με το γεγονός τα στοιχεία αυτά να παραμείνουν κρυφά, ώστε να μην μπορεί κάποιος τρίτος να αποσπάσει τα προσωπικά δεδομένα και πληροφορίες για τη συμπεριφορά των χρηστών.
- **Προσαρμοστικότητα (Adaptivity)** – Η προσαρμοστικότητα αφορά τη δυνατότητα του συστήματος να προσαρμόζει τις συστάσεις του στην αλλαγή ενδιαφερόντων του χρήστη. Για παράδειγμα, σε μια εφαρμογή ειδήσεων μπορεί κάποια άρθρα να έχουν γενικά πολύ μικρό ενδιαφέρον, το οποίο όμως κάποια περίοδο να αυξηθεί κατακόρυφα λόγω ενός γεγονότος. Τέτοιες μεταβολές ενδιαφέροντος πρέπει να γίνονται αντιληπτές από το σύστημα. Επίσης, αλλαγή ενδιαφερόντων από την πλευρά του χρήστη μπορεί να υπάρχει όταν αυτός αλλάξει τις πληροφορίες στο προφίλ του. Το σύστημα πρέπει να έχει τη δυνατότητα να προσαρμόζεται και σε αυτού του είδους τις αλλαγές.
- **Κλιμάκωση (Scalability)** – Η κλιμάκωση υποδεικνύει την ικανότητα του συστήματος να παράγει αποτελέσματα στον ίδιο περίπου χρόνο όταν το μέγεθος των δεδομένων αυξάνεται. Οι χρήστες απαιτούν γενικά οι συστάσεις να παράγονται γρήγορα, όμως όσο περισσότερα δεδομένα υπάρχουν στο σύστημα τόσο περισσότεροι υπολογισμοί χρειάζονται. Ένα σύστημα με καλή κλιμάκωση δεν θα πρέπει να αργεί όταν ο όγκος δεδομένων είναι πολύ μεγάλος και ούτε, επίσης, να περιορίζει την ακρίβεια και ποιότητα των αποτελεσμάτων του, ώστε να πετύχει καλύτερο χρόνο.

3. ΑΡΧΕΣ ΔΙΑΔΙΚΤΥΟΥ

3.1 Αρχιτεκτονικές Web Εφαρμογών

Ο όρος διαδικτυακή (web) εφαρμογή αναφέρεται σε εφαρμογές, δηλαδή προγράμματα, που τρέχουν επάνω στο δίκτυο. Στις μέρες μας υπάρχουν χιλιάδες ίσως και εκατομμύρια τέτοιες εφαρμογές. Μπορούμε να διαχωρίσουμε τις διαδικτυακές εφαρμογές σε δύο κατηγορίες, με βάση τον τρόπο με τον οποίο δομούνται επάνω στα τερματικά συστήματα.

Η πρώτη κατηγορία είναι το μοντέλο πελάτη-εξυπηρετή (client-server) [29]. Το μοντέλο αυτό είναι μια κεντριοποιημένη προσέγγιση κατά την οποία υπάρχει ένας συγκεκριμένος υπολογιστής υπηρεσίας (εξυπηρετή ή server), ο οποίος είναι πάντα ενεργός και εξυπηρετεί αιτήσεις για υπηρεσίες άλλων χρηστών-πελατών. Οι πελάτες δεν είναι απαραίτητα πάντα ενεργοί. Οι εξυπηρετές έχουν συνήθως σταθερή και γνωστή διεύθυνση, ώστε να μπορούν οι πελάτες να επικοινωνήσουν μαζί τους. Επίσης, έχουν εγκατεστημένο ειδικό λογισμικό για αυτή τη δουλειά. Στις client-server εφαρμογές όταν οι πελάτες θέλουν να επικοινωνήσουν μεταξύ τους επικοινωνούν μέσω του εξυπηρετή.

Συχνά είναι δύσκολο ένας εξυπηρετής να ικανοποιήσει όλες τις αιτήσεις, για αυτό το λόγο συνηθίζεται να χρησιμοποιείται ένα σύμπλεγμα υπολογιστών που ονομάζεται φάρμα εξυπηρετών (server farm), οι οποίοι δημιουργούν έναν εικονικό εξυπηρετή. Με αυτό το τέχνασμα αυξάνεται ο αριθμός των αιτήσεων που μπορούν να ικανοποιηθούν.

Μερικές πολύ γνωστές εφαρμογές με αρχιτεκτονική πελάτη-εξυπηρετή είναι ο παγκόσμιος ιστός (World Wide Web – www) [30], η εφαρμογή μετάδοσης αρχείων επάνω στο δίκτυο (FTP) [31] και ηλεκτρονική αλληλογραφία (e-mail) [32].

Η άλλη πολύ διαδεδομένη αρχιτεκτονική ονομάζεται αρχιτεκτονική ομοτίμων (peer-to-peer ή p2p) [33]. Η peer-to-peer είναι μια μη κεντριοποιημένη προσέγγιση, δηλαδή δεν υπάρχει κάποιος συγκεκριμένος υπολογιστής που είναι πάντα ενεργός και στον οποίο να απευθύνονται όλες οι αιτήσεις. Το γεγονός αυτό υποδεικνύει ότι οι ομότιμοι έχουν απευθείας επικοινωνία μεταξύ τους. Οι ομότιμοι συνήθως είναι απλοί υπολογιστές χρηστών σε σπίτια, πανεπιστήμια και γραφεία.

Το πλεοέκτημα της αρχιτεκτονικής ομοτίμων έναντι της πελάτη-εξυπηρετή είναι ότι αυτο-κλιμακώνεται (self-scalable). Για παράδειγμα, σε μια εφαρμογή διαμοιρασμού αρχείων κάθε ομότιμος στέλνει αιτήσεις προς τους άλλους

ομότιμους για τα πακέτα δεδομένων, όμως παίζει και το ρόλο του εξυπηρετή προς όλους τους υπόλοιπους ομότιμους για τα αρχεία που έχει λάβει. Δηλαδή, κατ' αντιστοιχία με την αριτεκτονική client-server, κάθε ομότιμος παίζει το ρόλο και του πελάτη και του εξυπηρετή. Η peer-to-peer αρχιτεκτονική έχει μικρό κόστος καθώς δεν απαιτείται ισχυρό λογισμικό, υλικό και γρήγορη προσβαση στο δίκτυο.

Γνωστές διαδικτυακές εφαρμογές με p2p αρχιτεκτονική είναι το BitTorrent [34], το LimeWire [35], και το Skype [36].

Πολλές εφαρμογές χρησιμοποιούν υβριδική αρχιτεκτονική. Για παράδειγμα, σε εφαρμογές ανταλλαγής μηνυμάτων χρησιμοποιούνται εξυπηρετές ώστε να αποθηκεύονται οι πληροφορίες των χρηστών (όπως η διεύθυνση IP), αλλά τα μηνύματα από χρήστη σε χρήστη στέλνονται απευθείας από τον έναν στον άλλον χωρίς τη μεσολάβηση κάποιου ενδιάμεσου εξυπηρετή.

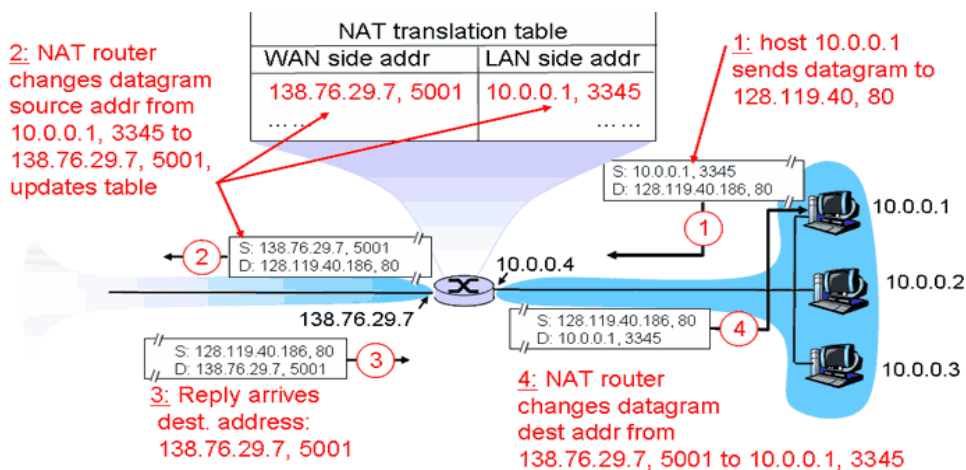
3.2 Μεταφραστές Διευθύνσεων Δικτύου

Η ταχύτατη εξάπλωση του διαδικτύου από το 1990 και μετά καθώς και η συνεχής σύνδεση νέων συσκευών σε αυτό κατέστησε σαφές, ότι αργά ή γρήγορα οι διαθέσιμες IPv4 [37] διευθύνσεις δεν θα ήταν αρκετές ώστε να καλύψουν όλες τις συσκευές. Η λύση δόθηκε μέσω της Μετάφρασης Διευθύνσεων Δικτύου (Network Address Translation) [38].

Πλέον δυο διαφορετικοί υπολογιστές στο διαδίκτυο μπορούν να έχουν την ίδια IP διεύθυνση, αρκεί να βρίσκονται πίσω από διαφορετικό δρομολογητή. Οι δρομολογητές με δυνατότητες NAT ορίζουν αυθαίρετες διευθύνσεις στους υπολογιστές του υποδικτύου τους (ιδιωτικές IP ή private IP) οι οποίοι επικοινωνούν μεταξύ τους με αυτές τις IP. Επίσης, αποθηκεύουν αντιστοιχίες (πίνακας μετάφρασης) ανάμεσα στις διευθύνσεις του υποδικτύου που ορίζουν και σε θύρες προορισμού, έτσι ώστε όταν δεχτούν ένα πακέτο να ελέγχουν τη θύρα προορισμού του και να το προωθούν στον κατάλληλο υπολογιστή.

