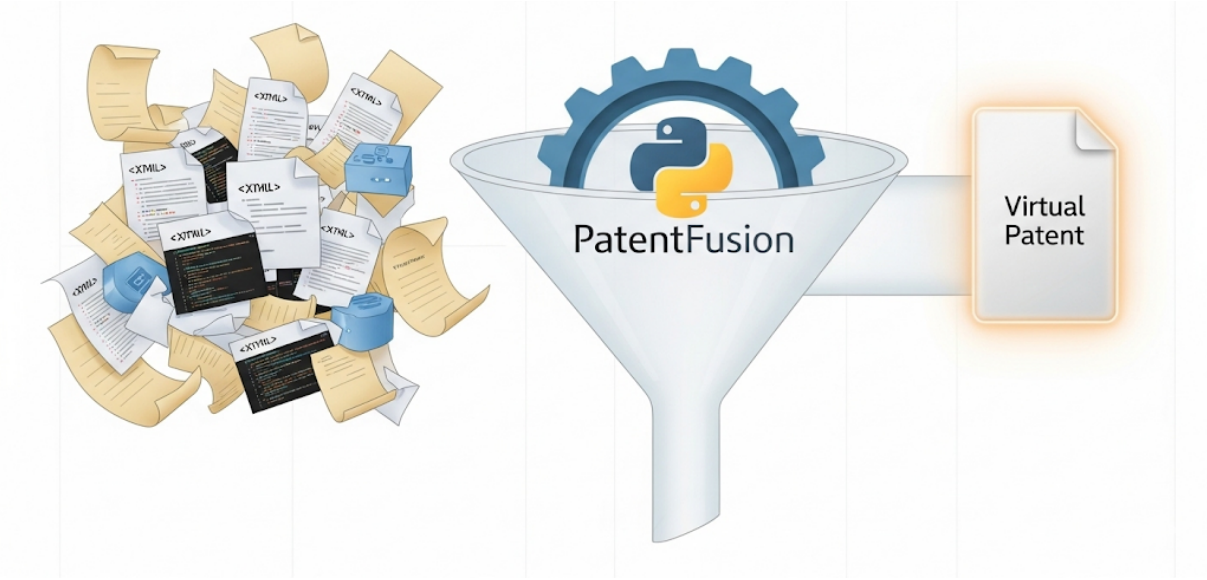
ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Ανάπτυξη Συστήματος Επεξεργασίας & Ανάλυσης Μεγάλου Όγκου Δεδομένων: Εφαρμογή σε Πατέντες»



Του φοιτητή
Παπαδόπουλου Χρήστου
Αρ. Μητρώου: 10/2022

Επιβλέπων
Σαλαμπάσης Μιχάλης
Βαθμίδα Καθηγητής

Ημερομηνία 31/08/2025

Τίτλος Δ.Ε. Ανάπτυξη Συστήματος Επεξεργασίας & Ανάλυσης Μεγάλου Όγκου Δεδομένων: Εφαρμογή σε Πατέντες
Κωδικός Δ.Ε. 25177
Ονοματεπώνυμο φοιτητή/των Παπαδόπουλος Χρήστος
Ονοματεπώνυμο εισηγητή Σαλαμπάσης Μιχάλης
Ημερομηνία ανάληψης Δ.Ε. 12/03/2025
Ημερομηνία περάτωσης Δ.Ε. 31/08/2025

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

«Στην πολυαγαπημένη μου σύζυγο και στα λατρεμένα μας παιδιά»

Πρόλογος

Η επιλογή της συγκεκριμένης διπλωματικής εργασίας προέκυψε από την ανάγκη να αντιμετωπίσω ένα πραγματικό και επίκαιρο πρόβλημα στον τομέα της πληροφορικής. Κατά τη διάρκεια των σπουδών μου στο μεταπτυχιακό πρόγραμμα «Ευφυής Τεχνολογίες Διαδικτύου», συνειδητοποίησα ότι η επεξεργασία μεγάλων όγκων δεδομένων αποτελεί κεντρική πρόκληση για τους σύγχρονους μηχανικούς πληροφορικής. Η ευκαιρία να εργαστώ με δεδομένα διπλωμάτων ευρεσιτεχνίας, που χαρακτηρίζονται από εξαιρετική πολυπλοκότητα και όγκο, μου προσέφερε ένα ιδανικό πλαίσιο για την εφαρμογή και επέκταση των γνώσεών μου.

Η εργασία αυτή με βοήθησε να αναπτύξω βαθιά κατανόηση των σύγχρονων αρχιτεκτονικών επεξεργασίας δεδομένων, της streaming τεχνολογίας, και των τεχνικών διαχείρισης μνήμης. Επιπλέον, η συνεργασία μου με διεθνείς ερευνητές στο πλαίσιο του WPI+ μου έδωσε πολύτιμη εμπειρία στην ακαδημαϊκή έρευνα και την ανάπτυξη συστημάτων που χρησιμοποιούνται από την παγκόσμια ερευνητική κοινότητα. Τέλος, η πρακτική εφαρμογή προηγμένων τεχνικών πληροφορικής σε πραγματικά προβλήματα με μετρήσιμα αποτελέσματα ενίσχυσε σημαντικά τις τεχνικές μου δεξιότητες και την αυτοπεποίθησή μου ως μηχανικός πληροφορικής.

Περίληψη

Η παρούσα διπλωματική εργασία παρουσιάζει το PatentFusion, ένα καινοτόμο σύστημα για την επεξεργασία και ανάλυση μεγάλου όγκου δεδομένων διπλωμάτων ευρεσιτεχνίας. Το κεντρικό πρόβλημα που αντιμετωπίζει, είναι ο κατακερματισμός της πληροφορίας που προκύπτει από την ύπαρξη πολλαπλών εκδόσεων του ίδιου διπλώματος (διαφορετικά kind codes), γεγονός που δημιουργεί σημαντικές προκλήσεις στην έρευνα πατεντών.

Η λύση που προτείνεται είναι η δημιουργία "Εικονικών Πατεντών" (Virtual Patents), τα οποία συνδυάζουν την πλέον επικαιροποιημένη πληροφορία από όλες τις διαθέσιμες εκδόσεις ενός διπλώματος σε ένα ενοποιημένο έγγραφο. Το σύστημα διατηρεί την πλήρη ιεραρχία XML και την ιχνηλασιμότητα της προέλευσης του κάθε στοιχείου μέσω των kind-source attributes.

Το PatentFusion υλοποιεί προηγμένες τεχνικές όπως πολυεπεξεργασία ροής (streaming multiprocessing) για τη διαχείριση μεγάλων datasets, συγχώνευση βάση προτεραιότητας (priority-based merging) για την επιλογή της βέλτιστης πληροφορίας από κάθε έκδοση, και ευφυή ανίχνευση διπλότυπων (intelligent duplicate detection) για την αποφυγή επαναλήψεων. Η αρχιτεκτονική του, αφού έχει δοκιμαστεί σε επεξεργασία datasets άνω των 150GB, μπορεί να ανταπεξέλθει σε απεριόριστου μεγέθους datasets με ελάχιστη χρήση μνήμης, χάρις την αρχιτεκτονική ροής που υλοποιεί.

Η πειραματική αξιολόγηση στη συλλογή WPI αποδεικνύει ότι το σύστημα μειώνει τον αριθμό των αρχείων κατά 22-35% χωρίς απώλεια πληροφορίας. Το PatentFusion ενσωματώθηκε στο πλαίσιο του WPI+, συμβάλλοντας στην τυποποίηση και αναπαραγωγιμότητα των πειραμάτων στον τομέα των διπλωμάτων ευρεσιτεχνίας.

Το σύστημα αντιμετωπίζει θεμελιώδεις προκλήσεις στην ανάκτηση πληροφοριών πατεντών και στην επεξεργασία φυσικής γλώσσας παρέχοντας στους ερευνητές ενοποιημένα, πλήρη έγγραφα που εξαλείφουν την πολυπλοκότητα της διαχείρισης πολλαπλών εκδόσεων πατεντών. Αυτό βελτιώνει σημαντικά την αποδοτικότητα της έρευνας διατηρώντας πλήρη ακεραιότητα και ιχνηλασιμότητα της πληροφορίας. Η εργασία αποτελεί σημαντική συνεισφορά στον τομέα της ανάλυσης πατεντών και αποδεικνύει την πρακτική της εφαρμογή σε διάφορα ερευνητικά σενάρια συμπεριλαμβανομένων της αναζήτησης prior-art, της ταξινόμησης πατεντών, και των διαγλωσσικών εργασιών ανάκτησης.

«System Development for Processing & Analyzing Large Volumes of Data: Application to Patents»

«Christos Papadopoulos»

Abstract

This thesis presents PatentFusion, an innovative system for processing and analyzing large volumes of patent data. The core problem addressed is information fragmentation arising from the existence of multiple versions of the same patent (different kind codes), which creates significant challenges in patent research.

The proposed solution is the creation of "Virtual Patents," which combine the most up-to-date information from all available versions of a patent into a unified document. The system maintains complete XML hierarchy and traceability of each element's origin through kind-source attributes.

PatentFusion implements advanced techniques including streaming multiprocessing for handling large datasets, priority-based merging for selecting optimal information from each version, and intelligent duplicate detection to avoid repetitions. Its architecture, having been tested on datasets larger than 150GB, can handle datasets of unlimited size with minimal memory usage, thanks to its streaming architecture.

Experimental evaluation on the WPI collection demonstrates that the system reduces the number of files by 22-35% without information loss. PatentFusion was integrated into the WPI+ framework, contributing to standardization and reproducibility of experiments in the patent domain.

The system addresses fundamental challenges in patent information retrieval and natural language processing by providing researchers with consolidated, complete documents that eliminate the complexity of managing multiple patent versions. This significantly improves research efficiency while maintaining full information integrity and traceability. The work represents a substantial contribution to the field of patent analysis and demonstrates practical applicability across various research scenarios including prior-art search, patent classification, and cross-lingual retrieval tasks.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την οικογένειά μου, τη γυναίκα μου και τα παιδιά μου, για την υποστήριξη, την υπομονή τους, και το κουράγιο που μου έδιναν καθ' όλη την διάρκεια της εκπόνησης αυτής της εργασίας. Χωρίς την δική τους συμπαράσταση δεν θα ήταν δυνατόν να ολοκληρωθεί η διπλωματική μου εργασία. Τον επιβλέποντα Καθηγητή κ. Μιχάλη Σαλαμπάση που με την πολύ μεγάλη ακαδημαϊκή του εμπειρία και πολύτιμη καθοδήγηση, συνέβαλε τα μέγιστα για το άρτιο αποτέλεσμα αυτής της εργασίας. Τέλος θα ήθελα να ευχαριστήσω την υποψήφια διδάκτορα κα Ελένη Καματέρη για τις ατελείωτες ώρες συνεργασίας, καθοδήγησης, και βαθιάς γνώσης του αντικειμένου, υπήρξε οδηγός στην όλη πορεία της διπλωματικής εργασίας, από την έναρξη, έως και την ολοκλήρωσή της.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες

IR	Information Retrieval
WPI	World Patent Information
CPU	Central Processing Unit
EPO	European Patent Office
USPTO	United States Patent and Trademark Office
WIPO	World Intellectual Property Organization
JPO	Japan Patent Office
KIPO	Korean Intellectual Property Office
CNIPA	China National Intellectual Property Administration
XML	eXtensible Markup Language
NLP	Natural Language Processing
API	Application Programming Interface
REST	Representational State Transfer
PCT	Patent Cooperation Treaty
CPC	Cooperative Patent Classification
IPC	International Patent Classification
IPCR	International Patent Classification Reform
CSV	Comma-Separated Values
JSON	JavaScript Object Notation
GB	Gigabyte
TB	Terabyte
RAM	Random Access Memory
SSD	Solid State Drive
PCIe	Peripheral Component Interconnect Express
NVMe	Non-Volatile Memory Express
USB	Universal Serial Bus
GIL	Global Interpreter Lock
VP	Virtual Patents
CN	China (country code)
JP	Japan (country code)
KR	Korea (country code)

Κεφάλαιο 1ο: Εισαγωγή

1.1 Το πρόβλημα της διαχείρισης πολλαπλών εκδόσεων διπλωμάτων ευρεσιτεχνίας

Η διαχείριση και επεξεργασία διπλωμάτων ευρεσιτεχνίας αποτελεί μια από τις πλέον σύνθετες προκλήσεις στον τομέα της ανάκτησης πληροφοριών (Information Retrieval) και της επεξεργασίας φυσικής γλώσσας. Η πολυπλοκότητα αυτή πηγάζει από την ιδιαίτερη φύση των εγγράφων διπλωμάτων ευρεσιτεχνίας, τα οποία χαρακτηρίζονται από εξειδικευμένη τεχνική ορολογία, πολύπλοκες νομικές διατυπώσεις, και πολυγλωσσικό περιεχόμενο που εκτείνεται σε όλους τους τεχνολογικούς τομείς.

Ένα ιδιαίτερα προβληματικό χαρακτηριστικό του οικοσυστήματος των διπλωμάτων ευρεσιτεχνίας είναι η ύπαρξη πολλαπλών εκδόσεων για το ίδιο δίπλωμα, οι οποίες διακρίνονται μέσω των λεγόμενων kind codes. Τα kind codes αποτελούν τυποποιημένους κωδικούς που υποδεικνύουν το στάδιο και τον τύπο της δημοσίευσης ενός διπλώματος ευρεσιτεχνίας. Για παράδειγμα, ένα δίπλωμα μπορεί αρχικά να δημοσιευτεί ως αίτηση (A1), στη συνέχεια να υποστεί τροποποιήσεις (A2, A4), και τελικά να χορηγηθεί (B1), με πιθανές μεταγενέστερες διορθώσεις (B2, B9). Κάθε έκδοση μπορεί να περιέχει διαφορετικές ή συμπληρωματικές πληροφορίες, δημιουργώντας έναν κατακερματισμό της συνολικής πληροφορίας που αφορά ένα συγκεκριμένο δίπλωμα ευρεσιτεχνίας.

Αυτός ο κατακερματισμός δημιουργεί σημαντικές προκλήσεις τόσο για τους ερευνητές όσο και για τους επαγγελματίες του χώρου. Οι ερευνητές που εργάζονται σε tasks όπως η prior art αναζήτηση, η ταξινόμηση διπλωμάτων ευρεσιτεχνίας, ή η αυτόματη περίληψη, συχνά αντιμετωπίζουν δυσκολίες στην επιλογή της κατάλληλης έκδοσης για την ανάλυσή τους. Παράλληλα, η ύπαρξη πολλαπλών αρχείων για το ίδιο δίπλωμα επιβαρύνει υπολογιστικά τα συστήματα επεξεργασίας και δυσχεραίνει τη διασφάλιση της συνέπειας στα πειραματικά αποτελέσματα.

1.2 Σκοπός και στόχοι της εργασίας

Η παρούσα διπλωματική εργασία στοχεύει στην αντιμετώπιση του προβλήματος του κατακερματισμού της πληροφορίας στα διπλώματα ευρεσιτεχνίας μέσω της ανάπτυξης ενός καινοτόμου συστήματος δημιουργίας Εικονικών Πατεντών (Virtual Patents). Οι Virtual Patents αποτελούν ενοποιημένα έγγραφα που συνδυάζουν την πλέον επικαιροποιημένη πληροφορία από όλες τις διαθέσιμες εκδόσεις ενός διπλώματος ευρεσιτεχνίας, διατηρώντας παράλληλα την πλήρη ιεραρχική δομή XML και την ιχνηλασιμότητα της προέλευσης κάθε στοιχείου.

Οι επιμέρους στόχοι της εργασίας περιλαμβάνουν:

Πρώτον, την ανάλυση και τεκμηρίωση των προκλήσεων που προκύπτουν από την επεξεργασία διπλωμάτων ευρεσιτεχνίας σε επίπεδο kind code σε αντιδιαστολή με την επεξεργασία σε επίπεδο patent family. Αυτή η ανάλυση παρέχει το θεωρητικό υπόβαθρο για την κατανόηση της αναγκαιότητας της δημιουργίας των Virtual Patents.