Για παράδειγμα, ας υποθέσουμε ότι σε ένα οικιακό δίκτυο ο υπολογιστής με διεύθυνση υποδικτύου 10.0.0.1 (private IP) ζητά μια ιστοσελίδα που βρίσκεται σε κάποιον εξυπηρετή Web (θύρα 80) με διεύθυνση IP 128.119.30.186. Ο υπολογιστής 10.0.0.1 εκχωρεί αυθαίρετα τον αριθμό θύρας προέλευσης 3345 και στέλνει το δεδομένογράμμα μέσα στο LAN. Ο δρομολογητής NAT δέχεται το δεδομένογράμμα, αντικαθιστά τη διεύθυνση προέλευσης IP με τη διεύθυνση IP στη μεριά του WAN 138.76.29.7 και αντικαθιστά τον αρχικό αριθμό θύρας προέλευσης 3345 με τον αριθμό 5001. Το NAT στον

δρομολογητή προσθέτει επίσης μια νέα καταχώρηση στον πίνακα μετάφρασης του. Ο εξυπηρέτης Web, ο οποίος δεν γνωρίζει ότι η διεύθυνση και η θύρα προέλευσης του αφικνούμενου δεδομενογράμματος έχουν τροποποιηθεί από τον δρομολογητή NAT, απαντά στην αίτηση HTTP με ένα νέο δεδομένογράμμο με διεύθυνση προορισμού τη διεύθυνση του δρομολογητή NAT και θύρα προορισμού 5001. Όταν αυτό το δεδομένογράμμο φτάσει στον δρομολογητή NAT, ο δρομολογητής ταξινομεί τον πίνακα μετάφρασης NAT χρησιμοποιώντας τη διεύθυνση IP προορισμού ώστε να πάρει την κατάλληλη διεύθυνση προορισμού (10.0.0.1) και τον αριθμό θύρας προορισμού (3345) για τον φυλλομετρητή του υπολογιστή του οικιακού δικτύου. Στη συνέχεια, ο δρομολογητής ξαναγράφει τη διεύθυνση προορισμού και τον αριθμό θύρας προορισμού του δεδομενογράμματος και προωθεί το δεδομένογράμμο στο οικιακό δίκτυο. Η λειτουργία που περιγράφηκε φαίνεται στην εικόνα 1.



Εικόνα 1: Απεικόνιση της λειτουργίας ενός NAT δρομολογητή

Παρά την ευρεία χρήση του NAT και τη λύση που έδωσε στο πρόβλημα της διευθυνσιοδότησης, υπάρχουν αρκετοί θεωρητικοί που αντιτίθενται στη χρησιμοποίησή του για τους παρακάτω λόγους [39]:

- Οι αριθμοί θυρών πρέπει να χρησιμοποιούνται για διευθυνσιοδότηση διεργασιών και όχι για διευθυνσιοδότηση υπολογιστών.
- Οι δρομολογητές δεν πρέπει να τροποποιούν τα πακέτα επιπέδου μεταφοράς.
- Το πρωτόκολλο NAT παραβιάζει την “αρχή-από-άκρο-σε-άκρο”, σύμφωνα με την οποία οι υπολογιστές πρέπει να συνομιλούν απευθείας μεταξύ τους χωρίς οι παρεμβαλλόμενοι κόμβοι να τροποποιούν τα πακέτα.
- Υποστηρίζουν ότι η λύση στο πρόβλημα της διευθυνσιοδότησης θα

πρέπει να δοθεί μέσω της χρησιμοποίησης του πρωτοκόλλου IPv6 [40].

4. ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP) [41] [42] είναι κλάδος της τεχνητής νοημοσύνης. Αντικείμενο της είναι η επικοινωνία ανθρώπου μηχανής. Η Επεξεργασία Φυσικής Γλώσσας αποτελείται από δύο συνιστώσες την Κατανόηση Φυσικής Γλώσσας (Natural Language Understanding – NLU) [43] [44] και την Παραγωγή Φυσικής Γλώσσας (Natural Language Generation – NLG) [45].

Η NLU περιλαμβάνει την αντιστοίχιση της φυσικής γλώσσας σε εσωτερική αναπαράσταση του συστήματος με τρόπο τέτοιο ώστε να είναι κατανοητή από αυτό. Επιπλέον, αντικείμενο της NLU είναι η ανάλυση διαφόρων παραμέτρων της γλώσσας.

Η NLG σχετίζεται κυρίως με τη διαδικασία σχηματισμού κειμένου ή προτάσεων με νόημα οι οποίες προκύπτουν από κάποια εσωτερική αναπαράσταση. Τρία βασικά στάδια της διαδικασίας αυτής είναι:

- η προσκόμιση των κατάλληλων δεδομένων από τη βάση (text planning).
- η επιλογή των κατάλληλων λέξεων και ο σχηματισμός φράσεων υιοθετώντας το κατάλληλο ύφος (sentence planning).
- ο σχηματισμός προτάσεων που προκύπτει από τη συνένωση των φράσεων που σχηματίστηκαν στο προηγούμενο στάδιο (text realization).

Η είσοδος ενός συστήματος Επεξεργασίας Φυσικής Γλώσσας μπορεί να είναι είτε κείμενο φυσικής γλώσσας, είτε φωνή. Η προσέγγιση που χρησιμοποιούν σήμερα

οι αλγόριθμοι βασίζεται στη μηχανική μάθηση και τη στατιστική, σε αντίθεση με παλιότερα που βασιζόνταν στη χειρονακτική εισαγωγή κανόνων στο σύστημα. Η μηχανική μάθηση είναι μια προσέγγιση που στοχεύει στην παραγωγή νέων κανόνων και συμπερασμάτων από ήδη υπάρχοντες κανόνες.

Μερικές περιοχές στις οποίες έχει αφιερωθεί μεγάλο κομμάτι της έρευνας της Επεξεργασίας Φυσικής Γλώσσας είναι οι παρακάτω:

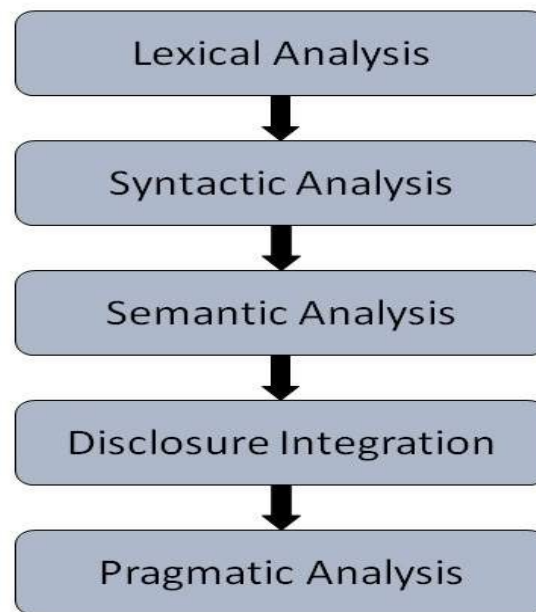
- **Αυτόματη Παραγωγή Περίληψης (Automatic summarization):** Σκοπός είναι η παραγωγή περίληψης για ένα κείμενο φυσικής γλώσσας.
- **Ανάλυση Αναφοράς (Coreference resolution) :** Προσδιορισμός λέξεων που αναφέρονται στην ίδια οντότητα μέσα σε ένα κείμενο.
- **Μηχανική Μετάφραση (Machine Translation) :** Σκοπός είναι η αυτόματη μετάφραση ενός κειμένου από μια ανθρώπινη γλώσσα σε μια

άλλη. Είναι από τα πιο δύσκολα προβλήματα, ανήκει στην κατηγορία "AI-complete" [46] προβλημάτων.

- **Παραγωγή Φυσικής Γλώσσας (Natural language generation)** :Μετατροπή πληροφοριών από τη βάση δεδομένων σε γλώσσα κατανοητή από τον άνθρωπο.
- **Κατανόηση Φυσικής Γλώσσας (Natural language understanding)** : Μετατροπή ενός κειμένου σε κάποια εσωτερική αναπαράσταση κατανοητή από τον υπολογιστή.
- **Οπτική Αναγνώριση Χαρακτήρων (Optical character recognition – OCR)** : Δίνεται ως είσοδος μια εικόνα με κείμενο και σκοπός είναι η αναγνώριση του κειμένου.
- **Αναγνώριση Μερών-του-λόγου (Part-of-speech Tagging)** : Δεδομένης μιας πρότασης να βρεθεί τί μέρος του λόγου είναι κάθε λέξη. Η δυσκολία είναι ότι πολλές φορές υπάρχουν ασάφειες, όπως η λέξη "book" μπορεί να είναι ρήμα ή ουσιαστικό.
- **Parsing** : Προσδιορισμός του parse tree [47] μιας πρότασης. Η δυσκολία με τις φυσικές γλώσσες είναι ότι κάθε πρόταση μπορεί να έχει πολλές διαφορετικές μορφές.

Στην Εικόνα 2 φαίνονται τα κύρια στάδια της Επεξεργασίας Φυσικής Γλώσσας. Αυτά διακρίνονται σε:

- **Λεκτική Ανάλυση (Lexical Analysis)**: Είναι η διαδικασία προσδιορισμού και ανάλυσης της δομής του κειμένου. Το κείμενο διαμερίζεται σε παραγράφους, προτάσεις και λέξεις.
- **Συντακτική Ανάλυση (Syntactic Analysis)**: Είναι η ανάλυση των προτάσεων με τρόπο ώστε να φαίνεται η σχέση που έχουν οι λέξεις μεταξύ τους. Η πρόταση "The school goes to boy" πρέπει να απορρίπτεται από τον συντακτικό αναλυτή της αγγλικής γλώσσας.
- **Σημασιολογική Ανάλυση (Semactic Analysis)**: Προσδιορίζει το ακριβές νόημα του κειμένου. Το κείμενο ελέγχεται για την ύπαρξη νοήματος. Η πρόταση "hot ice-cream" θα πρέπει να απορρίπτεται από έναν αγγλικό σημασιολογικό αναλυτή, λόγω απουσίας νοήματος.
- **Disclosure Integration**: Αποτελεί τη διαδικασία που επαναπροσδιορίζει το νόημα μιας πρότασης με βάση τις προτάσεις που προηγούνται και έπονται.
- **Πραγματολογική Ανάλυση (Pragmatic Analysis)**: Κατά τη διαδικασία αυτή επαναπροσδιορίζεται το νόημα του κειμένου με βάση γνώση που απαιτείται από τον πραγματικό κόσμο.



Εικόνα 2: Τα κύρια στάδια της NLP

4.1 Μηχανική Μετάφραση [48]

Η Μηχανική Μετάφραση (Machine Translation-MT) αποτελεί πεδίο μελέτης της Επεξεργασίας Φυσικής Γλώσσας. Σκοπός είναι ένας υπολογιστής να δέχεται ως είσοδο κείμενο σε φυσική γλώσσα και να παράγει ένα κείμενο με ίδιο νόημα σε άλλη γλώσσα. Σημαντικές παράμετροι σε μια τέτοια μετατροπή είναι:

- η σημασιολογική ακεραιότητα, δηλαδή το αρχικό και τελικό κείμενο να έχουν ίδιο νόημα.
- η ακεραιότητα ύφους, δηλαδή το αρχικό και τελικό κείμενο να έχουν το ίδιο ύφος.
- η συντακτική ακεραιότητα, που σημαίνει ότι το αρχικό και το μεταφρασμένο κείμενο έχουν την ίδια σύνταξη
- η λεκτική ακεραιότητα, δηλαδή η μετάφραση κάθε λέξης του αρχικού κειμένου.
- η χωρική ακεραιότητα, η οποία αφορά στην διατήρηση των χαρακτηριστικών - όπως το μέγεθος και η θέση των λέξεων μέσα στο κείμενο-του αρχικού κειμένου και στο τελικό κείμενο.

Στις μέρες μας οι αλγόριθμοι μηχανικής μετάφρασης στοχεύουν κυρίως στη σημασιολογική ακεραιότητα, ενώ όταν η μηχανική μετάφραση βρισκόταν στα πρώτα της στάδια στόχος ήταν η λεκτική ακεραιότητα, δηλαδή η μετάφραση των λέξεων μία προς μία. Σύντομα, όμως, έγινε σαφές ότι η προσέγγιση αυτή δεν οδηγούσε σε ποιοτικές μεταφράσεις, καθώς κατά τη διαδικασία

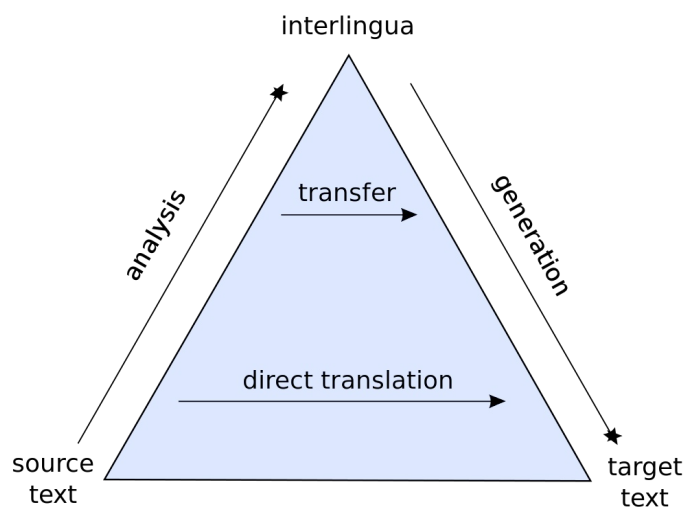
μετάφρασης θα πρέπει να λαμβάνονται υπόψιν παράγοντες όπως οι συντακτικοί κανόνες και οι ιδιωτισμοί της κάθε γλώσσας. Έχουν αναπτυχθεί αρκετές διαφορετικές προσεγγίσεις για τη διαδικασία της μηχανικής μετάφρασης οι οποίες αναλύονται παρακάτω.

4.1.1 Μετάφραση Βασισμένη Σε Κανόνες (Rule-based Machine Translation)

Η προσέγγιση αυτή στηρίζεται σε λεξικά και γραμματικούς κανόνες. Γίνεται μορφολογική, σημασιολογική και συντακτική ανάλυση της αρχικής και της τελικής γλώσσας και χρησιμοποιούνται parse trees για να γίνει η σύνδεση μεταξύ των δυο γλωσσών. Σημαντικό μειονέκτημα αυτή της μεθόδου είναι ότι πρέπει όλοι οι κανόνες να εισαχθούν στο σύστημα χειρονακτικά, γεγονός που δυσκολεύει πολύ τη διαδικασία συντήρησης. Η προσέγγιση αυτή έχει τρεις βασικές υποκατηγορίες:

- **Απευθείας Μετάφραση (Direct Translation):** Είναι η αρχαιότερη τεχνική. Η μετάφραση βασίζεται κυρίως σε λεξικά και στην κατά λέξη μετάφραση του αρχικού κειμένου, χωρίς να δίνεται ιδιαίτερη σημασία στο νόημα που προκύπτει. Η συντακτική δομή της πρότασης εξετάζεται σε πολύ βασικό επίπεδο.
- **Μετάφραση βασισμένη στη μεταφορά (Transfer-based Machine Translation):** Αποτελεί μια από τις περισσότερο χρησιμοποιούμενες μεθόδους. Η διαδικασία μετάφρασης διακρίνεται σε τρία στάδια:
 - α) Η ενδιάμεση αναπαράσταση εξαρτάται από το ζευγάρι αρχικής-τελικής γλώσσας. Για την παραγωγή κειμένου χρησιμοποιούνται λεξικά και γραμματικές. Η μέθοδος αυτή έχει υψηλή ακρίβεια μετάφρασης.
 - β) την αναπαράσταση του κειμένου σε μια ενδιάμεση μορφή που βοηθάει την παραγωγή κειμένου στην τελική γλώσσα (target language)
 - γ) την παραγωγή κειμένου στην τελική γλώσσα (target language).Η ενδιάμεση αναπαράσταση εξαρτάται από το ζευγάρι αρχικής-τελικής γλώσσας. Για την παραγωγή κειμένου χρησιμοποιούνται λεξικά και γραμματικές. Έχει υψηλή ακρίβεια μετάφρασης.
- **Interlingual Machine Translation:** Η προσέγγιση αυτή μοιάζει πολύ με την transfer-based MT με τη διαφορά ότι η ενδιάμεση αναπαράσταση δεν εξαρτάται από το ζευγάρι αρχικής-τελικής γλώσσας (language independent).

Στην Εικόνα 3 απεικονίζεται η πυραμίδα του Bernard Vauquois. Η πυραμίδα συγκρίνει για τις τρεις μεθόδους που αναφέρθηκαν το βάθος στο οποίο αναλύεται η αρχική γλώσσα προτού παραχθεί το μεταφρασμένο κείμενο στην τελική γλώσσα. Όσο πιο ψηλά στην πυραμίδα βρίσκεται μια μέθοδος τόσο περισσότερο αναλύει το αρχικό κείμενο. Η απευθείας μετάφραση βρίσκεται στη βάση της πυραμίδας, εφόσον δεν πραγματοποιεί ουσιαστική ανάλυση κειμένου. Στην κορυφή βρίσκεται η μέθοδος Interlingual η οποία πραγματοποιεί την πιο βαθειά μορφολογική, συντακτική και σημασιολογική ανάλυση του αρχικού κειμένου.



Εικόνα 3: Η πυραμίδα του Bernard Vauquois

4.1.2 Μετάφραση Βασισμένη Στην Στατιστική (Statistical Machine Translation - SMT) [49]

Η μετάφραση με αυτή τη μέθοδο γίνεται με βάση στατιστικές μεθόδους που βασίζονται σε δίγλωσσα corpora [50]. Όταν τα corpora είναι πολλά η μετάφραση μπορεί να είναι πολύ καλή, όμως τα corpora είναι δύσκολο να φτιαχτούν και επίσης για αρκετά ζευγάρια γλωσσών είναι λίγα.

Τα πλεονεκτήματα Στατιστικής Μετάφρασης είναι :

- Δεν επικεντρώνεται στη μετάφραση συγκεκριμένων ζευγαριών γλωσσών.
- Δεν χρειάζεται πολύς κώδικας “στο χέρι”, όπως για τους κανόνες των rule-based μεθόδων.
- Οι μεταφράσεις είναι αρκετά καλές.

Ωστόσο, η μετάφραση με βάση τη στατιστική συνοδεύεται από μειονεκτήματα,

που περιλαμβάνουν:

- Τα corpora που αποτελούν το βασικό εργαλείο είναι ακριβά.
- Υπάρχουν λάθη που είναι δύσκολο να επιλυθούν.
- Η μέθοδος αυτή δεν λειτουργεί καλά για ευγάρια γλωσσών με πολύ διαφορετική δομή και σύνταξη.
- Αν εξαιρέσουμε τις γλώσσες της δυτικής Ευρώπης δεν υπάρχουν αρκετά corpora.