Δεύτερον, τη σχεδίαση και υλοποίηση του συστήματος PatentFusion, ενός υψηλής απόδοσης εργαλείου που αυτοματοποιεί τη δημιουργία Virtual Patents. Το σύστημα ενσωματώνει προηγμένες τεχνικές όπως streaming multiprocessing για τη διαχείριση μεγάλων datasets, priority-based merging για την επιλογή της βέλτιστης πληροφορίας από κάθε έκδοση, και intelligent duplicate detection για την αποφυγή επαναλήψεων.

Τρίτον, την πειραματική αξιολόγηση της αποτελεσματικότητας των Virtual Patents μέσω στατιστικής ανάλυσης της μείωσης του αριθμού των αρχείων και της διατήρησης της πληρότητας της πληροφορίας.

Η αξιολόγηση περιλαμβάνει σύγκριση των Virtual Patents με τα αρχικά verticals της συλλογής WPI για διάφορα patent offices.

Τέταρτον, τη συνεισφορά στην ευρύτερη ερευνητική κοινότητα μέσω της ενσωμάτωσης του συστήματος στο πλαίσιο του WPI+, μιας επέκτασης της συλλογής WPI που στοχεύει στη βελτίωση της συγκρισιμότητας και αναπαραγωγιμότητας των πειραμάτων στον τομέα της ανάκτησης πληροφοριών από διπλώματα ευρεσιτεχνίας.

1.3 Συνεισφορά της εργασίας στο πλαίσιο του WPI+

Η παρούσα εργασία εντάσσεται στο ευρύτερο πλαίσιο του WPI+ (World Patent Information Plus), μιας πρωτοβουλίας που στοχεύει στην επέκταση και βελτίωση της συλλογής WPI για την υποστήριξη αξιόπιστης και αναπαραγώγιμης έρευνας στον τομέα των διπλωμάτων ευρεσιτεχνίας. Η συλλογή WPI, που δημοσιεύτηκε το 2019 από τους Lupu et al. [1], αποτελεί τη μεγαλύτερη δημόσια διαθέσιμη συλλογή διπλωμάτων ευρεσιτεχνίας, περιλαμβάνοντας 6.3 εκατομμύρια έγγραφα από έξι μεγάλες αρχές διπλωμάτων ευρεσιτεχνίας (USPTO, EPO, WIPO, JPO, KIPO, CNIPA) για τα έτη 2014-2015. Η συλλογή ακολουθεί μια «οριζόντια» προσέγγιση, καλύπτοντας όλους τους τεχνικούς τομείς για μια συγκεκριμένη χρονική περίοδο, σε αντίθεση με τις παραδοσιακές «κάθετες» συλλογές που εστιάζουν σε έναν τομέα για πολλά έτη [2].

Παρά τον πλούτο των δεδομένων της, η αρχική συλλογή WPI αντιμετώπισε περιορισμένη υιοθέτηση από την ερευνητική κοινότητα, κυρίως λόγω της έλλειψης του ground truth για τις βασικές εργασίες, αλλά και της πολυπλοκότητας στη διαχείριση των πολλαπλών εκδόσεων των πατεντών. Το WPI+ αντιμετωπίζει αυτές τις προκλήσεις μέσω της παροχής επιπρόσθετων πόρων, όπως προκαθορισμένα verticals (υποσύνολα της συλλογής), ground truth για διάφορες εργασίες, και εργαλεία ανάλυσης, ένα από αυτά είναι και το PatentFusion.

Η συνεισφορά του συστήματος PatentFusion στο WPI+ είναι πολυδιάστατη. Καταρχάς, τα Virtual Patents που δημιουργεί αποτελούν ένα νέο, εξειδικευμένο vertical (#VP) που απλοποιεί σημαντικά την επεξεργασία και ανάλυση των δεδομένων. Αντί να διαχειρίζονται πολλαπλά αρχεία για κάθε δίπλωμα, οι ερευνητές μπορούν να εργάζονται με ένα ενιαίο, ολοκληρωμένο έγγραφο που περιέχει την πλέον επικαιροποιημένη πληροφορία.

Επιπλέον, το σύστημα υποστηρίζει την αρχή της "soft standardization" που προωθεί το WPI+. Αυτή η προσέγγιση στοχεύει στη βελτίωση της συγκρισιμότητας και αναπαραγωγιμότητας των πειραμάτων χωρίς να επιβάλλει υπερβολικά περιοριστικά πρότυπα. Το PatentFusion επιτυγχάνει αυτόν τον στόχο παρέχοντας σαφείς, τεκμηριωμένες μεθόδους για τη δημιουργία Virtual Patents, διασφαλίζοντας ότι διάφοροι ερευνητές που χρησιμοποιούν το ίδιο εργαλείο θα παράγουν συνεπή αποτελέσματα.

Τέλος, η ανοιχτή διάθεση του κώδικα του συστήματος μέσω του GitHub repository του WPI+ συμβάλλει στη διαφάνεια και επεκτασιμότητα της έρευνας. Άλλοι ερευνητές μπορούν να εξετάσουν, να τροποποιήσουν, και να βελτιώσουν τη μεθοδολογία, προωθώντας τη συνεργατική ανάπτυξη εργαλείων για την κοινότητα των διπλωμάτων ευρεσιτεχνίας. Το πλήρες dataset με όλα τα verticals που αντιστοιχούν στα ανά τον κόσμο Γραφεία Διπλωμάτων Ευρεσιτεχνίας έχουν ανέβει το αποθετήριο του TU Vien https://researchdata.tuwien.ac.at/.

1.4 Δομή της διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία οργανώνεται σε οκτώ κεφάλαια που καλύπτουν τόσο το θεωρητικό υπόβαθρο όσο και την πρακτική υλοποίηση του συστήματος PatentFusion.

Το Κεφάλαιο 2 παρουσιάζει το θεωρητικό υπόβαθρο και τη βιβλιογραφική ανασκόπηση, εστιάζοντας στην εξέλιξη από τη συλλογή WPI στην WPI+, την έννοια της soft standardization, και τη σύγκριση με άλλες υφιστάμενες συλλογές διπλωμάτων ευρεσιτεχνίας.

Το Κεφάλαιο 3 εμβαθύνει στην τεχνική ανάλυση της επεξεργασίας διπλωμάτων σε επίπεδο kind code και patent family, εξετάζοντας τη δομή των XML εγγράφων, το σύστημα των kind codes, και την αναγκαιότητα των Virtual Patents.

Το Κεφάλαιο 4 περιγράφει λεπτομερώς την αρχιτεκτονική και υλοποίηση του συστήματος PatentFusion, αναλύοντας τα επιμέρους modules, τον αλγόριθμο δημιουργίας των Virtual Patents, και τις βελτιστοποιήσεις απόδοσης.

Το Κεφάλαιο 5 παρουσιάζει τη μεθοδολογία υλοποίησης, συμπεριλαμβανομένων των τεχνολογιών που χρησιμοποιήθηκαν, της streaming architecture για μεγάλα datasets, και των διαδικασιών testing και validation.

Το Κεφάλαιο 6 αναλύει τα αποτελέσματα της πειραματικής αξιολόγησης, παρέχοντας στατιστικά στοιχεία για τα Virtual Patents ανά patent office και συγκρίσεις με τα αρχικά verticals του WPI.

Το Κεφάλαιο 7 περιγράφει συνοπτικά τις πρακτικές εφαρμογές των Virtual Patents σε διάφορες ερευνητικές εργασίες όπως prior-art search, patent classification, και summarization.

Τέλος, το Κεφάλαιο 8 συνοψίζει τις κύριες συνεισφορές της εργασίας, αναγνωρίζει τους περιορισμούς της τρέχουσας υλοποίησης, και προτείνει κατευθύνσεις για μελλοντική έρευνα και ανάπτυξη.

Κεφάλαιο 2ο: Θεωρητικό Υπόβαθρο και Βιβλιογραφική Ανασκόπηση

2.1 Εισαγωγή στη συλλογή WPI

Η συλλογή WPI (World Patent Information) αποτελεί ορόσημο στην εξέλιξη των test collections για την έρευνα στον τομέα των διπλωμάτων ευρεσιτεχνίας. Δημοσιεύτηκε το 2019 ως αποτέλεσμα συνεργασίας ακαδημαϊκών και βιομηχανικών φορέων, με στόχο την υποστήριξη αξιόπιστων και αναπαραγώγιμων πειραμάτων στους τομείς της ανάκτησης πληροφοριών (Information Retrieval), της επεξεργασίας φυσικής γλώσσας (Natural Language Processing), και της μηχανικής μάθησης (Machine Learning) [1].

Η συλλογή περιλαμβάνει 6,313,165 έγγραφα διπλωμάτων ευρεσιτεχνίας σε μορφή XML και 55,231,022 συμπληρωματικά αρχεία, όπως εικόνες, χημικές δομές και διαγράμματα. Τα δεδομένα προέρχονται από τις έξι μεγαλύτερες αρχές διπλωμάτων ευρεσιτεχνίας παγκοσμίως: το United States Patent and Trademark Office (USPTO), το European Patent Office (EPO), το World Intellectual Property Organization (WIPO), το Japan Patent Office (JPO), το Korean Intellectual Property Office (KIPO), και το China National Intellectual Property Administration (CNIPA). Η χρονική περίοδος που καλύπτει η συλλογή είναι τα έτη 2014 και 2015, παρέχοντας μια σύγχρονη βάση δεδομένων για ερευνητικούς σκοπούς.

Το καινοτόμο χαρακτηριστικό της WPI είναι η «οριζόντια» προσέγγισή της στη συλλογή δεδομένων. Σε αντίθεση με προηγούμενες συλλογές που εστίαζαν σε συγκεκριμένους τεχνικούς τομείς ή γεωγραφικές περιοχές για μεγάλες χρονικές περιόδους (δηλαδή οι λεγόμενες «κάθετες» συλλογές), η WPI καλύπτει όλους τους τεχνικούς τομείς από όλες τις μεγάλες αρχές διπλωμάτων ευρεσιτεχνίας για μια συγκεκριμένη, περιορισμένη χρονική περίοδο. Αυτή η προσέγγιση επιτρέπει τη διεξαγωγή συγκριτικών μελετών μεταξύ διαφορετικών αρχών και τεχνολογικών πεδίων, διατηρώντας παράλληλα τη χρονική συνέπεια των δεδομένων.

Η συλλογή διατίθεται δωρεάν για ερευνητικούς σκοπούς μέσω της πλατφόρμας Zenodo, με τα XML αρχεία να είναι άμεσα προσβάσιμα και τα συμπληρωματικά αρχεία να διατίθενται κατόπιν αιτήματος λόγω του μεγάλου όγκου τους (περίπου 2TB). Η δομή των αρχείων ακολουθεί την επέκταση του προτύπου ST-36 της WIPO, διασφαλίζοντας τη συμβατότητα με υπάρχοντα εργαλεία επεξεργασίας διπλωμάτων ευρεσιτεχνίας.

2.2 Η εξέλιξη προς το WPI+: Κίνητρα και στόχοι

Παρά τον πλούτο και την ποιότητα των δεδομένων της, η αρχική συλλογή WPI αντιμετώπισε περιορισμένη υιοθέτηση από την ερευνητική κοινότητα. Η ανάλυση των αιτιών αυτής της περιορισμένης χρήσης οδήγησε στην ανάπτυξη του WPI+, μιας επέκτασης που στοχεύει να αντιμετωπίσει τα κύρια εμπόδια που εντοπίστηκαν [3].

Το πρώτο και σημαντικότερο εμπόδιο είναι η απουσία ground truth για βασικές ερευνητικές εργασίες. Ενώ η συλλογή περιέχει πλήρη κείμενα και μεταδεδομένα, δεν παρέχει αξιολογήσεις συνάφειας (relevance assessments) για αναζήτηση prior-art, δεδομένα με ετικέτα (labeled data) για εργασίες ταξινόμησης (classification tasks), ή περιλήψεις αναφοράς (reference summaries) για περιλήψεις

αποσπασμάτων (abstractive summarization). Αυτή η έλλειψη καθιστά δύσκολη τη χρήση της συλλογής για την αξιολόγηση νέων μεθόδων και την σύγκριση με υπάρχουσες προσεγγίσεις.

Το δεύτερο εμπόδιο σχετίζεται με την πολυπλοκότητα της δομής των δεδομένων. Κάθε δίπλωμα ευρεσιτεχνίας μπορεί να αντιπροσωπεύεται από πολλαπλά έγγραφα με διαφορετικά kind codes, δημιουργώντας σύγχυση σχετικά με το ποια έκδοση πρέπει να χρησιμοποιηθεί για κάθε εργασία. Επιπλέον, η έλλειψη προκαθορισμένων υποσυνόλων (verticals) δυσχεραίνει τη διεξαγωγή στοχευμένων πειραμάτων.

Το τρίτο εμπόδιο αφορά την απουσία υποστηρικτικών εργαλείων και τεκμηρίωσης. Οι ερευνητές πρέπει να αναπτύξουν από την αρχή κώδικα για βασικές λειτουργίες όπως την ανάλυση (parsing) των XML αρχείων, την εξαγωγή συγκεκριμένων πεδίων, και τη στατιστική ανάλυση της συλλογής. Αυτές η επαναλαμβανόμενη ανάπτυξη εργαλείων και τεκμηρίωσης οδηγεί σε αναποτελεσματική χρήση πόρων και αυξημένο κίνδυνο σφαλμάτων.

Το WPI+ αντιμετωπίζει συστηματικά αυτές τις προκλήσεις μέσω τεσσάρων βασικών πυλώνων:

Πρώτον, παρέχει ground truth για διάφορες εργασίες, συμπεριλαμβανομένων 2,592 θεμάτων (topics) για αναζήτηση prior-art με αξιολογήσεις συνάφειας βασισμένα σε παραπομπές (citations), τέσσερα test sets για ταξινόμηση πατεντών (patent classification) με διαφορετικές κατανομές labels, και 1,819 πατέντες με περιλήψεις που έχουν εξαχθεί για εργασίες περίληψης.

Δεύτερον, ορίζει σαφώς προκαθορισμένα verticals με μοναδικούς κωδικούς και hash values, όπως το #(EPO,WO,US) που περιλαμβάνει όλα τα αγγλόφωνα διπλώματα από τρεις αρχές, και το #VP (Virtual Patents) που δημιουργείται από το σύστημα PatentFusion που αναπτύχθηκε στο πλαίσιο αυτής της εργασίας.

Τρίτον, παρέχει εκτενή κώδικα σε Python για την ανάλυση και χρήση της συλλογής, συμπεριλαμβανομένων notebooks με παραδείγματα εφαρμογής μεθόδων μηχανικής μάθησης, scripts για την εξαγωγή verticals, και εργαλεία για στατιστική ανάλυση.

Τέταρτον, οργανώνει όλους τους πόρους σε ένα κεντρικό GitHub repository, διευκολύνοντας την πρόσβαση, τη συνεργασία, και τη συνεχή ενημέρωση των διαθέσιμων εργαλείων και δεδομένων.

2.3 Η έννοια του "soft standardization" στα πειράματα με διπλώματα ευρεσιτεχνίας

Το WPI+ εισάγει την έννοια του "soft standardization" ως μια ισορροπημένη προσέγγιση στην τυποποίηση των πειραμάτων με διπλώματα ευρεσιτεχνίας. Αυτή η προσέγγιση αναγνωρίζει την ανάγκη για συγκρισιμότητα και αναπαραγωγιμότητα, χωρίς να επιβάλλει υπερβολικά περιοριστικά πλαίσια που θα μπορούσαν να εμποδίσουν την καινοτομία [3].

Η soft standardization λειτουργεί σε τρία επίπεδα:

Στο επίπεδο των δεδομένων, παρέχονται προκαθορισμένα verticals και ground truths, αλλά οι ερευνητές παραμένουν ελεύθεροι να δημιουργήσουν τα δικά τους υποσύνολα ή να τροποποιήσουν τα υπάρχοντα σύμφωνα με τις ανάγκες τους. Η μόνη απαίτηση είναι η σαφής τεκμηρίωση των επιλογών τους για να διασφαλίζεται η αναπαραγωγιμότητα.