4.1.3 Μετάφραση Βασισμένη Σε Παραδείγματα (Example-based Machine Translation - EBMT)

Η συγκεκριμένη προσέγγιση χρησιμοποιεί δίγλωσσα corpora (bilingual corpora). Οι μεταφράσεις γίνονται με βάση την αναλογία προς προηγούμενες μεταφράσεις και προς ανάλογες μεταφράσεις που περιλαμβάνει το corpora. Η λογική είναι ότι η πρόταση υποδιαιρείται σε φράσεις καθεμιά από τις οποίες έχει ένα νόημα και το τελικό νόημα προκύπτει από τη συνένωση των υπονοημάτων.

Στο πίνακάκι που ακολουθεί φαίνεται ένα πιθανό περιεχόμενο κάποιου δίγλωσσου corpus, που χρησιμοποιείται για τη μετάφραση από τα αγγλικά στα ιαπωνικά.

English	Japanese
How much is that red umbrella ?	Ano akai kasa wa ikura desu ka.
How much is that small camera ?	Ano chiisai kamera wa ikura desu ka.

Η διατριβή των Davide Turcato και Fred Popowich [51] είναι εξαιρετική πηγή για τα συστήματα μετάφρασης βασισμένης σε παραδείγματα.

4.1.4 Υβριδικές Μέθοδοι Μετάφρασης

Συχνά χρησιμοποιούνται υβριδικές μέθοδοι για την μετάφραση κειμένου. Οι υβριδικές μέθοδοι βασίζονται κυρίως στη στατιστική μετάφραση, χρησιμοποιούν όμως και συντακτικούς κανόνες. Τέτοιες τεχνικές έχουν μελετηθεί αρκετά, όμως δεν έχουν επιτυχία γιατί είναι δύσκολο να γίνει μίξη δυο τόσο διαφορετικών τεχνικών, όπως η rule-based και η statistics-based.

Οι πιο σημαντικές κατηγορίες υβριδικών μεθόδων είναι:

- **Κανόνες ακολουθούμενοι από στατιστική (Rules post-processed by statistics):** Οι μεταφράσεις πραγματοποιούνται χρησιμοποιώντας rule-based αλγορίθμους. Στη συνέχεια χρησιμοποιείται στατιστική με την προοπτική να γίνει διόρθωση-βελτίωση του αποτελέσματος [52].
- **Στατιστική οδηγούμενη από κανόνες (Statistics guided by rules):** Το βασικό μέρος της μετάφρασης γίνεται χρησιμοποιώντας στατιστική. Επιπλέον, χρησιμοποιούνται κανόνες για προεπεξεργασία του κειμένου αποσκοπώντας στην υποβοήθηση της στατιστικής μετάφρασης, αλλά και για την επεξεργασία του αποτελέσματος της στατιστικής μετάφρασης.

4.1.5 Μετάφραση Google

Το Google Translate [53] κάνει μετάφραση βασισμένη στη στατιστική. Σαρώνει εκατοντάδες χιλιάδες αρχεία - που έχουν ήδη μεταφραστεί από ανθρώπους – αναζητώντας συγκεκριμένα πρότυπα και πραγματοποιεί εφυείς επιλογές βασισμένες στα πρότυπα αυτά. Γενικά, όσο περισσότερα μεταφρασμένα αρχεία υπάρχουν, τόσο καλύτερη θα είναι η ποιότητα της μετάφρασης.

Το Google δεν μεταφράζει απευθείας από μια αρχική γλώσσα L1 σε μια τελική γλώσσα L2. Αυτό που κάνει, συνήθως, είναι να μεταφράζει πρώτα την αρχική γλώσσα στα αγγλικά και στη συνέχεια τα αγγλικά στην τελική γλώσσα (L1->Αγγλικά->L2). Πολλές φορές προκύπτουν λάθη στη μετάφραση, επειδή στα αγγλικά υπάρχουν ασάφειες (όπως σε όλες τις φυσικές γλώσσες), αλλά και λόγω της στατιστικής μετάφρασης. Συχνά λάθη που παρατηρούνται είναι η μετάφραση λέξεων της αρχικής γλώσσας σε λέξεις της τελικής γλώσσας που έχουν παρόμοιο αλλά όχι ίδιο νόημα, καθώς και η αντιστροφή του νοήματος μιας πρότασης.

Σύμφωνα με τον Franz Josef Och [54] για να δημιουργηθεί ένα ισχυρό σύστημα μεταφράσεων μεταξύ δυο γλωσσών από την αρχή χρειάζεται ένα δίγλωσσο corpus με τουλάχιστον 150-200 εκατομμύρια λέξεις και επιπλέον δυο μονόγλωσσα corpora με περισσότερες από ένα δισεκατομμύριο λέξεις. Για να βρει η Google αυτή την τεράστια ποσότητα πληροφορίας χρησιμοποίησε τα αρχεία Ηνωμένων Εθνών. Ο οργανισμός Ηνωμένων Εθνών (ΟΗΕ) δημοσιεύει τακτικά αρχεία για κάθε μια από τις έξι επίσημες γλώσσες του [55] (Αγγλικά, Γαλλικά, Αραβικά, Ισπανικά, Κινέζικα, Ρωσικά) και με αυτόν τον τρόπο έχουν δημιουργηθεί τεράστια corpora για αυτές τις γλώσσες.

5. ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ RESTAURANT FINDER

Βασικό μέρος της παρούσας πτυχιακής ήταν η επέκταση του συστήματος Restaurant Finder, το οποίο είναι μια διαδικτυακή εφαρμογή που υλοποιεί ένα σύστημα συστάσεων για εστιατόρια του Λονδίνου. Ο κύριος φόρτος εργασίας για την ανάπτυξη αυτού του συστήματος αφορούσε στη συλλογή και αξιοποίηση των πληροφοριών για τα εστιατόρια.

Περίληπτικά, η δημιουργία του Restaurant Finder έγινε ως εξής. Οι πληροφορίες των εστιατορίων του Λονδίνου συλλέχτηκαν - με τη βοήθεια της java βιβλιοθήκης jsoup [56] - από τέσσερα διαφορετικά συστήματα συστάσεων, τα “tripadvisor.com”, “yelp.com”, “zomato.com” και “hardens.com”. Το jsoup διαβάζοντας τον html κώδικα της σελίδας πήρε για κάθε εστιατόριο πληροφορίες (μεταξύ άλλων) για το όνομα, τη διεύθυνση, τα είδη κουζινών καθώς και για τις αγγλικές κριτικές του. Η πληροφορία αυτές συγκεντρώθηκαν σε ένα αρχείο για κάθε ζεύγος εστιατορίου και ιστοτόπου και οργανώθηκαν με συγκεκριμένη μορφή μέσα σε αυτό. Στη συνέχεια τα αρχεία δόθηκαν ως είσοδος στο πρόγραμμα GATE [57], το οποίο προσέθεσε ετικέτες (tags) σε κατάλληλα σημεία στο αρχείο, κυρίως στις κριτικές. Οι ετικέτες αυτές παρέχουν πληροφορία χρήσιμη στη διαδικασία αξιολόγησης διαφόρων παραμέτρων του εστιατορίου και συγκεκριμένα της ατμόσφαιρας, του φαγητού, της εξυπηρέτησης και της σχέσης ποιότητας-τιμής. Αφού έγινε η προσθήκη ετικετών ένα συγκεκριμένο πρόγραμμα διάβαζε τα αρχεία και περνούσε τις πληροφορίες των εστιατορίων στη βάση. Επιπλέον, το πρόγραμμα αυτό χρησιμοποιούσε τα tags - που είχε προσθέσει το GATE πιο πριν - ώστε να παράξει βαθμολογίες για το κάθε εστιατόριο, ως προς τις παραμέτρους που αναφέρθηκαν και να τις αποθηκεύσει στη βάση. Μετά από αυτό η διαδικτυακή εφαρμογή είχε τα κατάλληλα δεδομένα ώστε να λειτουργήσει.

Παρακάτω περιγράφονται αναλυτικά δομικά μέρη του Restaurant Finder. Στη εικόνα Κ απεικονίζεται η δομή του συστήματος.

• ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

Η συνιστώσα αυτή είναι υπεύθυνη για τη συλλογή όλων των απαραίτητων δεδομένων για τα εστιατόρια και τις κριτικές τους. Χρησιμοποιώντας το jsoup api, σαρώνει τον html κώδικα της εκάστοτε ιστοσελίδας, επιλέγει τα χρήσιμα δεδομένα και στη συνέχεια τα αποθηκεύει στο αρχείο που αντιστοιχεί στο ζεύγος εστιατορίου-ιστοτόπου. Υπάρχει συγκεκριμένη μορφή-πρότυπο που ακολουθεί για

να αποθηκεύσει τα δεδομένα στο αρχείο, έτσι ώστε αυτά να μπορούν να αξιοποιηθούν από τα υπόλοιπα δομικά μέρη του συστήματος. Οι ιστότοποι από τους οποίους συλλέγει πληροφορίες είναι οι tripadvisor.com, hardens.com, yelp.com, zomato.com , όμως πολύ εύκολα γίνεται να οριστεί οποιοσδήποτε άλλος ιστότοπος.

- **ΕΞΟΡΥΞΗ ΠΛΗΡΟΦΟΡΙΑΣ**

Η συγκεκριμένη συνιστώσα της εφαρμογής δέχεται ως είσοδο τα αρχεία δεδομένων που δημιουργεί το προηγούμενο στάδιο, η ΣΥΛΛΟΓΗ ΠΛΗΡΟΦΟΡΙΩΝ. Η αποστολή αυτής της συνιστώσας είναι να γίνει επεξεργασία των δεδομένων που συλλέχθηκαν σε προηγούμενο στάδιο και η μετατροπή τους από φυσική γλώσσα σε μια μορφή κατανοήσιμη από τον υπολογιστή. Η μέθοδος που ακολουθείται στην εφαρμογή είναι η προσθήκη ετικετών (tags) απευθείας στο κείμενο φυσικής γλώσσας, η οποία σε συνδυασμό με τη χρήση κανόνων και λεξικών καθιστούν δυνατή την εξόρυξη πληροφορίας από το κείμενο. Οι ετικέτες εκφράζουν μια κλίμακα βαθμολόγησης για διάφορα στοιχεία του εστιατορίου. Η παρούσα συνιστώσα χρησιμοποιεί το λογισμικό GATE για την εισαγωγή ετικετών.

- **ΕΠΕΞΕΡΓΑΣΙΑ ΚΑΙ ΑΠΟΘΗΚΕΥΣΗ ΠΛΗΡΟΦΟΡΙΩΝ**

Η συνιστώσα αυτή δουλεύει επάνω στα σχολιασμένα (annotated) αρχεία που δημιουργεί η προηγούμενη συνιστώσα, η ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ. Σκοπός της είναι η εισαγωγή όλων των πληροφοριών στη βάση δεδομένων της εφαρμογής. Χρησιμοποιείται ένας αλγόριθμος υπολογισμού βαθμολογίας εστιατορίων, ο οποίος διαβάζει τις ετικέτες των αρχείων και δημιουργεί μια τελική βαθμολογία για κάθε ένα από τα βαθμολογούμενα χαρακτηριστικά (σχέση ποιότητα-τιμής, εξυπηρέτηση, ατμόσφαιρα, ποιότητα φαγητού) κάθε εστιατορίου. Επίσης, εξάγονται από τα αρχεία και αποθηκεύονται στη βάση οι βασικές πληροφορίες κάθε εστιατορίου, όπως το όνομα και η διεύθυνση του.

- **ΕΥΡΕΣΗ ΣΥΣΤΑΣΕΩΝ**

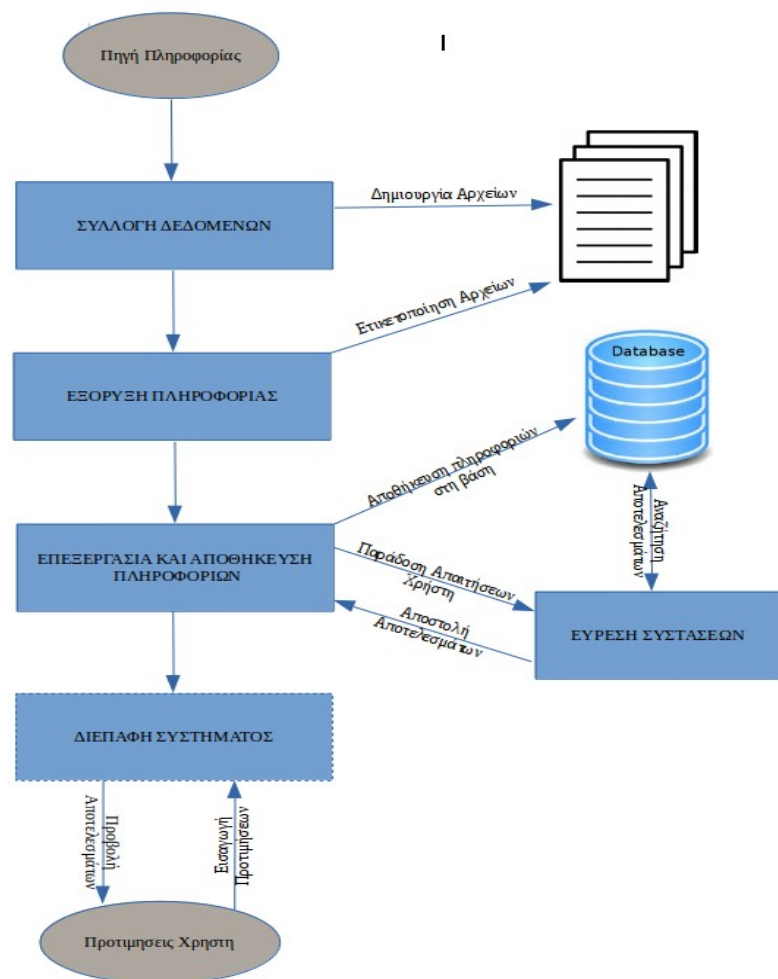
Το συγκεκριμένο κομμάτι του συστήματος δέχεται ως είσοδο τις προτιμήσεις που εισάγει ο χρήστης από τη διεπαφή και ανατρέχει στη βάση δεδομένων για να υπολογίσει τις κατάλληλες συστάσεις. Ο αλγόριθμος λαμβάνει υπόψιν τόσο τις προτιμήσεις που δήλωσε ο

χρήστης, όσο και τα χαρακτηριστικά που δεν δήλωσε ότι τον ενδιαφέρουν και τελικά προσπαθεί να κατατάξει και να παρουσιάσει τα αποτελέσματα-συστάσεις με την καλύτερη δυνατή σειρά (στην κορυφή των αποτελεσμάτων μπαίνει το εστιατόριο που ανταποκρίνεται καλύτερα στις προτιμήσεις του χρήστη).

• ΔΙΕΠΑΦΗ ΣΥΣΤΗΜΑΤΟΣ

Πρόκειται για το σύνολο επικοινωνίας μεταξύ του συστήματος και του χρήστη. Ο χρήστης δηλώνει στο σύστημα της προτιμήσεις του μέσω της διεπαφής, ενώ το σύστημα παρουσιάζει τα αποτελέσματα του στον χρήστη μέσω της διεπαφής. Εναλλακτικά, με βάση τα δομικά μέρη του συστήματος, οι προτιμήσεις του χρήστη γίνονται είσοδος στη ΕΥΡΕΣΗ ΣΥΣΤΑΣΕΩΝ και τα αποτελέσματα της ΕΥΡΕΣΗΣ ΣΥΣΤΑΣΕΩΝ εμφανίζονται στη διεπαφή.

Αναλυτική περιγραφή και λεπτομέρειες υλοποίησης για την εφαρμογή αυτή βρίσκεται στην αναφορά της πτυχιακής της Δωροθέας Τσιμπίδη [1].



Εικόνα 4: Η δομή του Restaurant Finder

6. ΕΠΕΚΤΑΣΗ ΤΟΥ RESTAURANT FINDER

Συνολικά πραγματοποιήθηκαν τρεις επεκτάσεις στο σύστημα Restaurant Finder. Συγκεκριμένα, η μεταφορά της εφαρμογής σε web server, ώστε να είναι προσβάσιμη από το διαδίκτυο, η επέκταση της βάσης δεδομένων με εστιατόρια της Αθήνας και η αξιοποίηση των κριτικών όλων των γλωσσών. Οι επεκτάσεις αυτές περιγράφονται αναλυτικά πιο κάτω.

6.1 Μεταφορά Του Restaurant Finder Σε Εξυπηρέτη

Ως web εξυπηρέτης χρησιμοποιήθηκε ο προσωπικός μου υπολογιστής. Χρειάστηκε κάποια διαδικασία ώστε να γίνει αυτό εφικτό. Αρχικά ο εφοδιασμός του υπολογιστή με το κατάλληλο λογισμικό και στη συνέχεια κάποιες ρυθμίσεις στον δρομολογητή για να γίνει το μηχάνημα προσπελάσιμο από το δίκτυο.

Ξεκινώντας από το λογισμικό, έγινε εγκατάσταση του λειτουργικού συστήματος “Ubuntu-Server 14.04.4 LTS” [58], το οποίο είναι ένα λειτουργικό σύστημα ανοικτού κώδικα (open source) ειδικά σχεδιασμένο για τη λειτουργία ενός εξυπηρέτη. Το Ubuntu-Server δεν έχει γραφικό περιβάλλον. Αφού η εγκατάσταση ολοκληρώθηκε με επιτυχία ακολούθησε η εγκατάσταση ειδικού λογισμικού για τη φιλοξενία ιστοσελίδων. Συγκεκριμένα, εγκαταστάθηκε το “Apache Tomcat” [59], δεδομένου ότι η εφαρμογή αναπτύχθηκε σε java, και το “mysql-server” [60]. Ο Apache Tomcat είναι ένα λογισμικό ανοικτού κώδικα που υλοποιεί web εξυπηρέτη για java εφαρμογές. Το λογισμικό mysql-server είναι και αυτό ανοιχτού κώδικα και αποτελεί ένα πρόγραμμα διαχείρισης σχεσιακών βάσεων δεδομένων [61] (relational database management system – RDBMS).

Μετά την εγκατάσταση του κατάλληλου λογισμικού έπρεπε να γίνει προώθηση θύρας (port forwarding), δηλαδή να δηλωθεί μια θύρα από την οποία ο δρομολογητής θα προωθεί τα αφικνούμενα πακέτα στον server. Σε αυτό το κομμάτι συμβαίνει ό,τι ακριβώς περιγράφηκε στο κεφάλαιο 4 για τους μεταφραστές διευθύνσεων δικτύου. Όσον αφορά τα διαδικαστικά, πρέπει να γίνει είσοδος στις ρυθμίσεις του δρομολογητή στη διεύθυνση 192.168.1.1 μέσω του φυλλομετρητή. Στη συνέχεια, πρέπει να γίνει εισαγωγή εγγραφής στον πίνακα μετάφρασης NAT (NAT Translation Table) του router, συνεπώς δηλώνονται η private ip του server, η θύρα του server στην οποία “περιμένει” πακέτα ο Tomcat και η θύρα του δρομολογητή η οποία θα αντιστοιχεί στον server. Στην εικόνα 5 φαίνεται αυτή η διαδικασία. Στην εικόνα 6 φαίνεται η απομονωμένη εφαρμογή και στην εικόνα 7 η σύνδεση του Restaurant Finder στο διαδίκτυο και η δυνατότητα απομακρυσμένης προσπέλασης.