Στο επίπεδο των μεθόδων, παρέχονται βασικές (baseline) εφαρμογές και καλές πρακτικές, αλλά ενθαρρύνεται η ανάπτυξη νέων προσεγγίσεων. Οι ερευνητές καλούνται να συγκρίνουν τις μεθόδους τους με τα παρεχόμενα baselines για να διασφαλίζεται η συγκρισιμότητα των αποτελεσμάτων.

Στο επίπεδο της αξιολόγησης, προτείνονται τυποποιημένες μετρικές και πρωτόκολλα αξιολόγησης, αλλά αναγνωρίζεται ότι διαφορετικές εργασίες μπορεί να απαιτούν διαφορετικές προσεγγίσεις. Η έμφαση δίνεται στη διαφάνεια και την πλήρη τεκμηρίωση των διαδικασιών αξιολόγησης.

Αυτή η προσέγγιση αντιμετωπίζει ένα χρόνιο πρόβλημα στην έρευνα των διπλωμάτων ευρεσιτεχνίας: την έλλειψη συγκρισιμότητας μεταξύ διαφορετικών μελετών. Όπως τεκμηριώνεται στο submission document του WPI+ [3], πολλές δημοσιευμένες εργασίες χρησιμοποιούν διαφορετικά υποσύνολα των ίδιων συλλογών, διαφορετικές διαδικασίες προεπεξεργασίας, ή ακόμα και διαφορετικούς ορισμούς για βασικές έννοιες όπως το τι συνιστά μια "US patent".

2.4 Σύγκριση WPI με Υφιστάμενες συλλογές διπλωμάτων ευρεσιτεχνίας

Η συλλογή WPI διαφοροποιείται σημαντικά από τις υφιστάμενες συλλογές διπλωμάτων ευρεσιτεχνίας τόσο σε επίπεδο περιεχομένου όσο και σε επίπεδο σχεδιαστικής φιλοσοφίας. Μια συγκριτική ανάλυση με τις κύριες εναλλακτικές συλλογές αναδεικνύει τα μοναδικά χαρακτηριστικά και πλεονεκτήματα της WPI.

Η συλλογή CLEF-IP [4-8], που αναπτύχθηκε στο πλαίσιο των ομώνυμων evaluation campaigns (2009-2013), περιλαμβάνει περίπου 3.5 εκατομμύρια έγγραφα από το EPO και το WIPO. Ενώ παρέχει πλούσια ground truths για ανάκτηση και εργασίες κατηγοριοποίησης, περιορίζεται γεωγραφικά στην Ευρώπη και χρονικά σε έγγραφα πριν το 2009. Επιπλέον, υποστηρίζει μόνο τρεις γλώσσες (Αγγλικά, Γαλλικά, Γερμανικά), σε αντίθεση με τις έξι γλώσσες της WPI.

Το USPTO-2M dataset [9], με περίπου 2 εκατομμύρια έγγραφα από το USPTO για την περίοδο 2006-2015, εστιάζει αποκλειστικά στην ταξινόμηση διπλωμάτων με βάση το CPC σύστημα. Δεν περιλαμβάνει πλήρη βιβλιογραφικά δεδομένα ή συμπληρωματικά αρχεία, περιορίζοντας τη χρησιμότητά του για πολυδιάστατες αναλύσεις.

Το BigPatent dataset [10], με 1.3 εκατομμύρια έγγραφα από το USPTO (1971-2018), σχεδιάστηκε ειδικά για abstractive summarization. Περιλαμβάνει μόνο abstracts και descriptions, χωρίς claims ή βιβλιογραφικά δεδομένα, καθιστώντας το ακατάλληλο για άλλες εργασίες όπως prior-art search ή classification.

Το HUPD (Harvard USPTO Patent Dataset) [11], με 4.5 εκατομμύρια έγγραφα από το USPTO (2004-2018), υποστηρίζει πολλαπλές εργασίες αλλά περιορίζεται σε μία αρχή και μία γλώσσα. Παρέχει μερικά βιβλιογραφικά δεδομένα αλλά όχι τα πλήρη XML αρχεία ή συμπληρωματικά media.

Η συλλογή WIPO-alpha [12], με περίπου 75,000 έγγραφα από το WIPO (1998-2002), είναι μια από τις παλαιότερες διαθέσιμες συλλογές. Ο μικρός της όγκος και η παλαιότητα των δεδομένων την καθιστούν λιγότερο σχετική για σύγχρονες εφαρμογές μηχανικής μάθησης που απαιτούν μεγάλα datasets.

Η WPI υπερτερεί σε πολλαπλά σημεία: είναι η μόνη συλλογή που καλύπτει και τις έξι μεγάλες αρχές διπλωμάτων ευρεσιτεχνίας, υποστηρίζει έξι γλώσσες συμπεριλαμβανομένων των Κορεατικών, Ιαπωνικών και Κινεζικών, παρέχει πλήρη XML δομή με όλα τα πεδία των διπλωμάτων, περιλαμβάνει

όλα τα συμπληρωματικά αρχεία (εικόνες, χημικές δομές), και διατηρεί τα αρχικά PDF αρχεία για κάθε δίπλωμα.

Ωστόσο, η WPI έχει και περιορισμούς. Η χρονική της κάλυψη (2014-2015) είναι περιορισμένη σε σύγκριση με άλλες συλλογές που καλύπτουν δεκαετίες. Επίσης, η αρχική έκδοση δεν παρείχε ground truths, ένα κενό που καλύπτει το WPI+ με τις επεκτάσεις που περιγράφονται σε αυτή την εργασία.

Κεφάλαιο 3ο: Επεξεργασία Πατεντών σε Επίπεδο Kind Code και Patent Family

3.1 Ανατομία ενός διπλώματος ευρεσιτεχνίας: Δομή XML και μεταδεδομένα

Τα σύγχρονα διπλώματα ευρεσιτεχνίας αποθηκεύονται και διανέμονται κυρίως σε μορφή XML, ακολουθώντας το πρότυπο ST.36 της World Intellectual Property Organization (WIPO) [13]. Αυτό το πρότυπο ορίζει μια ιεραρχική δομή που οργανώνει την πολύπλοκη πληροφορία ενός διπλώματος ευρεσιτεχνίας σε διακριτές ενότητες, διευκολύνοντας την αυτοματοποιημένη επεξεργασία και ανάλυση.

Η βασική δομή ενός XML εγγράφου διπλώματος ευρεσιτεχνίας αποτελείται από τέσσερις κύριες ενότητες. Η πρώτη ενότητα, bibliographic-data, περιέχει τα βιβλιογραφικά μεταδεδομένα του διπλώματος, συμπεριλαμβανομένων των στοιχείων δημοσίευσης (publication-reference), των στοιχείων αίτησης (application-reference), των δεδομένων προτεραιότητας (priority-claims), των ταξινομήσεων (classifications-ipcr, classifications-cpc), και των στοιχείων των εφευρετών και δικαιούχων. Αυτά τα μεταδεδομένα είναι κρίσιμα για την ταυτοποίηση, την αναζήτηση και την ταξινόμηση των διπλωμάτων.

Η δεύτερη ενότητα περιλαμβάνει το abstract, μια συνοπτική περιγραφή της εφεύρεσης που συνήθως δεν υπερβαίνει τις 150 λέξεις. Το abstract μπορεί να υπάρχει σε πολλαπλές γλώσσες, με κάθε έκδοση να χαρακτηρίζεται από το αντίστοιχο lang attribute. Η τρίτη ενότητα, description, αποτελεί τη λεπτομερή τεχνική περιγραφή της εφεύρεσης, συμπεριλαμβανομένων του τεχνικού πεδίου, του υποβάθρου της τέχνης, της περίληψης της εφεύρεσης, και της αναλυτικής περιγραφής των προτιμώμενων ενσωματώσεων. Η τέταρτη ενότητα, claims, περιέχει τις νομικές αξιώσεις που ορίζουν το εύρος της προστασίας του διπλώματος.

Πέρα από αυτές τις κύριες ενότητες, τα XML έγγραφα μπορεί να περιλαμβάνουν πρόσθετα στοιχεία όπως drawings-reference για αναφορές σε σχέδια, amended-claims για τροποποιημένες αξιώσεις, και copyright για πληροφορίες πνευματικών δικαιωμάτων. Κάθε στοιχείο μπορεί να έχει πολλαπλά attributes που παρέχουν πρόσθετες πληροφορίες, όπως το ucid (unique citation identifier) που αποτελεί μοναδικό αναγνωριστικό του εγγράφου.

Η πολυπλοκότητα αυτής της δομής επαυξάνεται από το γεγονός ότι διαφορετικές αρχές διπλωμάτων ευρεσιτεχνίας μπορεί να χρησιμοποιούν ελαφρώς διαφορετικές υλοποιήσεις του προτύπου ST.36. Για παράδειγμα, το USPTO μπορεί να συμπεριλαμβάνει πρόσθετα πεδία για συγκεκριμένες αμερικανικές απαιτήσεις, ενώ το EPO μπορεί να έχει διαφορετική δομή για τις αναφορές σε prior art. Αυτές οι διαφορές καθιστούν αναγκαία την ανάπτυξη ευέλικτων συστημάτων επεξεργασίας που μπορούν να χειριστούν τις ιδιαιτερότητες κάθε αρχής.

3.2 Το σύστημα Kind Codes και η σημασία τους

Τα kind codes αποτελούν θεμελιώδες στοιχείο του διεθνούς συστήματος διπλωμάτων ευρεσιτεχνίας, παρέχοντας τυποποιημένη πληροφορία σχετικά με το στάδιο και τον τύπο κάθε δημοσίευσης. Αποτελούνται από έναν ή δύο χαρακτήρες (συνήθως γράμμα και αριθμό) που προστίθενται στον αριθμό του διπλώματος για να δημιουργήσουν έναν μοναδικό προσδιοριστή για κάθε έκδοση του εγγράφου [14].

3.2.1 Κατηγορίες Kind Codes (A1, A2, B1, B2, κλπ.)

Το σύστημα kind codes ακολουθεί μια ιεραρχική λογική όπου το πρώτο γράμμα υποδεικνύει τη βασική κατηγορία του εγγράφου. Τα έγγραφα κατηγορίας Α αντιπροσωπεύουν δημοσιευμένες αιτήσεις διπλωμάτων ευρεσιτεχνίας. Συγκεκριμένα, το A1 υποδηλώνει την πρώτη δημοσίευση μιας αίτησης με έκθεση έρευνας, το A2 την πρώτη δημοσίευση χωρίς έκθεση έρευνας, το A3 τη μεταγενέστερη δημοσίευση της έκθεσης έρευνας, και το A4 τη συμπληρωματική έκθεση έρευνας. Υπάρχουν επίσης τα A8 και A9 που υποδηλώνουν διορθώσεις στα βιβλιογραφικά δεδομένα και διορθώσεις στο πλήρες έγγραφο αντίστοιχα.

Τα έγγραφα κατηγορίας Β αντιπροσωπεύουν χορηγημένα διπλώματα ευρεσιτεχνίας. Το B1 είναι το πρώτο χορηγημένο δίπλωμα, το B2 το χορηγημένο δίπλωμα με τροποποιήσεις μετά από αντίρρηση ή επανεξέταση, το B3 το δίπλωμα με περιορισμένες αξιώσεις, και τα B8, B9 διορθώσεις στο χορηγημένο δίπλωμα. Η κατανομή των kind codes στη συλλογή WPI δείχνει ότι τα A1 και B2 είναι τα πιο συχνά, αντιπροσωπεύοντας περίπου το 63% και 31% των εγγράφων αντίστοιχα [1].

3.2.2 Η εξέλιξη ενός διπλώματος μέσω διαφορετικών Kind Codes

Η πορεία ενός διπλώματος ευρεσιτεχνίας από την αρχική κατάθεση μέχρι την τελική χορήγηση και πιθανές μεταγενέστερες τροποποιήσεις δημιουργεί μια χρονική ακολουθία εγγράφων με διαφορετικά kind codes. Μια τυπική εξέλιξη ξεκινά με τη δημοσίευση της αίτησης (A1 ή A2), ακολουθείται από πιθανές διορθώσεις ή συμπληρώσεις (A4, A8, A9), και καταλήγει στη χορήγηση του διπλώματος (B1). Μεταγενέστερες τροποποιήσεις μπορεί να οδηγήσουν σε νέες εκδόσεις (B2, B9).

Κάθε νέα έκδοση μπορεί να περιέχει σημαντικές αλλαγές στο περιεχόμενο του διπλώματος. Για παράδειγμα, οι αξιώσεις μπορεί να τροποποιηθούν για να ξεπεραστούν αντιρρήσεις του εξεταστή, η περιγραφή μπορεί να εμπλουτιστεί με πρόσθετα παραδείγματα, ή τα βιβλιογραφικά δεδομένα μπορεί να διορθωθούν. Αυτή η δυναμική φύση των διπλωμάτων ευρεσιτεχνίας δημιουργεί την ανάγκη για συστήματα που μπορούν να παρακολουθούν και να ενσωματώνουν όλες τις εκδόσεις για να παρέχουν μια ολοκληρωμένη εικόνα του διπλώματος.

3.3 Patent Families: Ορισμός και σημασία

Μια patent family αποτελείται από ένα σύνολο διπλωμάτων ευρεσιτεχνίας που σχετίζονται μεταξύ τους μέσω κοινών αιτήσεων προτεραιότητας. Όταν μια εφεύρεση κατατίθεται για προστασία σε πολλαπλές χώρες ή περιοχές, δημιουργείται μια οικογένεια διπλωμάτων που περιλαμβάνει όλες τις σχετικές αιτήσεις και χορηγήσεις. Η έννοια της patent family είναι κρίσιμη για την κατανόηση του παγκόσμιου τοπίου προστασίας μιας εφεύρεσης και για την αποφυγή επικαλύψεων στην ανάλυση διπλωμάτων ευρεσιτεχνίας.

Υπάρχουν διάφοροι ορισμοί για τις patent families, με τους πιο κοινούς να είναι η simple family (όλα τα έγγραφα μοιράζονται ακριβώς τις ίδιες προτεραιότητες) και η extended family (τα έγγραφα μοιράζονται τουλάχιστον μία κοινή προτεραιότητα). Στο πλαίσιο της συλλογής WPI, η πληροφορία για τις patent families παρέχεται μέσω των priority-claims στα βιβλιογραφικά δεδομένα κάθε εγγράφου.

Η επεξεργασία σε επίπεδο patent family επιτρέπει την αποφυγή της υπερεκπροσώπησης εφευρέσεων που έχουν κατατεθεί σε πολλαπλά γραφεία πατεντών. Για παράδειγμα, μια σημαντική εφεύρεση μπορεί να έχει αιτήσεις σε 20 ή περισσότερες χώρες, δημιουργώντας δεκάδες έγγραφα στη βάση δεδομένων.

Χωρίς την ομαδοποίηση σε επίπεδο family, αυτή η εφεύρεση θα είχε δυσανάλογη βαρύτητα σε στατιστικές αναλύσεις ή μοντέλα μηχανικής μάθησης.

3.4 Σύγκριση επεξεργασίας σε επίπεδο Kind Code vs Patent Family

Η επιλογή μεταξύ επεξεργασίας σε επίπεδο kind code και επεξεργασίας σε επίπεδο patent family εξαρτάται από τους στόχους της ανάλυσης και τις απαιτήσεις της εφαρμογής. Κάθε προσέγγιση έχει διακριτά πλεονεκτήματα και περιορισμούς που πρέπει να λαμβάνονται υπόψη.

Η επεξεργασία σε επίπεδο kind code παρέχει τη μέγιστη λεπτομέρεια και επιτρέπει την παρακολούθηση της εξέλιξης ενός διπλώματος ευρεσιτεχνίας στο χρόνο. Είναι απαραίτητη για εφαρμογές που απαιτούν πρόσβαση σε συγκεκριμένες εκδόσεις, όπως η νομική ανάλυση που πρέπει να γνωρίζει τις ακριβείς αξιώσεις που ίσχυαν σε μια συγκεκριμένη ημερομηνία. Επίσης, επιτρέπει τη μελέτη των διαδικασιών εξέτασης και των αλλαγών που επιβάλλονται από τους εξεταστές. Ωστόσο, αυτή η προσέγγιση οδηγεί σε σημαντική πολυπλοκότητα λόγω του μεγάλου αριθμού εγγράφων και της ανάγκης διαχείρισης των σχέσεων μεταξύ τους.

Η επεξεργασία σε επίπεδο patent family απλοποιεί σημαντικά την ανάλυση ομαδοποιώντας σχετικά έγγραφα και αποφεύγοντας επικαλύψεις. Είναι ιδανική για στατιστικές αναλύσεις τεχνολογικών τάσεων, για τη δημιουργία datasets για μηχανική μάθηση όπου η ανεξαρτησία των δειγμάτων είναι σημαντική, και για cross-lingual retrieval όπου διαφορετικές γλωσσικές εκδόσεις της ίδιας εφεύρεσης πρέπει να αντιμετωπίζονται ως ενότητα. Ο περιορισμός αυτής της προσέγγισης είναι ότι μπορεί να χάσει σημαντικές λεπτομέρειες και διαφοροποιήσεις μεταξύ των μελών της οικογένειας.

3.5 Η ανάγκη για Virtual Patents

Η έννοια των Virtual Patents αναδύεται ως λύση στο δίλημμα μεταξύ της λεπτομέρειας του επιπέδου kind code και της απλότητας του επιπέδου patent family. Ένα Virtual Patent είναι ένα συνθετικό έγγραφο που δημιουργείται συνδυάζοντας την πλέον επικαιροποιημένη και πλήρη πληροφορία από όλες τις διαθέσιμες εκδόσεις (kind codes) ενός συγκεκριμένου διπλώματος ευρεσιτεχνίας, διατηρώντας παράλληλα την ιχνηλασιμότητα της προέλευσης κάθε στοιχείου [3].

Η ανάγκη για Virtual Patents προκύπτει από πρακτικές προκλήσεις που αντιμετωπίζουν οι ερευνητές και οι επαγγελματίες. Πρώτον, η επιλογή της "σωστής" έκδοσης ενός διπλώματος για ανάλυση είναι συχνά αυθαίρετη και μπορεί να οδηγήσει σε ασυνέπειες μεταξύ διαφορετικών μελετών. Δεύτερον, σημαντική πληροφορία μπορεί να είναι διασκορπισμένη σε πολλαπλές εκδόσεις, με κάποιες εκδόσεις να περιέχουν μόνο τις αλλαγές και όχι το πλήρες κείμενο. Τρίτον, ο όγκος των δεδομένων και η υπολογιστική επιβάρυνση από την επεξεργασία πολλαπλών εκδόσεων καθιστούν δύσκολη την κλιμάκωση αυτών των αναλύσεων.

Οι Virtual Patents αντιμετωπίζουν αυτές τις προκλήσεις παρέχοντας ένα ενιαίο, ολοκληρωμένο έγγραφο που αντιπροσωπεύει την τρέχουσα κατάσταση του διπλώματος. Χρησιμοποιώντας priority-based selection, το σύστημα επιλέγει αυτόματα την πλέον κατάλληλη πληροφορία από κάθε διαθέσιμη έκδοση, με τα έγγραφα κατηγορίας B να έχουν προτεραιότητα έναντι των εγγράφων κατηγορίας A, και τα νεότερα έγγραφα να προτιμώνται από τα παλαιότερα εντός της ίδιας κατηγορίας. Αυτή η προσέγγιση διασφαλίζει ότι το Virtual Patent περιέχει τις πλέον επικαιροποιημένες και νομικά έγκυρες πληροφορίες.

Η υλοποίηση των Virtual Patents στο πλαίσιο του WPI+ μέσω του συστήματος PatentFusion αποτελεί σημαντική συνεισφορά στην απλοποίηση και τυποποίηση της έρευνας στα διπλώματα ευρεσιτεχνίας. Μειώνει τον αριθμό των εγγράφων που πρέπει να επεξεργαστούν κατά περίπου 22-35% διατηρώντας παράλληλα την πληρότητα της πληροφορίας, όπως θα αναλυθεί λεπτομερώς στα επόμενα κεφάλαια.

Κεφάλαιο 4ο: Το Σύστημα PatentFusion - Αρχιτεκτονική και Υλοποίηση

4.1 Επισκόπηση της αρχιτεκτονικής του συστήματος

Το PatentFusion αποτελεί ένα καινοτόμο σύστημα υψηλής απόδοσης για την επεξεργασία εικονικών διπλωμάτων ευρεσιτεχνίας (virtual patents) που αντιμετωπίζει θεμελιώδεις προκλήσεις στην μεγάλης κλίμακας επεξεργασία των δεδομένων των διπλωμάτων αυτών. Η αρχιτεκτονική του συστήματος βασίζεται σε σύγχρονες αρχές κατανεμημένων συστημάτων (distributed systems) και επεξεργασία ροής δεδομένων (streaming data processing), όπως περιγράφονται στη βιβλιογραφία για επεξεργασία ροής μεγάλων δεδομένων (big data stream processing) [15].

Θεωρητικό Υπόβαθρο Streaming Architecture: Η streaming αρχιτεκτονική που υιοθετεί το PatentFusion ακολουθεί τις σύγχρονες αρχές που παρουσιάζονται σε νέες και σύγχρονες έρευνας περί big data stream processing systems. Σύμφωνα με τη σχετική βιβλιογραφία [16], τα streaming systems πρέπει να διαχειρίζονται αποδοτικά τρεις κρίσιμες διαστάσεις: την out-of-order επεξεργασία δεδομένων, τη διαχείριση κατάστασης (state management), και την ανοχή στα σφάλματα (fault tolerance). Το PatentFusion υιοθετεί αυτές τις αρχές προσαρμόζοντάς τες στις ιδιαίτερες απαιτήσεις της πληροφορίας των πατεντών .

Αρθρωτό Πρότυπο Σχεδιασμού (Modular Design Pattern): Η αρχιτεκτονική του συστήματος ακολουθεί το μοντέλο του αρθρωτού σχεδιασμού, αποτελούμενη από εννέα κύρια modules που συνεργάζονται για την επίτευξη της λειτουργικότητας του συστήματος. Αυτή η προσέγγιση παίρνει έμπνευση από τις αρχές των αρχιτεκτονικών μικροϋπηρεσιών (microservices architectures) και επιτρέπει την ανεξάρτητη ανάπτυξη, δοκιμή, και βελτιστοποίηση κάθε component [17].

Κύρια Modules του Συστήματος:

1. **Configuration Manager**: Κεντρική διαχείριση παραμετροποίησης και επικύρωση συστήματος
2. **XML Parser**: Δημιουργία εικονικών πατεντών με διατήρηση της πλήρους XML ιεραρχίας
3. **File System**: Αποδοτικός εντοπισμός και διαχείριση αρχείων με στατιστική ανάλυση
4. **Data Processor**: Μετα-επεξεργασία εικονικών διπλωμάτων ευρεσιτεχνίας
5. **Memory Manager**: Streaming multiprocessing με αποδοτική διαχείριση μνήμης
6. **Output Manager**: Παραγωγή πολλαπλών μορφών εξόδου με παράλληλη επεξεργασία
7. **Parallel Processor**: Συντονισμός πολυεπεξεργασίας με XML serialization
8. **Utils**: Κοινόχρηστα εργαλεία και βοηθητικές λειτουργίες συστήματος
9. **Constants**: Κοινόχρηστες σταθερές και προεπιλογές παραμετροποίησης

Το Παράδειγμα Πολυεπεξεργασίας Ροής (Streaming Multiprocessing Paradigm): Το σύστημα έχει σχεδιαστεί με βάση τις αρχές του streaming multiprocessing, επιτρέποντας την επεξεργασία απεριόριστου μεγέθους datasets χωρίς περιορισμούς μνήμης. Αυτή η προσέγγιση αντιμετωπίζει τη θεμελιώδη πρόκληση των big data systems όπως περιγράφεται στη σύγχρονη βιβλιογραφία: την ανάγκη για κλιμακούμενες αρχιτεκτονικές (scalable architectures) που μπορούν να επεξεργαστούν μεγάλους όγκους δεδομένων με σταθερή απόδοση και χαμηλή χρήση μνήμης [18].

Καινοτομίες στην Επεξεργασία Δεδομένων Πατεντών: Η επεξεργασία δεδομένων διπλωμάτων ευρεσιτεχνίας παρουσιάζει μοναδικές προκλήσεις που το PatentFusion αντιμετωπίζει μέσω 2 στρατηγικών, της έξυπνης αναγνώρισης διπλότυπων πεδίων (intelligent duplicate field detection) και της στρατηγικής συγχώνευσης μέσω προτεραιοτήτων (priority-based merging). Σύμφωνα με τις νέες έρευνες στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing) στο τομέα διπλωμάτων ευρεσιτεχνίας (patent domain) [19], η αποδοτική επεξεργασία δεδομένων πατεντών απαιτεί εξειδικευμένες τεχνικές που λαμβάνουν υπόψη τη γλωσσική ποικιλομορφία, την τεχνική πολυπλοκότητα, και τη χρονική εξέλιξη των εγγράφων. Έτσι το σύστημα που έχει δημιουργηθεί, απαντά σε όλες τις παραπάνω απαιτήσεις.

4.2 Βασικά modules και η λειτουργία τους

4.2.1 Configuration Manager

Το Configuration Manager (config_manager.py) αποτελεί το κεντρικό σημείο διαχείρισης της παραμετροποίησης του συστήματος, υιοθετώντας τις αρχές της κεντρικοποιημένης διαχείρισης (centralized configuration management) που αποτελούν βασική προϋπόθεση για την υλοποίηση επεκτάσιμων και κατανεμημένων συστημάτων. Η σημασία της κεντρικής διαχείρισης παραμετροποίησης έχει τεκμηριωθεί στη βιβλιογραφία ως κρίσιμος παράγοντας για την συντήρηση και αξιοπιστία σύγχρονων και σύνθετων συστημάτων [15].

Αρχιτεκτονική Παραμετροποίησης: Η σχεδίαση του Configuration Manager ακολουθεί το pattern της ιεραρχικής διαχείρισης παραμέτρων, επιτρέποντας την οργάνωση των ρυθμίσεων σε λογικές ομάδες που αντιστοιχούν στις διαφορετικές λειτουργικές ενότητες του συστήματος. Αυτή η προσέγγιση διευκολύνει την κατανόηση και διαχείριση της πολυπλοκότητας που προκύπτει από τη μεγάλη ποικιλία παραμέτρων που απαιτεί ένα σύστημα επεξεργασίας δεδομένων πατεντών.

Κύριες Λειτουργίες και Καινοτομίες:

- **Ευφυής επικύρωση ρυθμίσεων**: Αυτόματη επικύρωση της συνοχής των παραμέτρων με δυναμικούς ελέγχους βάσει των δυνατοτήτων του συστήματος
- **Προσαρμοστικός υπολογισμός παραμέτρων**: Αυτόματος υπολογισμός των παραμέτρων των λειτουργιών που προσαρμόζονται δυναμικά στο hardware
- **Ρυθμίσεις εξαγωγής σε πολλαπλά format**: Ευέλικτη διαχείριση πολλαπλών μορφών εξόδου. Υποστήριξη για XML, CSV, JSON
- **Ρυθμίσεις με γνώμονα την απόδοση**: Αυτόματη βελτιστοποίηση παραμέτρων απόδοσης βάσει διαθέσιμων πόρων του συστήματος

Τεχνική Υλοποίηση και Βελτιστοποιήσεις: Το module αξιοποιεί το configparser library της Python, επεκτείνοντας τη βασική λειτουργικότητα με εξειδικευμένους validators και adaptive calculators. Η προσέγγιση αυτή εξασφαλίζει ότι το σύστημα μπορεί να λειτουργεί αποδοτικά σε διαφορετικά περιβάλλοντα hardware χωρίς να απαιτείται η χειροκίνητη επιλογή παραμέτρων, ένα χαρακτηριστικό που είναι ιδιαίτερα σημαντικό για τη διαχείριση μεγάλης κλίμακας datasets πατεντών.

4.2.2 XML Parser και η διατήρηση της ιεραρχίας XML

Το XML Parser (xml_parser.py) αποτελεί τον πυρήνα του συστήματος PatentFusion και εμπεριέχει τις πιο σύνθετες αλγοριθμικές καινοτομίες της πλατφόρμας. Η σχεδίαση του βασίζεται σε αποδεδειγμένες

αρχές για την επεξεργασία μεγάλης κλίμακας XML δεδομένων και την αποδοτική διαχείριση πολύπλοκων structured documents [16].

Θεμελιώδεις Αρχές XML Processing: Η επεξεργασία μεγάλων XML datasets παρουσιάζει ιδιαίτερες τεχνικές προκλήσεις που σχετίζονται με τη διαχείριση μνήμης, την απόδοση της parsing τεχνικής, και τη διατήρηση δομικής ακεραιότητας. Το PatentFusion αντιμετωπίζει αυτές τις προκλήσεις μέσω της χρήσης της βιβλιοθήκης lxml, η οποία παρέχει απόδοση επιπέδου γλώσσας C με την ευελιξία της γλώσσας Python, επιτρέποντας την αποδοτική επεξεργασία εκατομμυρίων XML elements χωρίς σημαντική υποβάθμιση απόδοσης.

Αρχιτεκτονική Δημιουργίας Virtual Patent: Η δημιουργία εικονικών διπλωμάτων ευρεσιτεχνίας αντιπροσωπεύει μια καινοτόμα προσέγγιση στην ενοποίηση δεδομένων πατεντών που λαμβάνει υπόψη τόσο τη χρονική εξέλιξη των δικαιωμάτων όσο και τη γλωσσική ποικιλομορφία περιεχομένου των διπλωμάτων. Η στρατηγική του priority-based merging που εφαρμόζεται, αντικατοπτρίζει την πραγματική ιεραρχία της σημασίας των διαφόρων kind codes στο διεθνές σύστημα πατεντών.

Κύριες Αλγοριθμικές Συνεισφορές:

- **Ευφυής ομαδοποίηση**: Προηγμένη ομαδοποίηση των αρχείων βάσει patent identifiers με χειρισμός σφαλμάτων για ονόματα αρχείων που περιέχουν λάθη
- **Επεξεργασία βάσει προτεραιότητας**: Προεπιλεγμένη ιεραρχική ταξινόμηση βάσει προεπιλεγμένης προτεραιότητας kind codes (B9 > B8 > ... > A1 > A)
- **Επεξεργασία βάσει προτεραιότητας που μπορεί να διαμορφωθεί από τον χρήστη**: Ιεραρχική ταξινόμηση με βάση την ιεραρχία που έχει επιλέξει ο χρήστης-ερευνητής, πχ (A4>A1>B4>B1)
- **Συγχώνευση με διατήρηση δομής**: Διατήρηση πλήρους XML ιεραρχίας με πεδία που μπορεί να επιλέξει προς διατήρηση ή μη, ο χρήστης-ερευνητής
- **Μετασχηματισμός Attribute**: Ελεγχόμενος μετασχηματισμός ucid attributes για την εύκολη αναγνώριση των virtual patent
- **Βελτίωση μεταδεδομένων**: Εμπλουτισμός μέσω της εισαγωγής του kind-source attribute για πλήρη ιχνηλασιμότητα της προέλευσης της πληροφορίας

Πολυεπίπεδος Αλγόριθμος Δημιουργίας:

```
Επίπεδο 1: group_files_by_patent() - Ταξινομική οργάνωση των αρχείων

Επίπεδο 2: sort_files_by_priority() – Ιεραρχική κατανομή βάσει της
νομικής προτεραιότητας

Επίπεδο 3: create_virtual_patent() – Ευφυής συγχώνευση με διατήρηση της
δομής των πατεντών

Επίπεδο 4: transform_ucid_attributes() – Σημασιολογική μετατροπή για την
αναγνώριση των Virtual Patents

Επίπεδο 5: update_kind_to_kind_merging() - Βελτίωση μεταδεδομένων για
ιχνηλασιμότητα μέσω της μετατροπής του kind σε kind_merging

Επίπεδο 6: apply_config_filtering() – Επιπλέον προσαρμοζόμενες επιλογές
περιεχομένου που ορίζονται από τον χρήστη-ερευνητή
```

Τεχνολογικά Πλεονεκτήματα της βιβλιοθήκης lxml: Η επιλογή της βιβλιοθήκης lxml μόνο τυχαία δεν μπορεί να χαρακτηριστεί, καθώς βασίζεται σε τεκμηριωμένα τεχνικά πλεονεκτήματα: παρέχει απόδοση native C κώδικα μέσω της libxml2, υποστηρίζει XPath expressions για αποδοτική επιλογή των elements, και διατηρεί namespace awareness που είναι κρίσιμο για τη σωστή επεξεργασία των πολύπλοκων σχημάτων (schemes) XML των πατεντών. Επιπλέον, η lxml επιτρέπει παράλληλη επεξεργασία με Global Interpreter Lock (GIL) release, ένα χαρακτηριστικό που αξιοποιείται εκτεταμένα στην αρχιτεκτονική παράλληλης επεξεργασίας του PatentFusion, παρακάμπτοντας έτσι ένα σημαντικό μειονέκτημα γλωσσών διερμηνείας (interpreted languages) όπως η Python.