Advanced

Interface Setup

Advanced Setup

Access Management

Maintenance

Status

Firewall

Routing

NAT

ADSL

QoS

PortBinding

IPv6 Routing

Virtual Server

Virtual Server for : Single IPs Account/ PVC0

Application : HTTP_Ser HTTP_Server

Protocol : TCP

Start Port Number : 8080

End Port Number : 8080

Local IP Address : 192.168.1.5

Start Port Number(Local) : 8080

End Port Number(Local) : 8080

Virtual Server Listing

Rule	Application	Protocol	Start Port	End port	Local IP Address	Start Port Local	End Port Local	Edit	Drop
0	HTTP_Server	TCP	8080	8080	192.168.1.5	8080	8080		
1	N/A	N/A	N/A	N/A	N/A	N/A	N/A		
2	N/A	N/A	N/A	N/A	N/A	N/A	N/A		
3	N/A	N/A	N/A	N/A	N/A	N/A	N/A		
4	N/A	N/A	N/A	N/A	N/A	N/A	N/A		
5	N/A	N/A	N/A	N/A	N/A	N/A	N/A		
6	N/A	N/A	N/A	N/A	N/A	N/A	N/A		
7	N/A	N/A	N/A	N/A	N/A	N/A	N/A		
8	N/A	N/A	N/A	N/A	N/A	N/A	N/A		
9	N/A	N/A	N/A	N/A	N/A	N/A	N/A		

APPLY

BACK

CANCEL

Εικόνα 5: Στον NAT Address Table η πρώτη εγγραφή ορίζει ότι τα αφικνούμενα πακέτα στη θύρα 8080 του δρομολογητή θα προωθούνται στη θύρα 8080 του υπολογιστή με διεύθυνση υποδικτύου 192.168.1.5



Εικόνα 6: Restaurant Finder σε τοπικό υπολογιστή. Δεν είναι δυνατή η προσπέλαση του από το διαδίκτυο



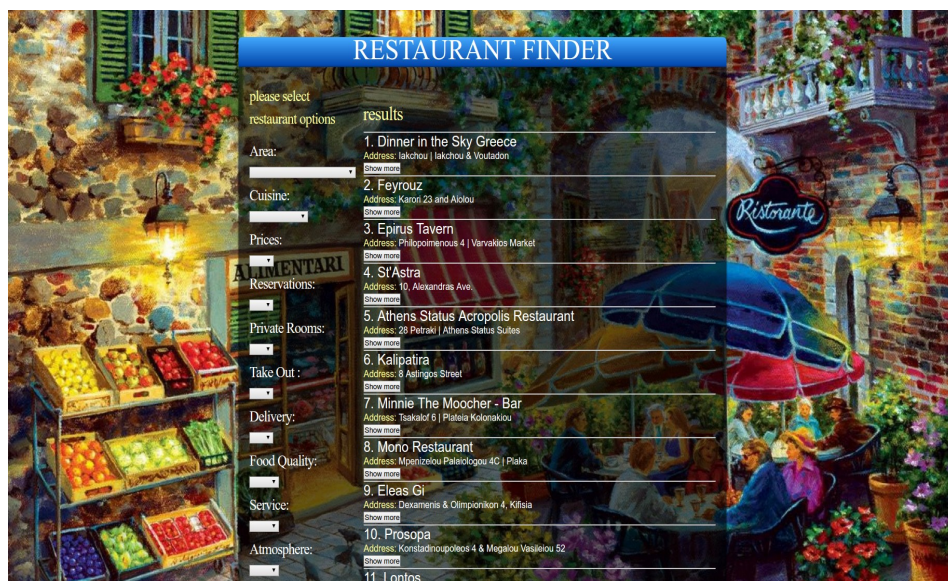
Εικόνα 7: Μεταφορά του Restaurant Finder σε εξυπηρέτη. Η προσπέλαση του μέσω του διαδικυου είναι εφικτή.

6.2 Προσθήκη Πληροφοριών Για Εστιατόρια Της Αθήνας

Η συγκεκριμένη επέκταση αποσκοπούσε κυρίως στον έλεγχο της φορητότητας (portability) της εφαρμογής. Χρησιμοποιήθηκαν τα προγράμματα και η μέθοδοι που της αρχικής έκδοσης του Restaurant Finder. Οι πληροφορίες για τα εστιατόρια συλλέχθηκαν μόνο από το tripadvisor.com, καθώς τα yelp.com, hardens.com και zomato.com δεν περιλαμβάνουν ελληνικά εστιατόρια.

Το πιο σημαντικό βήμα της διαδικασίας είναι αλλαγή του συνδέσμου που διαβάζει το jsoup, έτσι ώστε αντί για τα εστιατόρια του Λονδίνου να διαβάζει αυτά της Αθήνας. Επίσης, χρειάστηκαν μερικές αλλαγές σε συγκεκριμένα σημεία του κώδικα (για παράδειγμα όπου αναμενόταν να διαβαστεί “London” έπρεπε να μετατραπεί σε “Athens”), ώστε γίνει το parsing της σελίδας σωστά. Γενικά, ο HTML κώδικας των ιστοσελίδων για τα εστιατόρια της Αθήνας δεν είχε σημαντικές διαφορές από τον αντίστοιχο κώδικα για το Λονδίνο, οπότε δεν χρειάστηκαν ουσιαστικές αλλαγές.

Το συμπέρασμα που προέκυψε είναι ότι η εφαρμογή της Δωροθέας είναι portable όταν η πληροφορία συλλέγεται από κάποιον ιστότοπο από τους tripadvisor.com, τα yelp.com, hardens.com και zomato.com. Αυτό οφείλεται στο γεγονός ότι για κάθε έναν από αυτούς τους ιστότοπους είναι ήδη έτοιμος ο κώδικας που διαβάζει τις ιστοσελίδες. Ωστόσο, η προσθήκη εστιατορίων από κάποιον άλλον ιστότοπο πέρα από αυτούς δεν θα είναι εύκολη, καθώς θα πρέπει να γραφτεί ο parsing κώδικας από την αρχή.



Εικόνα 8: Η αρχική σελίδα της εφαρμογής με τα αποτελέσματα της αναζήτησης σε ελληνικά εστιατόρια

Η συγκεκριμένη επέκταση τροποποιεί τη συνιστώσα του συστήματος “ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ”. Στις εικόνες 7 και 8 φαίνονται δυο screenshots της εφαρμογής, με εστιατόρια της Αθήνας.



Εικόνα 9: Η σελίδα με τις πληροφορίες του εστιατορίου "Dinner in the Sky"

6.3 Αξιοποίηση Κριτικών Από Πολλές Γλώσσες

Η τελευταία επέκταση αφορά στις κριτικές. Το αρχικό πρόγραμμα είχε πληροφορία μόνο για τις αγγλικές κριτικές. Μετά την επέκταση αυτή για κάθε εστιατόριο συλλέχθηκε πληροφορία για όλες τις κριτικές στα αγγλικά, ελληνικά, ιταλικά, γαλλικά, γερμανικά, σουηδικά, ολλανδικά, τούρκικα, δανέζικα, τσέχικα, φινλανδικά, ρωσικά, πορτογαλικά. Έγινε χρήση του εργαλείου google translate για java με όνομα "gtranslateapi-1.0.jar" [62], το οποίο χρησιμοποιήθηκε για απευθείας μετάφραση του κειμένου, προτού αποθηκευτεί η πληροφορία στα αρχεία. Δηλαδή, η επέκταση αυτή επίσης αφορά τη συνιστώσα “ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ”.

Όσον αφορά το τεχνικό μέρος, κάθε γλώσσα στο tripadvisor έχει ξεχωριστή κατάληξη στο url, για παράδειγμα τα αγγλικά βρίσκονται στο tripadvisor.com, τα ελληνικά στο tripadvisor.com.gr, τα ιταλικά στο tripadvisor.it κ.ο.κ. Συνεπώς, προκειμένου να διαβαστούν οι κριτικές από διαφορετικές γλώσσες

δημιουργήθηκε ένας βρόγχος επανάληψης. Σε κάθε επανάληψη διαμορφωνόταν κατάλληλα η κατάληξη του url, άρα διαβαζόταν ο αντίστοιχος σύνδεσμος για την εκάστοτε γλώσσα.

Σχετικά με το `gtranslateapi-1.0.jar`, προκειμένου να χρησιμοποιηθεί έπρεπε να δημιουργηθεί λογαριασμός στο `console.developers.google.com` και να δηλωθεί η χρήση του εργαλείου. Το πακέτο αυτό περιέχει κλάσεις και συναρτήσεις που διευκολύνουν εξαιρετικά την μεταγλώττιση κειμένου.

Από τη χρήση του εργαλείου αυτού προέκυψαν τα εξής συμπεράσματα για τις μεταφράσεις της google:

- Οι μεταφράσεις για τις γλώσσες της ευρώπης είναι πολύ καλές.
- Για γλώσσες με πολύ διαφορετική σύνταξη από τα αγγλικά οι μεταφράσεις δεν είναι αρκετά καλές (όπως αναφέρεται στο κεφάλαιο 4.1.5 η μετάφραση από μια γλώσσα σε μια άλλη δεν γίνεται απευθείας, αλλά πρώτα στα αγγλικά). Σε αυτές τις γλώσσες, μεταξύ άλλων, περιλαμβάνονται τα τούρκικα, κινέζικα, γιαπωνέζικα, αιγυπτιακά, αραβικά, φιλιππινέζικα, κορεάτικα, βιετναμέζικα, μαλαισιανά, φιλιππινέζικα. Η κακή ποιότητα της μετάφρασης σε αυτές τις γλώσσες απαιτούσε πολλές τροποποιήσεις του κώδικα και σε πολλές από αυτές δεν υπήρχαν κριτικές, για αυτούς τους δυο λόγους δεν εισήχθησαν στο πρόγραμμα. Μια ακόμη πιθανή αιτία της κακής μετάφρασης των συγκεκριμένων γλωσσών η απουσία ικανοποιητικού αριθμού από corpora.
- Το google translate πολλές φορές, αλλά όχι πάντα, δεν αντιλαμβάνεται τα ορθογραφικά και τυπογραφικά λάθη, καθώς επίσης και τα greeklish. Το γεγονός αυτό έχει άμεσο αντίκτυπο στην εφαρμογή, διότι το GATE δεν τοποθετήσει tag στις λέξεις που δεν έχουν μεταφραστεί σωστά και κατά συνέπεια θα επηρεαστεί η βαθμολόγηση που βγάζει το πρόγραμμα για το εκάστοτε εστιατόριο.