4.2.3 File System και διαχείριση αρχείων

Το File System module (file_system.py) διαχειρίζεται όλες τις λειτουργίες που σχετίζονται με τη διαχείριση αρχείων και καταλόγων.

Κύριες Λειτουργίες:

- Αναδρομικός εντοπισμός αρχείων διπλωμάτων ευρεσιτεχνίας (.xml)
- Δημιουργία δομής καταλόγων για μεμονωμένα εικονικά διπλώματα
- Διαχείριση προσωρινών αρχείων XML με άμεσο καθαρισμό τους
- Επικύρωση αρχείων και συλλογή στατιστικών

Τεχνική Υλοποίηση:

- Χρήση os.walk() για αναδρομική σάρωση καταλόγων
- Υπολογισμός μεγεθών αρχείων και συλλογή μετρικών
- Δημιουργία οργανωμένης δομής εξόδου (office/format/files)
- Άμεσος καθαρισμός temp files μετά την επεξεργασία για εξοικονόμηση χώρου

4.2.4 Data Processor

Το Data Processor module (data_processor.py) αναλαμβάνει τη μετα-επεξεργασία των εικονικών διπλωμάτων ευρεσιτεχνίας.

Κύριες Λειτουργίες:

- Απλή επικύρωση εικονικών διπλωμάτων
- Επεξεργασία XML elements
- Αποδοτική διαχείριση συλλογών εικονικών διπλωμάτων
- Βελτιστοποίηση garbage collection

Αλγόριθμος Μετα-επεξεργασίας:

```
def post_process_virtual_patents(virtual_patents, config):
    processed_patents = []
    for virtual_patent in virtual_patents:
        # Εφαρμογή φίλτρων βάσει παραμετροποίησης
        processed_patents.append(virtual_patent)
    gc.collect()  # Βελτιστοποίηση μνήμης
    return processed_patents
```

4.2.5 Memory Manager και streaming multiprocessing

Το Memory Manager (memory_manager.py) αποτελεί την πιο καινοτόμα συνεισφορά του PatentFusion στον τομέα της επεξεργασίας μεγάλης κλίμακας δεδομένων, υλοποιώντας μια προηγμένη streaming multiprocessing αρχιτεκτονική που επιλύει το θεμελιώδες πρόβλημα του scalability στην επεξεργασία μεγάλης κλίμακας patent datasets. Η προσέγγιση αυτή εμπνέεται από τις αρχές των μοντέρνων συστημάτων streaming [18] και προσαρμόζεται στις ιδιαίτερες απαιτήσεις της επεξεργασίας των πατεντών.

Θεωρητικό Πλαίσιο Streaming Architecture: Η streaming αρχιτεκτονική του PatentFusion βασίζεται στην αρχή της επεξεργασίας με περιορισμένη μνήμη (bounded memory processing), όπου η χρήση μνήμης παραμένει σταθερή ανεξάρτητα από το μέγεθος του dataset. Αυτό επιτυγχάνεται μέσω της εφαρμογής της αρχής του «ενός αρχείου για κάθε εργάτη» "one file per worker", που εξασφαλίζει ότι κάθε worker process διατηρεί στη μνήμη μόνο τα δεδομένα ενός προσωρινού αρχείου (temporary file) κάθε φορά, με αποτέλεσμα την πολύ μεγάλη εξοικονόμηση μνήμης που καθιστά ακόμη και οικιακά συστήματα υπολογιστών, ικανά για την επεξεργασία πολύ μεγάλων datasets.

Καινοτόμες Τεχνικές Διαχείρισης Μνήμης:

- **Προβλέψιμη χρήση μνήμης**: Εγγυημένη σταθερή χρήση μνήμης με σαφώς καθορισμένα όρια
- **Επιθετική στρατηγική καθαρισμού**: Άμεσος καθαρισμός προσωρινών αρχείων με εξαναγκασμένο garbage collection για την όσο τον δυνατόν περισσότερη διαθέσιμη μνήμη
- **Αλγόριθμος Load Balancing**: Δυναμική κατανομή εργασίας βάσει των διαθέσιμων πυρήνων του επεξεργαστή (CPU) αλλά και της πίεσης στην διαθέσιμη μνήμη
- **Ανίχνευση πίεσης μνήμης**: Προληπτική ανίχνευση μεγάλης χρήσης μνήμης με αυτόματη διαδικασία για τον καθαρισμό της και την απόδοση στο σύστημα
- **Παραλληλοποίηση Παρακολούθησης Προόδου**: Παρακολούθηση προόδου των παράλληλων διεργασιών επεξεργασίας σε παράλληλο επίπεδο και με ελάχιστο κόστος χρήσης μνήμης

Αρχιτεκτονική Επεξεργασίας σε τμήματα: Η τεχνική της επεξεργασίας σε τμήματα (chunked processing) που εφαρμόζεται στο PatentFusion αντιπροσωπεύει μια σύνθεση των καλύτερων πρακτικών από τη βιβλιογραφία των διανεμημένων συστημάτων (distributed systems), προσαρμοσμένη στις ανάγκες της επεξεργασίας δομημένων εγγράφων (structured documents) όπως τα έγγραφα πατεντών:

Αλγόριθμος: Προσαρμοστική στρατηγική τμηματοποίησης

```
1. Υπολογισμός τμημάτων με γνώμονα τη μνήμη: O(log n) πολυπλοκότητα
2. Ισορροπημένη κατανομή CPU: ανάθεση O(1) ανά worker (CPU core)
3. Διαδοχική επεξεργασία αρχείων: μνήμη O(k) ανά worker (k = αρχεία ανά τμήμα)
4. Άμεση εκκαθάριση: πολυπλοκότητα χώρου O(1)
5. Συγχρονισμός προόδου: επικοινωνία O(log w) (w = workers)
```

Επιστημονικές Συνεισφορές στη Διαχείριση της Μνήμης: Η διαχείριση μνήμης του PatentFusion εισάγει τρεις σημαντικές καινοτομίες στον τομέα της επεξεργασίας δεδομένων μεγάλης κλίμακας με αποδοτική χρήση μνήμης:

1. **Streaming μνήμης με όρια**: Μαθηματικά εγγυημένη σταθερή χρήση μνήμης για απεριόριστο μέγεθος των datasets
2. **Προσαρμοστικό Load Balancing**: Δυναμική αναδιανομή της εργασίας βάσει μετρικών συστήματος πραγματικού χρόνου
3. **Ιεραρχική παρακολούθηση Προόδου**: Παρακολούθηση εργασιών προόδου χωρίς σημαντική πίεση μνήμης

Τεστ Απόδοσης και Scalability Metrics: Η αρχιτεκτονική έχει δοκιμαστεί σε εκτεταμένα datasets που φτάνουν τα 100GB+ προσωρινών αρχείων και περισσότερα από 150GB πρωτογενών αρχείων πατεντών, αποδεικνύοντας την γραμμική επεκτασιμότητα με σταθερή χρήση μνήμης - ένα επίτευγμα που σπάνια συναντάται σε παραδοσιακά συστήματα batch processing.

4.2.6 Output Manager και υποστήριξη πολλαπλών formats

Το Output Manager (output_manager.py) είναι υπεύθυνο για την παραγωγή εξόδου των αρχείων των virtual patents σε πολλαπλές μορφές.

Υποστηριζόμενες Μορφές:

- **XML (hierarchical)**: Διατήρηση πλήρους ιεραρχίας των XML αρχείων
- **CSV (flat)**: Επίπεδη δομή με indexed fields
- **JSON (hierarchical)**: Ιεραρχική δομή σε JSON format

Κύριες Λειτουργίες:

- Δημιουργία μεμονωμένων αρχείων εικονικών διπλωμάτων ευρεσιτεχνίας
- Επιλογή εξόδου βάσει παραμετροποίησης
- Παράλληλη παραγωγή αρχείων εξόδου με XML serialization
- Οργανωμένη δομή καταλόγων (office/format/files)
- Ενσωματωμένη επιθεώρηση των συγχωνευμένων πατεντών κατά την αποθήκευση

Προηγμένο Language Filtering:

```
- parse_lang = ALL: Διατήρηση όλων των γλωσσικών εκδόσεων
- parse_lang = PRIMARY: Διατήρηση μόνο της κύριας γλώσσας
- parse_lang = EN: Διατήρηση μόνο του αγγλικού περιεχομένου
- parse_lang = EN,FR,DE: Διατήρηση συγκεκριμένων γλωσσών
```

4.2.7 Parallel Processor

Το Parallel Processor module (parallel_processor.py) συντονίζει την πολυεπεξεργασία για την βέλτιστη απόδοση του συστήματος δημιουργίας των Virtual Patents.

Κύριες Λειτουργίες:

- Παράλληλη batch επεξεργασία αρχείων διπλωμάτων ευρεσιτεχνιών
- Δημιουργία εικονικών διπλωμάτων XML στους επιλεγμένους/διαθέσιμους workers (CPU cores)
- Παρακολούθηση προόδου ανά worker
- XML serialization για συμβατότητα με multiprocessing για την πλήρη εκμετάλευση των διαθέσιμων resources του συστήματος

Αλγόριθμος Παραλληλισμού:

```
1. Διαίρεση αρχείων σε batches
2. Δημιουργία worker processes ανά διαθέσιμο ή επιλεγμένο αριθμό CPU
cores
3. Διανομή batches στους workers για παράλληλη επεξεργασία
4. Συλλογή και συνένωση αποτελεσμάτων
5. Παρακολούθηση προόδου με shared memory
```

4.3 Αλγόριθμος δημιουργίας Virtual Patents

4.3.1 Ομαδοποίηση αρχείων ανά patent number

Το πρώτο βήμα στη δημιουργία εικονικών διπλωμάτων ευρεσιτεχνιών είναι η ομαδοποίηση των αρχείων XML βάσει του patent number που εξάγεται από το όνομα του αρχείου.

Αλγόριθμος:

```
def group_files_by_patent(file_batch, test_patents_set=None):
    patent_groups = {}
    for file_path in file_batch:
        file_name = os.path.basename(file_path)
        patent_number = file_name.split(".")[0].split("-")[1]

        if patent_number not in patent_groups:
            patent_groups[patent_number] = []
        patent_groups[patent_number].append(file_path)

    return patent_groups
```

Τεχνικές Λεπτομέρειες:

- Εξαγωγή patent number από filename format: "EP-1234567-A1.xml"
- Δημιουργία dictionary με patent_number ως κλειδιού (key)
- Υποστήριξη test dataset exclusion για validation

4.3.2 Priority-based selection (B9>B8>...>A1>A)

Μετά την ομαδοποίηση, τα αρχεία ταξινομούνται βάσει της προτεραιότητας του kind code τους.
Default Global Priority List:

```
B9 > B8 > B6 > B3 > B2 > B1 > B > A9 > A8 > A6 > A5 > A4 > A3 > A2 > A1 > A
```

Αλγόριθμος Ταξινόμησης:

```
def sort_files_by_priority(file_list, global_priority):
    def get_priority_index(file_path):
        kind_code = extract_kind_code_from_file(file_path)
        try:
            return global_priority.index(kind_code)
        except ValueError:
            return len(global_priority)  # Χαμηλότερη προτεραιότητα

    return sorted(file_list, key=get_priority_index)
```

Ο αλγόριθμος βρίσκει για κάθε μοναδική πατέντα σύμφωνα με το patent number το αρχείο που έχει kind code με την υψηλότερη προτεραιότητα, και το ορίζει ως πρωτεύον. Αμέσως μετά βρίσκει το αρχείο με kind code με χαμηλότερη προτεραιότητα από πριν και το βάζει ως 2^{ο}, και ούτω καθεξής έως ότου εξαντλήσει τα kind codes που δόθηκαν. Η παραπάνω default τιμή είναι αυτή που έχει επιλεγεί με την λογική ότι κρατάμε κατά το δυνατόν τα νεότερα δεδομένα των πατεντών για την δημιουργία των VP προσθέτοντας πληροφορία από τις παλαιότερες μόνο αν οι προηγούμενες δεν είχαν αυτήν την πληροφορία. Το αποτέλεσμα θα είναι μια πληρέστερη VP που εμπεριέχει δεδομένα που μπορεί να ήταν διάσπαρτα σε 2-3 ή και παραπάνω αρχεία πατεντών. Παρόλα αυτά ο ερευνητής έχει την δυνατότητα να ορίσει οποιαδήποτε προτεραιότητα θέλει. Πχ A4 > A2 > A1 > B4 > B6 > B1.

4.3.3 Στρατηγική συγχώνευσης (merging strategy)

Το PatentFusion υλοποιεί μια ενοποιημένη στρατηγική συγχώνευσης (Unified Merging Strategy) με επιλεκτική ανίχνευση διπλότυπων που διαχειρίζεται διαφορετικά τα Level 1 elements.

Στρατηγική Ανά τύπο Element:

Bibliographic-data elements:

- Διατήρηση από το υψηλότερης προτεραιότητας δίπλωμα
- Στα elements επιπέδου 2 (publication-reference, application-reference κ.λπ.): Εκτελείται μια έξυπνη ιεραρχική συγχώνευση όπου διατηρούνται τα στοιχεία υψηλότερης προτεραιότητας, ενώ προθέτονται στοιχεία επιπέδου 2 και επιπέδου 3 που λείπουν από πρόσθετα διπλώματα ευρεσιτεχνίας.
- Συγχώνευση επιπέδου 3: Προσθέτει τα στοιχεία επιπέδου 3 που λείπουν εντός των κοντέινερ επιπέδου 2 (όπως, invention-title, citations within technical-data)

Abstract elements:

- Ευφυείς ανίχνευση διπλότυπων elements βάσει language, source, και περιεχόμενου κειμένου
- Διατήρηση μοναδικών abstracts από όλα τα priority levels
- Αποφυγή πραγματικών διπλότυπων elements

Άλλα Level 1 elements (description, claims, copyright):

- Πλήρη elements από το υψηλότερης προτεραιότητας δίπλωμα
- Προσθήκη μόνο elements που δεν έχουν ήδη βρεθεί

Αλγόριθμος Συγχώνευσης:

```
def merge_element_recursive(base_element, additional_element, config,
kind_code):
    for additional_child in additional_element:
        tag_name = additional_child.tag

        # Έλεγχος για duplicate
        if tag_name == 'abstract':
            # Intelligent duplicate detection για abstracts
            for base_child in base_children_list:
                if base_child.tag == tag_name:
                    if is_duplicate_element(base_child,
additional_child):
                        is_duplicate = True
                        break
        else:
            # Simple tag-based detection για άλλα elements
            for base_child in base_children_list:
                if base_child.tag == tag_name:
                    is_duplicate = True
                    break

        if not is_duplicate:
            # Προσθήκη στο base element
            base_element.append(copy.deepcopy(additional_child))
            add_kind_source_to_direct_children(additional_child,
kind_code)
```

4.3.4 Έξυπνη ανίχνευση διπλότυπων abstract

Το PatentFusion εισάγει μια πρωτοποριακή τεχνική ανίχνευσης διπλοτύπων για abstract elements που αντιπροσωπεύει σημαντική καινοτομία στον τομέα της επεξεργασίας του κειμένου των πατεντών. Αυτή η τεχνική βασίζεται σε ανάλυση πολλαπλών κριτηρίων που λαμβάνει υπόψη τόσο δομικά όσο και σημασιολογικά χαρακτηριστικά των abstracts, αντιμετωπίζοντας αποδοτικά τις προκλήσεις που προκύπτουν από τη γλωσσική ποικιλομορφία και τον πλεονασμό και την επανάληψη περιεχομένου στα δεδομένα πατεντών [19].

Θεωρητικό Πλαίσιο Ανάλυσης Πολλαπλών Κριτηρίων: Η προσέγγιση της ευφυής ανίχνευσης διπλοτύπων βασίζεται σε τρεις θεμελιώδεις διαστάσεις ανάλυσης που έχουν αναγνωριστεί στη βιβλιογραφία ως κρίσιμες για την αποδοτική επεξεργασία πολύγλωσσων τεχνικών κειμένων:

Πολυδιάστατα Κριτήρια Σύγκρισης:

1. **Γλωσσική διάσταση (lang attribute)**: Διάκριση γλωσσικών παραλλαγών με case-insensitive αντιστοίχιση
2. **Διάσταση προέλευσης (source attribute)**: Ταυτοποίηση πηγής περιεχομένου μέσω source και load-source attributes
3. **Διάσταση περιεχομένου (text content)**: Σημασιολογική ανάλυση πλήρους περιεχομένου κειμένου

Αλγοριθμική Καινοτομία στην Ανάλυση Περιεχομένου: Η τεχνική εξαγωγής και σύγκρισης του περιεχόμενου κειμένου που εφαρμόζεται υπερβαίνει τις παραδοσιακές προσεγγίσεις string matching, ενσωματώνοντας:

- **Ιεραρχική εξαγωγή κειμένου**: Αναδρομική συλλογή κειμένου από nested δομές XML
- **Ασαφής αντιστοίχιση περιεχομένου:** Ανίχνευση περικομμένου ή μερικώς τροποποιημένου περιεχομένου
- **Ομοιότητα βάσει κατωφλίου:** Προσαρμοστικά κατώφλια ομοιότητας για ισχυρή ανίχνευση διπλότυπων
- **Πολυγλωσσική ανοχή**: Γλωσσικά-ευαίσθητη σύγκριση με κανονικοποίηση Unicode

Αλγόριθμος Ανίχνευσης Διπλοτύπων:

```
def is_duplicate_element(element1, element2):
    # Σύγκριση lang attribute
    lang1 = element1.get('lang', '').lower()
    lang2 = element2.get('lang', '').lower()

# Σύγκριση source attribute
    source1 = element1.get('source', element1.get('load-source',
'')).lower()
    source2 = element2.get('source', element2.get('load-source',
'')).lower()

    # Εξαγωγή text content
    def get_element_text(elem):
        text_parts = []
        if elem.text and elem.text.strip():
            text_parts.append(elem.text.strip())
        for child in elem:
            if child.text and child.text.strip():
                text_parts.append(child.text.strip())
        return ' '.join(text_parts).strip()
    text1 = get_element_text(element1).lower()
    text2 = get_element_text(element2).lower()

    # Logic duplicate detection
    if lang1 and lang2 and lang1 != lang2:
        return False
    if source1 and source2 and source1 != source2:
        return False

    # Text similarity check
    if text1 == text2:
        return True

    # Truncated content detection
    if len(text1) > 50 and len(text2) > 50:
        shorter = text1 if len(text1) < len(text2) else text2
        longer = text2 if len(text1) < len(text2) else text1
        if shorter in longer and len(shorter) > len(longer) * 0.8:
            return True

    return False
```

Πλεονεκτήματα:

- Διατήρηση μοναδικού περιεχομένου από όλα τα priority levels
- Ανίχνευση πραγματικών διπλότυπων και απόρριψή τους
- Υποστήριξη πολύγλωσσου περιεχομένου
- Διαχείριση σεναρίων περικομμένου περιεχομένου

4.3.5 Kind-source attributes και ιχνηλασιμότητα

Το σύστημα εφαρμόζει διαφοροποιημένη στρατηγική kind-source με στόχο την πλήρη ιχνηλασιμότητα των merged elements. Εισαγάγει το attribute kind-source στις VP το οποίο καθιστά εύκολα ανιχνεύσιμη την προέλευση της κάθε πληροφορίας κάνοντας αναφορά στο kind code από το οποίο προήλθε. Έτσι γίνεται εξαιρετικά απλή η ιχνηλασιμότητα του κάθε πεδίου μιας virtual πανέντας.

Στρατηγική Ανά Element Type:

Bibliographic-data:

```
<bibliographic-data>
  <publication-reference kind-source="A4">
  <application-reference kind-source="A1">
</bibliographic-data>
```

Abstract, Description, Claims:

```
<abstract kind-source="A4">
<description kind-source="A4">
<claims kind-source="A1">
```

Copyright και άλλα childless elements:

```
<copyright kind-source="A4">
```

Τεχνική Υλοποίηση:

```
def add_kind_source_to_direct_children(xml_element, kind_code):
    for child in xml_element:
        tag_name = child.tag

        if tag_name == 'bibliographic-data':
            # Unified strategy: kind-source στα Level 2 children
            for level2_child in child:
                level2_child.set('kind-source', kind_code)
        else:
            # Differentiated strategy: kind-source στο Level 1 element
            child.set('kind-source', kind_code)
```

Οφέλη:

- Πλήρης ταυτοποίηση της πηγής κάθε merged element
- Βέλτιστη και λεπτομερής παρακολούθηση και ανίχνευση της κάθε πληροφορίας
- Εξισορρόπηση μεταξύ ανιχνευσιμότητας (traceability) και XML structure clarity
- Αποδοτικότερη περεταίρω επεξεργασία

4.4 Ψευδοκώδικας βασικών διαδικασιών

Κύρια Διαδικασία Επεξεργασίας:

```
ΔΙΑΔΙΚΑΣΙΑ PatentFusion_Main_Processing:
    1. ΦΟΡΤΩΣΗ configuration από config.ini
    2. ΕΚΚΙΝΗΣΗ logging και system monitoring

    3. FILE_DISCOVERY:
        ΣΑΡΩΣΗ vertical_origin_path για XML αρχεία
        ΣΥΛΛΟΓΗ στατιστικών και μετρικών
        ΔΗΜΙΟΥΡΓΙΑ file batches για parallel processing

    4. PARALLEL_PROCESSING:
        ΓΙΑ κάθε batch ΣΤΟ parallel:
            group_files_by_patent(batch)
            sort_files_by_priority(files, global_priority)
            create_virtual_patent(sorted_files)
        ΣΥΛΛΟΓΗ virtual patents από όλους τους workers

    5. MEMORY_EFFICIENT_OUTPUT:
        ΓΙΑ κάθε temp_file ΣΤΟ streaming:
            ΦΟΡΤΩΣΗ virtual patents από temp_file
            post_process_virtual_patents(patents)
            ΠΑΡΑΓΩΓΗ output σε XML, CSV, JSON formats
            ΔΙΑΓΡΑΦΗ temp_file αμέσως μετά την επεξεργασία

    6. CLEANUP_AND_REPORTING:
        ΑΝΑΦΟΡΑ timing statistics
        ΚΑΘΑΡΙΣΜΟΣ προσωρινών αρχείων
        REPORTING merged patents inspection
ΤΕΛΟΣ ΔΙΑΔΙΚΑΣΙΑΣ
```

Αλγόριθμος Δημιουργίας Virtual Patent:

```
ΔΙΑΔΙΚΑΣΙΑ create_virtual_patent(sorted_files, config):
    base_file = sorted_files[0]  // Υψηλότερη προτεραιότητα
    base_xml = ΦΟΡΤΩΣΗ_XML(base_file)
    base_kind_code = ΕΞΑΓΩΓΗ_KIND_CODE(base_file)

    // Προσθήκη kind-source στο base patent
    add_kind_source_to_direct_children(base_xml, base_kind_code)

    // Λίστα kind codes για merging tracking
    kind_codes = [base_kind_code]

    ΓΙΑ κάθε additional_file ΣΤΑ sorted_files[1:]:
        additional_xml = ΦΟΡΤΩΣΗ_XML(additional_file)
        additional_kind_code = ΕΞΑΓΩΓΗ_KIND_CODE(additional_file)

        // Συγχώνευση με unified strategy
        merge_xml_elements(base_xml, additional_xml, config,
additional_kind_code)

        kind_codes.ΠΡΟΣΘΗΚΗ(additional_kind_code)

    // Μετασχηματισμός σε virtual patent format
    transform_ucid_attributes(base_xml)  // Αλλαγή σε VP suffix
    update_kind_to_kind_merging(base_xml, kind_codes)  // kind="VP"
    reorder_xml_elements(base_xml)  // XML element ordering
    apply_config_filtering(base_xml, config)  // Configuration filtering

    ΕΠΙΣΤΡΟΦΗ base_xml
ΤΕΛΟΣ ΔΙΑΔΙΚΑΣΙΑΣ
```

Αλγόριθμος Παράλληλης Δημιουργίας Ροής Εξόδου:

```
ΔΙΑΔΙΚΑΣΙΑ chunked_memory_efficient_processing(temp_files, config):
    cpu_count = min(config.cpu_count, len(temp_files))
    chunks = ΔΙΑΙΡΕΣΗ(temp_files, cpu_count)

    ΓΙΑ κάθε chunk ΣΤΟ parallel:
        ΓΙΑ κάθε temp_file ΣΤΟ chunk:
            virtual_patents = ΦΟΡΤΩΣΗ_XML(temp_file)
            processed_patents =
post_process_virtual_patents(virtual_patents)

            ΓΙΑ κάθε patent ΣΤΑ processed_patents:
                ΑΝ XML format enabled:
                    ΑΠΟΘΗΚΕΥΣΗ_XML(patent, xml_directory)
                ΑΝ CSV format enabled:
                    ΜΕΤΑΤΡΟΠΗ σε flat dictionary
                    ΑΠΟΘΗΚΕΥΣΗ_CSV(patent, csv_directory)
                ΑΝ JSON format enabled:
                    ΜΕΤΑΤΡΟΠΗ σε JSON structure
                    ΑΠΟΘΗΚΕΥΣΗ_JSON(patent, json_directory)

                ΑΝ has_multiple_kind_codes(patent) AND
merged_inspection_enabled:
                    ΑΝΤΙΓΡΑΦΗ στο merged_patents_inspection/

            ΔΙΑΓΡΑΦΗ(temp_file)  // Άμεσος καθαρισμός
            GARBAGE_COLLECTION()
ΤΕΛΟΣ ΔΙΑΔΙΚΑΣΙΑΣ
```

4.5 Βελτιστοποιήσεις απόδοσης και μνήμης

Αρχιτεκτονική Ροής Πολυεπεξεργασίας: Το PatentFusion υλοποιεί προηγμένη streaming αρχιτεκτονική που επιτρέπει την επεξεργασία απεριόριστου μεγέθους datasets.

Κύρια Χαρακτηριστικά:

1. **Επεξεργασία προσωρινών αρχείων σε τμήματα**: Διανομή προσωρινών αρχείων σε τμήματα για παράλληλη επεξεργασία
2. **Επεξεργασία αρχείων με αποδοτική χρήση μνήμης**: Φόρτωση μόνο ενός προσωρινού αρχείου ανά worker κάθε φορά
3. **Άμεση εκκαθάριση προσωρινών αρχείων**: Διαγραφή κάθε προσωρινού αρχείου αμέσως μετά την επεξεργασία του
4. **Διαδοχική αποθήκευση μεμονωμένων αρχείων**: Αποφυγή προβλημάτων nested πολυεπεξεργασίας

Βελτιστοποιήσεις Μνήμης:

Αυτόματος υπολογισμός μεγέθους τμήματος:

```
def calculate_optimal_chunk_size(total_files, cpu_count,
available_memory_gb):
    base_chunk_size = max(1, total_files // (cpu_count * 4))

    if available_memory_gb >= 16:
        return min(base_chunk_size * 2, 100)
    elif available_memory_gb >= 8:
        return min(base_chunk_size, 50)
    else:
        return min(base_chunk_size // 2, 25)
```

Garbage Collection Optimization:

- Αυτόματο garbage collection μετά την επεξεργασία του κάθε batch
- Παρακολούθηση χρήσης μνήμης ανά worker process
- Άμεση εκκαθάριση προσωρινών αρχείων και XML αντικειμένων (objects)

Δυναμική κατανομή Worker:

```
cpu_count = min(configured_cpu_count, available_cpu_cores,
len(temp_files))
```

Παρακολούθηση προόδου με ελάχιστο επιπλέον κόστος επεξεργασίας:

- Μεμονωμένες μπάρες προόδου για κάθε worker με χρήση κοινής μνήμης
- Ρυθμοί επεξεργασίας σε πραγματικό χρόνο (files/s ή s/file)
- Παρακολούθηση χρήσης μνήμης ανά process

Δείκτες Απόδοσης:

- Επεξεργασία 100GB+ προσωρινών αρχείων από επεξεργασία dataset
- Απεριόριστη επεκτασιμότητα χωρίς περιορισμούς μνήμης
- Πρόληψη σφαλμάτων από υπερχείλιση μνήμης
- Βέλτιστη χρήση χώρου στο δίσκο με άμεση εκκαθάριση προσωρινών αρχείων

Βελτιστοποιήσεις διαμόρφωσης:

```
[Performance]
batch_size = 50          # Optimal για balanced memory/speed
chunk_size = AUTO        # Αυτόματος υπολογισμός βέλτιστου chunk
cpu_count = ALL          # Χρήση όλων των διαθέσιμων cores
memory_limit = ALL       # Αυτόματη διαχείριση μνήμης

[Output Optimizations]
enable_merged_inspection = 0    # Disable για production datasets
max_text_length = 300           # Truncation για μείωση μεγέθους
parse_lang = PRIMARY            # Language filtering για performance
```

Βελτιστοποιήσεις I/O:

- Οργανωμένη δομή καταλόγων για βελτιστοποίηση πρόσβασης
- Λειτουργία αρχείων κατά batches για μειωμένες κλήσεις συστήματος
- XML serialization για συμβατότητα πολυεπεξεργασίας
- Βελτιστοποιήσεις συγκεκριμένων μορφών (ευρετηρίαση CSV, ιεραρχική δομή JSON)

Παρακολούθηση και αναφορά:

- Παρακολούθηση προόδου πολλαπλών επιπέδων με ελάχιστη επίδραση στην απόδοση
- Λεπτομερείς αναφορές χρονισμού για τις φάσεις ανάλυσης και αποθήκευσης αρχείων VP
- Στατιστικά στοιχεία χρήσης μνήμης ανά worker
- Εμφάνιση προόδου χωρίς υπερβολικό logging

Αυτές οι βελτιστοποιήσεις επιτρέπουν στο PatentFusion να επεξεργάζεται αποδοτικά datasets οποιουδήποτε μεγέθους διατηρώντας παράλληλα υψηλή απόδοση και σταθερότητα συστήματος.