Ένα παράδειγμα μετάφρασης με χρήση του `gtranslateapi-1.0.jar` φαίνεται στις εικόνες 10, 11.

Απιθανη εμπειρια και αξιζει καποιος να το επισκεπτει.
Σιγουρα καποια στιγμή θα ξαναερθουμε να
απολαυσουμε τη θεα και το φαγητο που ηταν
εξαιρετικο.

Εικόνα 10: Αρχικό κείμενο στα ελληνικά

Cool experience and worth to visit someone. Certainly some point will come back to enjoy the view and the food was excellent.

Εικόνα 11: Μεταφρασμένο κείμενο από τα ελληνικά στα αγγλικά

7.ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα πτυχιακή έγινε θεωρητική μελέτη των συστημάτων συστάσεων και των τεχνικών αξιολόγησης τους, των αρχιτεκτονικών εφαρμογών διαδικτύου, του NAT πρωτοκόλλου και επίσης μελέτη του κλάδου της Επεξεργασίας Φυσικής Γλώσσας. Επιπλέον, σημαντικό μέρος της πτυχιακής αποτέλεσε η επέκταση του Restaurant Finder.

Η εφαρμογή μεταφέρθηκε σε web εξυπηρέτη και πλέον είναι προσβάσιμη από το διαδίκτυο. Επίσης, περιλαμβάνει πληροφορίες και για εστιατόρια της Αθήνας επιπρόσθετα σε αυτά του Λονδίνου που περιλάμβανε η αρχική εφαρμογή. Ακόμη, η βάση δεδομένων περιέχει πληροφορίες από κριτικές δεκαεπτά διαφορετικών γλωσσών.

Το συμπέρασμα που προέκυψε από την επέκταση της βάσης με αθηναϊκά εστιατόρια είναι ότι η εφαρμογή έχει υψηλό βαθμό φορητότητας (portability), όταν τα δεδομένα συλλέγονται από ιστοτόπους με έτοιμο κώδικα διαβάσματος ιστοσελίδας. Οι ιστοτόποι αυτοί είναι οι tripadvisor.com, hardens.com, yelp.com, zomato.com. Για οποιονδήποτε από αυτούς τους ιστοτόπους είναι αρκετά εύκολη η συλλογή δεδομένων. Αρκεί μέσα στον κώδικα να τροποποιηθεί ο αρχικός σύνδεσμος από τον οποίο διαβάζονται τα δεδομένα και ίσως χρειαστούν μερικές επιπλέον αλλαγές. Η συλλογή πληροφοριών από διαφορετικούς ιστοτόπους θα είναι πιο δύσκολη, διότι απαιτείται η συγγραφή κώδικα διαβάσματος από την αρχή.

Η εισαγωγή στη βάση περισσότερων κριτικών χάριν των μεταφράσεων οδήγησε αναμφίβολα σε πιο ακριβείς βαθμολογίες των εστιατορίων.

Όσον αφορά τις μεταφράσεις που έγιναν μέσω του εργαλείου `gtranslateapi-1.0.jar` έγινε φανερό ότι οι μεταφράσεις των ευρωπαϊκών γλωσσών της δυτικής Ευρώπης είναι πολύ ποιοτικές. Το γεγονός αυτό πιθανώς οφείλεται στην ομοιότητα των συγκεκριμένων γλωσσών με την αγγλική, καθώς και στην ύπαρξη ικανοποιητικού αριθμού εγγράφων στο διαδίκτυο. Οι μεταφράσεις για τις υπόλοιπες γλώσσες δεν είναι τόσο καλές.

Σχετικά με την εφαρμογή συστάσεων μένει να γίνει αξιολόγηση των αποτελεσμάτων που βγάξει και η σύγκρισή τους με τα αποτελέσματα άλλων συστημάτων συστάσεων, όπως το [tripadvisor](http://tripadvisor.com). Επίσης, θα μπορούσε να γίνει προσθήκη ακόμη περισσότερων εστιατορίων από διαφορετικές πόλεις και να προστεθεί η δυνατότητα εισαγωγής σχολίων για κάθε εστιατόριο.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	Ελληνικός Όρος
Recommender system	Σύστημα συστάσεων
Portability	Φορητότητα
Network Address Translation	Μετάφρασης Διευθύνσεων Δικτύου
Collaborative Filtering	Συνεργατικό Φιλτράρισμα
Content-Based Filtering	Φιλτράρισμα βασισμένο στο Περιεχόμενο
Knowledge-based Filtering	Φιλτράρισμα βασισμένο στη Γνώση
Demographic Filtering	Φιλτράρισμα βασισμένο στα Δημογραφικά Στοιχεία
Hybrid filtering	Υβριδικό φιλτράρισμα
Rule-based Machine Translation	Μετάφραση Βασισμένη Σε Κανόνες
Statistical Machine Translation	Μετάφραση Βασισμένη Στην Στατιστική
Example-based Machine Translation	Μετάφραση Βασισμένη Σε Παραδείγματα
Client-server architecture	Αρχιτεκτονική πελάτη-εξυπηρετή
Peer-to-peer architecture	Αρχιτεκτονική ομοτίμων
Information agents	Πράκτορες πληροφόρησης
Cold start	Αργή εκκίνηση
Scalability	Κλιμάκωση
Sparsity	Αραιότητα
Hybridization	Υβριδοποίηση
Switching Hybridization	Σταθμισμένη Υβριδοποίηση
Mixed Hybridization	Μεικτή Υβριδοποίηση
Feature Combination	Συνδυασμός Χαρακτηριστικών
Feature Augmentation	Αύξηση χαρακτηριστικών
Cascade Hybridization	Αλληλουχία υβριδοποίησης
Meta-level hybridization	Υβριδοποίηση μεταεπιπέδου
Root Square Mean Error	Τετραγωνική Ρίζα Μέσου Τετραγώνου

	του Σφάλματος
Mean Absolute Error	Μέση Απόλυτη Τιμή Σφάλματος
Normalized RMSE	Κανονικοποιημένη RMSE
Normalized MAE	Κανονικοποιημένη MAE
Average RMSE	Μέση RMSE
Average MAE	Μέση MAE
Coverage	Κάλυψη
Confidence	Αυτοπεποίθηση
Novelty	Καινοτομία
Diversity	Ποικιλία
Utility	Χρησιμότητα
Risk	Ρίσκο
Robustness	Ευρωστία
Privacy	Ιδιωτικότητα
Adaptivity	Προσαρμοστικότητα
Scalability	Κλιμάκωση
Server Farm	Φάρμα Εξυπηρετών
World Wide Web	Παγκόσμιος Ιστός
Self-scalable	Αυτο-κλιμακούμενο
Private IP	Ιδιωτική IP
Natural Language Processing	Επεξεργασία Φυσικής Γλώσσας
Natural Language Understanding	Κατανόηση Φυσικής Γλώσσας
Natural Language Generation	Παραγωγή Φυσικής Γλώσσας
Automatic summarization	Αυτόματη Παραγωγή Περίληψης
Coreference resolution	Ανάλυση Αναφοράς
Machine Translation	Μηχανική Μετάφραση
Optical Character Recognition	Οπτική Αναγνώριση Χαρακτήρων
Part-of-speech Tagging	Αναγνώριση Μερών-του-λόγου
Lexical Analysis	Λεκτική Ανάλυση
Syntactic Analysis	Συντακτική Ανάλυση
Semantic Analysis	Σημασιολογική Ανάλυση

Pragmatic Analysis	Πραγματολογική Ανάλυση
Direct Translation	Απευθείας Μετάφραση
Transfer-based Machine Translation	Μηχανική Μετάφραση βασισμένη στη μεταφορά
Rules post-processed by statistics	Κανόνες ακολουθούμενοι από στατιστική
Statistics guided by rules	Στατιστική οδηγούμενη από κανόνες
Tag	Ετικέτα
Relational Database Management System	Πρόγραμμα Διαχείρισης Σχεσιακών Βάσεων Δεδομένων
Port Forwarding	Προώθηση Θύρας
NAT Translation Table	Πίνακα Μετάφρασης NAT
Router	Δρομολογητής

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

NAT	Network Address Translation
CF	Collaborative Filtering
CBF	Content-Based Filtering
KBF	Knowledge-based Filtering
DF	Demographic Filtering
SMT	Statistical Machine Translation
EBMT	Example-based Machine Translation
P2P	Peer-to-peer
RSME	Root Square Mean Error
MAE	Mean Absolute Error
NRMSE	Normalized RMSE
NMAE	Normalized MAE
ARMSE	Average RMSE
AMAE	Average MAE
WWW	World Wide Web
NLP	Natural Language Processing
NLU	Natural Language Understanding
NLG	Natural Language Generation
OCR	Optical Character Recognition
MT	Machine Translation
RDBMS	Relational Database Management System