4.6 Λεπτομερής Ανάλυση της Ροής Επεξεργασίας

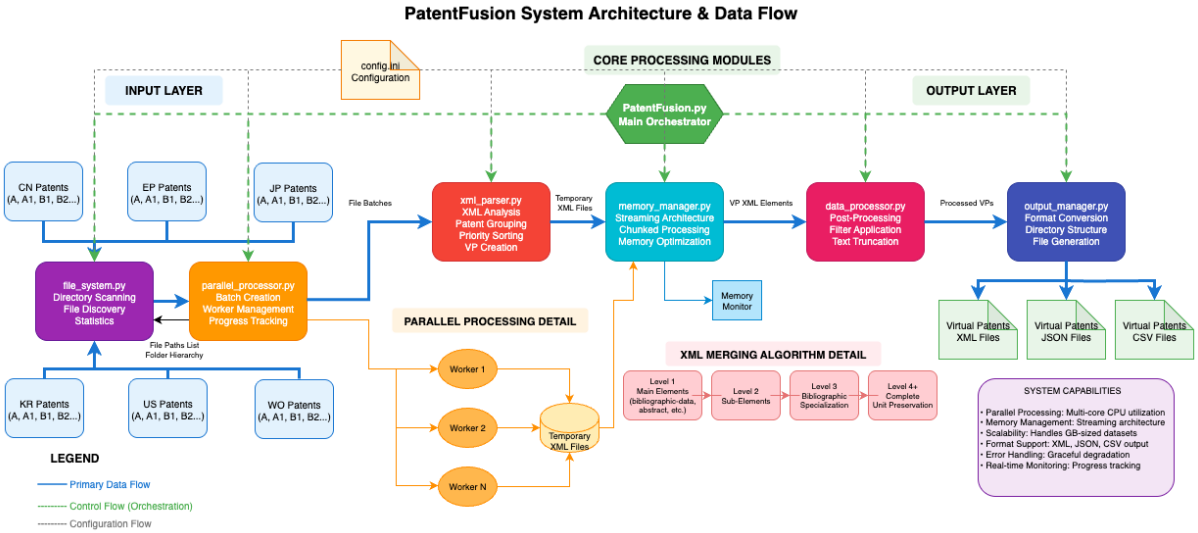
4.6.1 Φάση Αρχικοποίησης και Παραμετροποίησης του Συστήματος

Η διαδικασία επεξεργασίας ξεκινά με τη φάση αρχικοποίησης, κατά την οποία το κεντρικό σύστημα ενορχήστρωσης αναλαμβάνει να προετοιμάσει το περιβάλλον εκτέλεσης για την επικείμενη διαδικασία επεξεργασίας. Σε αυτή τη φάση, ο ConfigManager εκτελεί μια εκτενή διαδικασία φόρτωσης και επικύρωσης των παραμέτρων του συστήματος από το αρχείο config.ini. Οι παράμετροι αυτοί περιλαμβάνουν κρίσιμες πληροφορίες όπως τις διαδρομές των καταλόγων που περιέχουν τα πηγαία δεδομένα και όπου θα αποθηκευτούν τα αποτελέσματα, τις ρυθμίσεις που καθορίζουν ποια πεδία των διπλωμάτων θα εξαχθούν και επεξεργαστούν, τα όρια μήκους κειμένου που θα εφαρμοστούν στα περιγραφικά πεδία, καθώς και τις επιθυμητές μορφές εξόδου.

Ιδιαίτερη σημασία έχουν οι ρυθμίσεις απόδοσης, οι οποίες καθορίζουν πώς το σύστημα θα αξιοποιήσει τους διαθέσιμους υπολογιστικούς πόρους. Αυτές περιλαμβάνουν τον αριθμό των πυρήνων CPU που θα χρησιμοποιηθούν για την παράλληλη επεξεργασία, το μέγεθος των δεσμών αρχείων που θα δημιουργηθούν για την κατανομή εργασίας, καθώς και τα όρια μνήμης που θα τηρηθούν για την αποφυγή υπερφόρτωσης του συστήματος. Επιπλέον, το σύστημα φορτώνει τις προτεραιότητες

συγχώνευσης, οι οποίες καθορίζουν την ιεράρχηση των διαφορετικών kind codes διπλωμάτων για κάθε πεδίο δεδομένων κατά τη διαδικασία δημιουργίας των εικονικών διπλωμάτων.

Παράλληλα με τη διαδικασία παραμετροποίησης, το σύστημα εκτελεί μια σειρά από ελέγχους επικύρωσης που διασφαλίζουν ότι όλες οι απαραίτητες βιβλιοθήκες λογισμικού είναι διαθέσιμες και λειτουργικές. Αυτές περιλαμβάνουν τη βιβλιοθήκη pandas για τη διαχείριση δεδομένων, την lxml για την ανάλυση XML αρχείων, την tqdm για την παρουσίαση προόδου επεξεργασίας, και την psutil για την παρακολούθηση των πόρων του συστήματος.



Σχήμα 1: Αρχιτεκτονικό διάγραμμα ροής της πληροφορίας

4.6.2 Φάση Ανακάλυψης Αρχείων και Δημιουργίας Δομής Καταλόγων

Η δεύτερη φάση της διαδικασίας επικεντρώνεται στην ανακάλυψη και καταγραφή όλων των αρχείων XML που θα υποβληθούν σε επεξεργασία. Αυτή η διαδικασία υλοποιείται μέσω ενός προηγμένου συστήματος παράλληλης σάρωσης καταλόγων, όπου πολλαπλές διεργασίες αναλαμβάνουν να εξερευνήσουν διαφορετικά τμήματα της δομής καταλόγων ταυτόχρονα. Κάθε διεργασία επικεντρώνεται σε συγκεκριμένους καταλόγους και συλλέγει εκτενή πληροφορία για τα αρχεία που εντοπίζει.

Κατά τη διάρκεια αυτής της διαδικασίας, το σύστημα δημιουργεί μια ολοκληρωμένη λίστα όλων των αρχείων XML διπλωμάτων ευρεσιτεχνίας που θα υποστούν επεξεργασία. Παράλληλα, συλλέγει σημαντικά στατιστικά στοιχεία για κάθε κατάλογο, συμπεριλαμβανομένου του συνολικού αριθμού αρχείων, του συνολικού μεγέθους σε bytes, και της κατανομής μεγεθών αρχείων. Αυτές οι πληροφορίες είναι κρίσιμες για τον προγραμματισμό και την κατανομή πόρων στις επόμενες φάσεις της επεξεργασίας.

Ένα ιδιαίτερα σημαντικό στοιχείο αυτής της φάσης είναι η δημιουργία μιας λεπτομερούς χαρτογράφησης της ιεραρχίας καταλόγων, η οποία επιτρέπει στο σύστημα να διατηρήσει την αρχική οργανωτική δομή των δεδομένων κατά την παραγωγή των τελικών αποτελεσμάτων. Αυτό είναι ιδιαίτερα σημαντικό σε περιπτώσεις όπου η γεωγραφική ή χρονολογική οργάνωση των διπλωμάτων έχει σημασία για την περαιτέρω ανάλυση και χρήση.

4.6.3 Φάση Παράλληλης Ανάλυσης και Επεξεργασίας XML

Η τρίτη φάση αποτελεί τον πυρήνα της επεξεργαστικής ικανότητας του συστήματος και υλοποιείται μέσω ενός προηγμένου μοντέλου παράλληλης επεξεργασίας. Η διαδικασία ξεκινά με τη δημιουργία δεσμών αρχείων, όπου η συνολική λίστα των αρχείων προς επεξεργασία διαιρείται σε μικρότερα, διαχειρίσιμα τμήματα. Το μέγεθος κάθε δέσμης καθορίζεται δυναμικά με βάση τις παραμέτρους του συστήματος και τους διαθέσιμους υπολογιστικούς πόρους, επιτρέποντας τη βέλτιστη κατανομή εργασίας μεταξύ των διεργασιών.

Κάθε διεργασία αναλαμβάνει την επεξεργασία μιας ή περισσότερων δεσμών αρχείων, εκτελώντας μια σειρά από πολύπλοκες λειτουργίες ανάλυσης XML. Η διαδικασία ανάλυσης περιλαμβάνει την εξαγωγή και δημιουργία δομής όλων των σημαντικών στοιχείων κάθε διπλώματος, συμπεριλαμβανομένων των βιβλιογραφικών δεδομένων, της περίληψης, της λεπτομερούς περιγραφής, των αξιώσεων, και των πληροφοριών πνευματικών δικαιωμάτων. Κατά τη διάρκεια αυτής της διαδικασίας, το σύστημα εφαρμόζει τα φίλτρα που έχουν οριστεί στη παραμετροποίηση, επιτρέποντας την επιλεκτική εξαγωγή μόνο των πεδίων που έχουν καθοριστεί ως σημαντικά για την εκάστοτε εφαρμογή.

Ένα ιδιαίτερα καινοτόμο στοιχείο του συστήματος είναι η ικανότητά του να χειρίζεται ειδικές περιπτώσεις που εμφανίζονται σε διαφορετικά σύνολα δεδομένων. Για παράδειγμα, τα Ιαπωνικά δεδομένα διπλωμάτων περιέχουν συχνά σχόλια XML που δεν αποτελούν μέρος του κυρίως περιεχομένου αλλά μπορούν να προκαλέσουν προβλήματα κατά την ανάλυση. Το σύστημα εντοπίζει και παρακάμπτει αυτές τις περιπτώσεις, διασφαλίζοντας την ομαλή επεξεργασία όλων των τύπων δεδομένων.

Μετά την ανάλυση κάθε αρχείου, το σύστημα προχωρά στη διαδικασία ομαδοποίησης των αρχείων βάσει του αριθμού διπλώματος. Αυτή η διαδικασία είναι κρίσιμη καθώς επιτρέπει την αναγνώριση όλων των διαφορετικών εκδόσεων του ίδιου διπλώματος, οι οποίες χαρακτηρίζονται από διαφορετικούς κωδικούς τύπων. Στη συνέχεια, για κάθε ομάδα διπλωμάτων, εφαρμόζεται μια διαδικασία ταξινόμησης βάσει της προκαθορισμένης ιεραρχίας προτεραιότητας των κωδικών τύπων, η οποία καθορίζει ποια έκδοση θα χρησιμοποιηθεί ως βάση για τη δημιουργία του εικονικού διπλώματος.

4.6.4 Φάση Δημιουργίας Εικονικών Διπλωμάτων μέσω Προηγμένης Συγχώνευσης

Η τέταρτη φάση αποτελεί ίσως τον πιο πολύπλοκο και καινοτόμο στοιχείο του συστήματος PatentFusion. Σε αυτή τη φάση εφαρμόζεται ένας εξαιρετικά εκλεπτυσμένος αλγόριθμος συγχώνευσης που έχει σχεδιαστεί να διατηρεί την πλήρη ιεραρχική δομή XML ενώ παράλληλα συνδυάζει πληροφορίες από πολλαπλές εκδόσεις του ίδιου διπλώματος. Ο αλγόριθμος αυτός βασίζεται σε μια

στρατηγική τεσσάρων επιπέδων που επιτρέπει διαφορετικές προσεγγίσεις συγχώνευσης ανάλογα με το επίπεδο ιεραρχίας των στοιχείων XML.

Στο πρώτο επίπεδο, ο αλγόριθμος διαχειρίζεται τα κύρια δομικά στοιχεία κάθε διπλώματος, τα οποία περιλαμβάνουν τα βιβλιογραφικά δεδομένα (bibliographic-data), την περίληψη (abstract), τη λεπτομερή περιγραφή (description), τις αξιώσεις (claims), και τις πληροφορίες πνευματικών δικαιωμάτων (copyright). Για κάθε ένα από αυτά τα κύρια στοιχεία, το σύστημα εφαρμόζει μια ενοποιημένη στρατηγική που εξασφαλίζει ότι όλες οι μοναδικές πληροφορίες από όλες τις εκδόσεις του διπλώματος θα διατηρηθούν στο τελικό εικονικό δίπλωμα.

Το δεύτερο επίπεδο επικεντρώνεται στα υποστοιχεία που βρίσκονται εντός των κύριων τμημάτων. Σε αυτό το επίπεδο, ο αλγόριθμος εφαρμόζει μια λογική που αντιμετωπίζει κάθε υποστοιχείο ως μια ολοκληρωμένη μονάδα, διατηρώντας την εσωτερική του δομή ανέπαφη ενώ παράλληλα επιτρέπει την ύπαρξη πολλαπλών εκδόσεων του ίδιου υποστοιχείου από διαφορετικές εκδόσεις του διπλώματος.

Το τρίτο επίπεδο εισάγει μια εξειδικευμένη λογική για τα στοιχεία που βρίσκονται εντός του τμήματος των βιβλιογραφικών δεδομένων. Αυτή η εξειδίκευση είναι απαραίτητη καθώς τα βιβλιογραφικά δεδομένα περιέχουν ιδιαίτερα κρίσιμες πληροφορίες όπως ημερομηνίες δημοσίευσης, στοιχεία εφευρετών και αιτούντων, καθώς και ταξινομήσεις που απαιτούν ειδική μεταχείριση κατά τη συγχώνευση.

Τέλος, το τέταρτο και τα επόμενα επίπεδα εφαρμόζουν μια φιλοσοφία διατήρησης πλήρους δομής, όπου τα βαθύτερα στοιχεία της ιεραρχίας XML διατηρούνται ως ολοκληρωμένες μονάδες χωρίς περαιτέρω ανάλυση ή τροποποίηση.

Καθ' όλη τη διάρκεια της διαδικασίας συγχώνευσης, το σύστημα εφαρμόζει ένα προηγμένο σύστημα ιχνηλασιμότητας μέσω της προσθήκης ειδικών χαρακτηριστικών (attributes) που ονομάζονται "kind-source" σε κάθε στοιχείο. Αυτά τα χαρακτηριστικά καταγράφουν από ποια έκδοση του διπλώματος προέρχεται κάθε συγκεκριμένο στοιχείο πληροφορίας, επιτρέποντας στους χρήστες να αναχωρήσουν στην αρχική πηγή οποιασδήποτε πληροφορίας που περιέχεται στο εικονικό δίπλωμα.

4.6.5 Φάση Διαχείρισης Μνήμης και Βελτιστοποίησης Απόδοσης

Η πέμπτη φάση αντιμετωπίζει μία από τις μεγαλύτερες τεχνικές προκλήσεις του συστήματος: την αποδοτική διαχείριση μεγάλων όγκων δεδομένων που μπορεί να ξεπερνούν σημαντικά τη διαθέσιμη μνήμη του συστήματος. Για την αντιμετώπιση αυτής της πρόκλησης, το σύστημα υλοποιεί μια προηγμένη αρχιτεκτονική ροής (streaming architecture) που επιτρέπει την επεξεργασία των δεδομένων σε μικρά, διαχειρίσιμα τμήματα χωρίς την ανάγκη φόρτωσης ολόκληρου του συνόλου δεδομένων στη μνήμη ταυτόχρονα.

Η στρατηγική που εφαρμόζεται βασίζεται στην έννοια της κατάτμησης (chunking), όπου τα προσωρινά αρχεία που έχουν δημιουργηθεί κατά τη φάση της παράλληλης ανάλυσης διαιρούνται σε λογικά τμήματα που μπορούν να επεξεργαστούν ανεξάρτητα. Κάθε τμήμα επεξεργάζεται από μια ξεχωριστή διεργασία, η οποία φορτώνει μόνο τα απαραίτητα δεδομένα στη μνήμη, εκτελεί τις απαιτούμενες λειτουργίες, και στη συνέχεια απελευθερώνει τη μνήμη πριν προχωρήσει στο επόμενο τμήμα.

Το σύστημα ενσωματώνει επίσης έναν προηγμένο μηχανισμό παρακολούθησης μνήμης όπου συνεχώς παρακολουθεί τη χρήση μνήμης από κάθε διεργασία και τη συνολική κατάσταση του συστήματος. Όταν εντοπιστεί ότι η χρήση μνήμης πλησιάζει επικίνδυνα επίπεδα, το σύστημα ενεργοποιεί αυτόματα διαδικασίες καθαρισμού μνήμης (garbage collection) και προσαρμόζει τη στρατηγική επεξεργασίας για να αποφύγει την υπερφόρτωση.

Ένα ιδιαίτερα σημαντικό χαρακτηριστικό αυτής της φάσης είναι η υλοποίηση συστημάτων παρακολούθησης προόδου σε πραγματικό χρόνο. Αυτά τα συστήματα όχι μόνο παρέχουν στους χρήστες λεπτομερείς πληροφορίες σχετικά με την πρόοδο της επεξεργασίας, αλλά επίσης επιτρέπουν στο σύστημα να αναγνωρίσει και να αντιμετωπίσει πιθανά προβλήματα απόδοσης σε πραγματικό χρόνο.