ΑΝΑΦΟΡΕΣ

- [1] Δ. Τσιμπίδη, «Σύστημα Συστάσεων για Εστιατόρια: Η Περίπτωση των Εστιατορίων του Λονδίνου», Πτυχιακή Εργασία, Τμήμα Πληροφορικής & Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, 2015.
- [2] C.A. Knoblock and J.L. Ambite, *Agents for Information Gathering*, Information Sciences Institute and Department of Computer Science, California, 1997.
- [3] C.A. Gomez-Urbe and Neil Hunt, The Netflix Recommender System: Algorithms, Business Value and Innovation, *ACM Transactions on Management Information Systems (TMIS)*, vol 6, Jan. 2015
- [4] S. Phillips, "A brief history of Facebook", the Guardian ; <https://www.theguardian.com/technology/2007/jul/25/media.newmedia> [Προσπελάστηκε 16/5/16]
- [5] N. Carlson, "How Twitter Was Founded", Business Insider, Apr. 2011; <http://www.businessinsider.com/how-twitter-was-founded-2011-4?op=1> [Προσπελάστηκε 16/5/16]
- [6] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl: "GroupLens: an open architecture for collaborative filtering of netnews", *Proc. ACM Conf. Computer Supported Cooperative Work (CSCW 94)*, ACM, 1994, pp. 175–186.
- [7] J. Leskovec, A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2010, pp. 307-341
- [8] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, Recommender systems survey, *Knowledge-Based Systems*, vol. 46, 2013
- [9] Linyuan Lu, Matus Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang and Tao Zhou, *Recommender Systems*, Elsevier, 2012
- [10] D.H. Park, H.K. Kim, I.Y. Choi and J.K. Kim, "A Review and Classification of Recommender Systems Research", *International Conf. Social Science Humanity (ICSCH 2011)*, IACSIT Press, 2011
- [11] F. Ricci, L. Rokach, B. Shapira and P. Kantor, *Introduction to*

Recommender Systems Handbook, Springer-Verlag New York, 2011

[12] P. Melville and V. Sindhvani, *Recommender Systems*, Springer, 2010, pp 829-838

[13] M.D. Ekstrand, J.T. Riedl, and J.A. Konstan, Collaborative Filtering Recommender Systems, *Foundations and Trends*, vol. 4 , No. 2, 2011, pp. 81-173

[14] M.J. Pazzani, A Framework for Collaborative Content-Based and Demographic Filtering, *Artificial Intelligence Review*, vol. 13, No. 5-6, Dec. 1999, pp. 393-408

[15] S. Lee, J. Yang and S.Y. Park, *Discovery of Hidden Similarity on Collaborative Filtering to Overcome Sparsity Problem*, Discovery Science, 2007

[16] Mustansar Ali Ghazanfar and Adam Prugel-Bennett, *Fulfilling the Needs of Gray-Sheep Users in Recommender Systems A Clustering Solution*, 2011

[17] "Linkedin: A Short Historical Review", *Tiki-Toki*;
<http://www.tiki-toki.com/timeline/entry/347471/Linkedin-A-Short-Historical-Review/> [Προσπελάστηκε 17/5/16]

[18] R. Meteren and M. Someren, *Using Content-Based Filtering for Recommendation*,

[19] T. Ryan, "AN ORAL HISTORY OF RT, PART ONE: THE BEGINNING", June 2008; <http://editorial.rottentomatoes.com/article/an-oral-history-of-rt-part-one-the-beginning/> [Προσπελάστηκε 17/5/16]

[20] D.C. Chmielewski, "Col Needham created IMDb", *Los Angeles Times*, Jan. 2013 ; <http://articles.latimes.com/2013/jan/19/business/la-fi-himi-needham-20130120> [Προσπελάστηκε 17/5/16]

[21] A. Ostrow, "Jinni's Genius Way to Recommend Movies", Feb. 2009; <http://mashable.com/2009/02/11/jinni/#Gw4Ddr9kdZqu> [Προσπελάστηκε 17/5/16]

[22] C. Taylor, "Rovi Showcases Early Version of Media Guide and Solutions Portfolio for Digital Home Entertainment Products during CES 2010", Jan.

2010 ; <http://www.rovicorp.com/company/press-releases/archive/2010/1/7/rovi-showcases-early-versionmedia-guide-and-solutions-portfolio-.html> [Προσπελάστηκε 17/5/16]

[23] <http://jamanetwork.com/> [Προσπελάστηκε 17/5/16]

[24] <https://www.pandora.com/login.vm> [Προσπελάστηκε 17/5/16]

[25] R. Burke, Hybrid Recommender Systems: Survey and Experiments, *User Modeling and User-Adapted Interaction*, vol. 12, No. 4, Nov. 2002, pp. 331-370.

[26] R. Burke, "Hybrid web recommender systems", *The adaptive web*, Springer-Verlag, Berlin, pp. 377-408.

[27] A. Gupta, A. Mohapatra and T. Tennesi, *Towards a hybrid approach to Netflix Challenge*, Mar. 2009.

[28] G. Shani and A. Gunawardana, *Evaluating Recommendation Systems*, Springer, 2010, pp. 257-297.

[29] Haroon Shakirat Oluwatosin, Client-Server Model, *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 16, No. 1, Feb. 2014, pp. 67-71.

[30] "World Wide Web Timeline", Pews Research Center Mar. 2014;
<http://www.pewinternet.org/2014/03/11/world-wide-web-timeline/>
[Προσπελάστηκε 19/5/16]

[31] C. Grant and K. Wooding, "FTP Reviewed", Jul. 1998;
<http://pintday.org/whitepapers/ftp-review.shtml> [Προσπελάστηκε 19/5/16]

[32] Herbert P. Lockett, What's News: Electronic-mail delivery gets started, *Popular Science*, Vol. 202, No. 3. Mar. 1973, p. 85.

[33] *Peer-to-Peer Network*, The Government of the Hong Kong Special Administrative Region, 2008

[34] "BitTorrent"; <https://en.wikipedia.org/wiki/BitTorrent> [Προσπελάστηκε 19/5/16]

[35] "LimeWire"; <https://en.wikipedia.org/wiki/LimeWire> [Προσπελάστηκε

19/5/16]

[36] "Skype"; <https://en.wikipedia.org/wiki/Skype> [Προσπελάστηκε 19/5/16]

[37] "IPv4"; <https://en.wikipedia.org/wiki/IPv4> [Προσπελάστηκε 19/5/16]

[38] F. Audet and C. Jennings, *Network Address Translation (NAT) Behavioral Requirements for Unicast UDP*, Network Working Group , 2007

[39] J.F. Kurose, K.W. Ross, *Computer networking*, Pearson, 2008, pp. 353–356.

[40] "IPv6" ; <https://en.wikipedia.org/wiki/IPv6> [Προσπελάστηκε 19/5/16]

[41] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, 1999.

[42] "A brief history of Natural Language Processing"; http://www.cs.bham.ac.uk/~pjh/sem1a5/pt1/pt1_history.html [Προσπελάστηκε 19/5/16]

[43] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2009.

[44] E. Ovchinnikova, Integration of World Knowledge for Natural Language Understanding, *Atlantis Press*, vol. 3, 2012.

[45] E. Reiter and R. Dale, *Building Natural Language Generation Systems*, Cambridge University Press, 2000.

[46] S.C. Shapiro, *Artificial Intelligence*, Encyclopedia of Artificial Intelligence, 1992.

[47] "Parse tree"; https://en.wikipedia.org/wiki/Parse_tree [Προσπελάστηκε 19/5/16]

[48] M.W. Madsen, *The Limits of Machine Translation*, 2009.

[49] M. Collins, *Statistical Machine Translation: IBM Models 1 and 2*.

[50] K. Wolk and K. Marasek, *A Sentence Meaning Based Alignment Method*

for Parallel Text Corpora Preparation, Dept. Multimedia Polish Japanese Institute of Information Technology, 2015.

[51] D. Turcato and F. Popowich, ed. M. Carl and A. Way, *Recent Advances in Example-Based Machine Translation*, Springer Netherlands, 2003, pp 59-81.

[52] Mi.Simard, N. Ueffing, P. Isabelle and R. Kuhn, “Rule-based Translation With Statistical Phrase-based Post-editing”, *Proc. 2nd Workshop on Statistical Machine Translation (StatMT '07)*, 2007, pp. 203-206.

[53] “Google Translate”; https://en.wikipedia.org/wiki/Google_Translate
[Προσπελάστηκε 19/5/16]

[54] P. Koehn, F. J. Och and D. Marcu, “Statistical Phrase-Based Translation”, *Proc. 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03)*, Association for Computational Linguistics, 2003, pp. 48-54.

[55] “Official languages of the United Nations”;
https://en.wikipedia.org/wiki/Official_languages_of_the_United_Nations
[Προσπελάστηκε 19/5/16]

[56] Jonathan Hedley, “jsoup: Java HTML Parser”, 2009; <https://jsoup.org/>
[Προσπελάστηκε 19/5/16]

[57] “GATE” official website; <https://gate.ac.uk/> [Προσπελάστηκε 19/5/16]

[58] Ubuntu 14.04 LTS information;
http://cdimage.ubuntu.com/netboot/14.04/?_ga=1.49898978.558990943.1448034046 [Προσπελάστηκε 19/5/16]

[59] “Apache Tomcat” official website; <http://tomcat.apache.org/>
[Προσπελάστηκε 19/5/16]

[60] Download “MySQL Server” from here;
<https://dev.mysql.com/downloads/mysql/> [Προσπελάστηκε 19/5/16]

[61] “SQL - RDBMS Concepts”; <http://www.tutorialspoint.com/sql/sql-rdbms-concepts.htm> [Προσπελάστηκε 19/5/16]

[62] “Translating Text with REST ”;

https://cloud.google.com/translate/v2/using_rest [Προσπελάστηκε 19/5/16]