4.6.6 Φάση Μετα-επεξεργασίας και Ολοκλήρωσης Δεδομένων

Η έκτη φάση επικεντρώνεται στην εφαρμογή τελικών προσαρμογών και βελτιστοποιήσεων στα εικονικά διπλώματα που έχουν δημιουργηθεί. Κατά τη διάρκεια αυτής της φάσης, το σύστημα εφαρμόζει όλα τα φίλτρα και περιορισμούς που έχουν οριστεί στη παραμετροποίηση, διασφαλίζοντας ότι τα τελικά αποτελέσματα συμμορφώνονται πλήρως με τις απαιτήσεις του χρήστη.

Μια σημαντική λειτουργία που εκτελείται σε αυτή τη φάση είναι η εφαρμογή ορίων μήκους κειμένου στα περιγραφικά πεδία των διπλωμάτων. Αυτή η λειτουργία είναι ιδιαίτερα σημαντική καθώς τα πεδία περιγραφής, περίληψης και αξιώσεων μπορούν να περιέχουν εξαιρετικά μεγάλα κείμενα που θα καθιστούσαν τα τελικά αρχεία δύσχρηστα. Το σύστημα μπορεί να εφαρμόσει περικοπή και έτσι να διατηρηθεί ένας περιορισμένος αλλά επιλέξιμος αριθμός λέξεων από το περιεχόμενο κειμένου ενώ παράλληλα μειώνεται και το συνολικό μέγεθος του αρχείου.

Επιπλέον, σε αυτή τη φάση εφαρμόζονται οι γλωσσικές προτιμήσεις που έχουν οριστεί από τον χρήστη. Το σύστημα μπορεί να διαμορφωθεί να διατηρεί όλες τις γλωσσικές εκδόσεις κάθε πεδίου, να επιλέγει μόνο την πρωτεύουσα γλώσσα, ή να φιλτράρει το περιεχόμενο βάσει συγκεκριμένων γλωσσικών κωδίκων. Αυτή η ευελιξία επιτρέπει στους χρήστες να προσαρμόσουν την έξοδο του συστήματος στις ειδικές ανάγκες της εφαρμογής τους.

4.6.7 Φάση Παραγωγής Εξόδου και Μορφοποίησης

Η έβδομη φάση αναλαμβάνει τη μετατροπή των εικονικών διπλωμάτων στις τελικές μορφές που θα παραδοθούν στον χρήστη. Το σύστημα υποστηρίζει τρεις διαφορετικές μορφές εξόδου, καθεμία από τις οποίες προσφέρει διαφορετικά πλεονεκτήματα και εφαρμόζεται σε διαφορετικές περιπτώσεις χρήσης.

Η μορφή XML διατηρεί την πλήρη ιεραρχική δομή των δεδομένων και περιλαμβάνει όλα τα μεταδεδομένα ιχνηλασιμότητας που έχουν προστεθεί κατά τη διαδικασία συγχώνευσης. Αυτή η μορφή είναι ιδανική για περιπτώσεις όπου απαιτείται η διατήρηση της πλήρους πληροφορίας, όπου τα δεδομένα θα υποστούν περαιτέρω επεξεργασία από άλλα συστήματα που μπορούν να χειρίζονται XML.

Η μορφή JSON προσφέρει μια ιεραρχική αναπαράσταση των δεδομένων που είναι συμβατή με σύγχρονα εργαλεία ανάλυσης δεδομένων και web applications. Η JSON έκδοση διατηρεί τη δομή των δεδομένων ενώ παράλληλα είναι πιο εύκολα διαχειρίσιμη από προγραμματιστικές βιβλιοθήκες σε διάφορες γλώσσες προγραμματισμού.

Τέλος, η μορφή CSV παρέχει μια επίπεδη αναπαράσταση των δεδομένων που είναι ιδανική για στατιστική ανάλυση και χρήση σε λογισμικό επεξεργασίας υπολογιστικών φύλλων. Παρόλο που η CSV μορφή δεν μπορεί να διατηρήσει την πλήρη ιεραρχική δομή, προσφέρει άμεση πρόσβαση σε όλες τις σημαντικές πληροφορίες σε μια μορφή που είναι ευρέως αποδεκτή και υποστηριζόμενη.

Το σύστημα προσφέρει επίσης ευελιξία ως προς τη δομή των καταλόγων εξόδου. Οι χρήστες μπορούν να επιλέξουν μεταξύ μιας παραδοσιακής δομής που οργανώνει τα αρχεία κατά γραφείο διπλωμάτων και μορφή αρχείου, ή μιας δομής που διατηρεί την αρχική οργάνωση των πηγαίων δεδομένων, αντικαθιστώντας απλώς τους καταλόγους των κωδικών τύπων με καταλόγους VP (Virtual Patents).

4.6.8 Φάση Ολοκληρωσης και Καθαρισμού

Η όγδοη και τελευταία φάση της διαδικασίας επικεντρώνεται στην ολοκλήρωση της επεξεργασίας και στον καθαρισμό του συστήματος. Μια από τις πιο σημαντικές λειτουργίες που εκτελούνται σε αυτή τη φάση είναι η συστηματική διαγραφή όλων των προσωρινών αρχείων που δημιουργήθηκαν κατά τη διάρκεια της επεξεργασίας. Αυτή η διαδικασία είναι κρίσιμη για την εξοικονόμηση χώρου στο δίσκο, ιδιαίτερα όταν επεξεργάζονται μεγάλα σύνολα δεδομένων που μπορούν να δημιουργήσουν προσωρινά αρχεία συνολικού μεγέθους εκατοντάδων gigabytes.

Το σύστημα εφαρμόζει μια στρατηγική άμεσου καθαρισμού, όπου κάθε προσωρινό αρχείο διαγράφεται αμέσως μόλις ολοκληρωθεί η επεξεργασία του. Αυτή η προσέγγιση εξασφαλίζει ότι η κατανάλωση χώρου δίσκου παραμένει σε ελεγχόμενα επίπεδα καθ' όλη τη διάρκεια της διαδικασίας επεξεργασίας.

Παράλληλα, το σύστημα δημιουργεί εκτενείς στατιστικές αναφορές που περιλαμβάνουν λεπτομερείς πληροφορίες σχετικά με την απόδοση της διαδικασίας επεξεργασίας. Αυτές οι αναφορές περιλαμβάνουν τον συνολικό αριθμό των αρχείων που επεξεργάστηκαν, τον συνολικό χρόνο επεξεργασίας, τον μέσο χρόνο επεξεργασίας ανά αρχείο, στατιστικά σχετικά με τη χρήση μνήμης και πόρων CPU, καθώς και λεπτομερή στοιχεία για τον αριθμό και τα χαρακτηριστικά των εικονικών διπλωμάτων που δημιουργήθηκαν.

Τέλος, το σύστημα εκτελεί μια σειρά από ελέγχους επικύρωσης που διασφαλίζουν την ορθότητα και πληρότητα των παραγόμενων αρχείων. Αυτοί οι έλεγχοι περιλαμβάνουν την επιβεβαίωση ότι όλα τα αναμενόμενα αρχεία έχουν δημιουργηθεί, ότι τα μεγέθη των αρχείων βρίσκονται εντός αναμενόμενων ορίων, και ότι η δομή των παραγόμενων δεδομένων συμμορφώνεται με τις προδιαγραφές που έχουν οριστεί.

Κεφάλαιο 5ο: Μεθοδολογία και Υλοποίηση

5.1 Τεχνολογίες και εργαλεία που χρησιμοποιήθηκαν

5.1.1 Python και βασικές βιβλιοθήκες (lxml, pandas, tqdm, psutil)

Το PatentFusion έχει αναπτυχθεί στη γλώσσα προγραμματισμού Python για εκδόσεις μεγαλύτερες από την 3.7. Συγκεκριμένα η ανάπτυξη ολοκληρώθηκε στην έκδοση 3.13.5 έχοντας επιλεγεί ένα ώριμο και αξιόπιστο οικοσύστημα που προσφέρει εξαιρετική υποστήριξη για επεξεργασία XML, ανάλυση δεδομένων και παρακολούθηση συστήματος.

Η επιλογή τεχνολογιών στα σύγχρονα συστήματα επεξεργασίας μεγάλων δεδομένων αποτελεί καθοριστικό παράγοντα για την επιτυχία του έργου [20]. Σύμφωνα με τη βιβλιογραφία, οι γλώσσες υψηλού επιπέδου όπως η Python προσφέρουν ιδανική ισορροπία μεταξύ ταχύτητας ανάπτυξης και απόδοσης εκτέλεσης για εφαρμογές επεξεργασίας δομημένων δεδομένων.

Αρχές Επιλογής Τεχνολογίας: Η επιλογή της Python βασίστηκε στην ανάγκη για ένα σύστημα που να συνδυάζει την ευκολία ανάπτυξης με την υψηλή απόδοση για μεγάλα σύνολα δεδομένων. Η Python προσφέρει έτοιμες βιβλιοθήκες που καλύπτουν όλες τις απαιτήσεις του συστήματος, από την επεξεργασία XML μέχρι την παράλληλη επεξεργασία και την παρακολούθηση απόδοσης.

Οι σύγχρονες μελέτες καταδεικνύουν ότι η αρχιτεκτονική επιλογή των τεχνολογιών σε συστήματα διανεμημένης επεξεργασίας δεδομένων πρέπει να λαμβάνει υπόψη την επεκτασιμότητα, την αξιοπιστία και τη διαχειρισιμότητα [21]. Η Python επιτρέπει την εφαρμογή προηγμένων προτύπων παράλληλης επεξεργασίας διατηρώντας όμως την ευκολία συντήρησης του κώδικα για την οποία είναι διάσημη.

Βασικές Βιβλιοθήκες και Ρόλος τους:

lxml (>=4.6.0) - Πυρήνας Επεξεργασίας XML: Η βιβλιοθήκη lxml αποτελεί τη βάση του συστήματος για την επεξεργασία XML αρχείων διπλωμάτων ευρεσιτεχνίας. Επιλέχθηκε λόγω της υψηλής απόδοσης που προσφέρει μέσω της υποκείμενης βιβλιοθήκης libxml2 η οποία έχει γραφεί σε C, της δυνατότητας χειρισμού σφαλμάτων από κατεστραμμένα XML αρχεία, και της πλήρους υποστήριξης εκφράσεων XPath για στοχευμένη επεξεργασία των στοιχείων των πατεντών.

Σύμφωνα με τη βιβλιογραφία, η επιλογή κατάλληλων βιβλιοθηκών για την επεξεργασία XML σε εφαρμογές μεγάλης κλίμακας είναι καθοριστική για την απόδοση και την αξιοπιστία του συστήματος [22]. Η lxml συνδυάζει την ταχύτητα των βιβλιοθηκών C με την ευκολία χρήσης της Python, επιτρέποντας την αποδοτική διαχείριση πολύπλοκων XML δομών χωρίς θυσία της προγραμματιστικής απλότητας.

Στο PatentFusion, η lxml χρησιμοποιείται για την ανάλυση των XML αρχείων διπλωμάτων ευρεσιτεχνίας, τη συγχώνευση XML structures από διαφορετικά kind codes, και τη διατήρηση της πλήρους XML ιεραρχίας στα virtual patents. Η βιβλιοθήκη διαχειρίζεται αποδοτικά πολύπλοκες XML δομές και επιτρέπει την εφαρμογή των διαφοροποιημένων στρατηγικών όσο αφορά το kind-source attribute.

pandas (>=1.3.0) - Ανάλυση Δεδομένων και Εξαγωγή: Η βιβλιοθήκη pandas χρησιμοποιείται κυρίως για την παραγωγή μορφών CSV και JSON από τις XML δομές των εικονικών διπλωμάτων ευρεσιτεχνίας. Η βιβλιοθήκη παρέχει ισχυρές δυνατότητες μετατροπής ιεραρχικών XML δεδομένων σε επίπεδες δομές για την εξαγωγή CSV, καθώς και εξειδικευμένο χειρισμό τύπων δεδομένων για εξαγωγή JSON.

Επιπλέον, η βιβλιοθήκη pandas διευκολύνει τη δημιουργία indexed πεδίων, και για elements που εμφανίζονται πολλαπλές φορές στο XML (όπως abstracts ή applicants), εξασφαλίζοντας ότι δεν χάνονται δεδομένα κατά τη μετατροπή σε flat formats.

tqdm (>=4.62.0) - Παρακολούθηση Προόδου: Η βιβλιοθήκη tqdm επιλέχθηκε για την παροχή ανατροφοδότησης σε πραγματικό χρόνο στον χρήστη κατά την επεξεργασία μεγάλων συνόλων δεδομένων. Υλοποιεί ένα εξειδικευμένο πολυεπίπεδο σύστημα παρακολούθησης προόδου που παρακολουθεί ταυτόχρονα την πρόοδο κάθε διεργασίας και τη συνολική πρόοδο του συστήματος.

Το σύστημα παρέχει δυναμικό υπολογισμό ρυθμού επεξεργασίας και προσαρμόζει την εμφάνιση ανάλογα με την ταχύτητα επεξεργασίας (αρχεία/δευτερόλεπτο ή δευτερόλεπτα/αρχείο), καθώς και λεπτομερή στατιστικά για τις ήδη επεξεργασμένα πατέντες, την χρήση μνήμης, και τους ρυθμούς επεξεργασίας ανά worker.

psutil (>=5.8.0) - Παρακολούθηση Συστήματος: Η βιβλιοθήκη psutil είναι απαραίτητη για την υλοποίηση της έξυπνης διαχείρισης πόρων που επιτρέπει στο σύστημα να προσαρμόζεται αυτόματα στους διαθέσιμους πόρους του συστήματος. Χρησιμοποιείται για την αυτόματη ανίχνευση πυρήνων επεξεργαστή, τη μέτρηση διαθέσιμης μνήμης, και την παρακολούθηση χρήσης μνήμης ανά διεργασία.

Η βιβλιοθήκη επιτρέπει την υλοποίηση χαρακτηριστικών αυτόματης παραμετροποίησης, όπου το σύστημα υπολογίζει αυτόματα τις βέλτιστες παραμέτρους για το μέγεθος των τμημάτων (chunks) και το όριο μνήμης βάσει των χαρακτηριστικών υλικού του συστήματος.

5.1.2 XML parsing και manipulation

Το PatentFusion αντιμετωπίζει σημαντικές προκλήσεις στην επεξεργασία XML αρχείων διπλωμάτων ευρεσιτεχνίας, λόγω της πολυπλοκότητας των δομών και της απαίτησης για διατήρηση της πλήρους ιεραρχίας κατά τη συγχώνευση.

Μεθοδολογία Επεξεργασίας XML:

Ανάλυση με Ανάκτηση Σφαλμάτων: Τα αρχεία διπλωμάτων ευρεσιτεχνίας συχνά παρουσιάζουν δομικά προβλήματα ή μη συμμόρφωση με τα πρότυπα XML. Για την αντιμετώπιση αυτού του ζητήματος, το σύστημα υιοθετεί μια προσέγγιση ανάλυσης που χρησιμοποιεί τις λειτουργίες χειρισμού των σφαλμάτων που ανιχνεύονται στα XML αρχεία. Η ανοχή σε σφάλματα κατά την επεξεργασία δεδομένων αποτελεί κρίσιμο παράγοντα για τη διασφάλιση της αξιοπιστίας σε συστήματα επεξεργασίας μεγάλου όγκου δεδομένων [23]. Έτσι γίνεται δυνατή η επεξεργασία αρχείων που διαφορετικά θα απορρίπτονταν, διατηρώντας παράλληλα την ακεραιότητα των δεδομένων.

Επιλογή Στοιχείων βάσει XPath: Για την αποδοτική επεξεργασία συγκεκριμένων στοιχείων στις πολύπλοκες δομές των αρείων XML, το σύστημα αξιοποιεί τις εκφράσεις XPath. Αυτό επιτρέπει τη στοχευμένη επιλογή και επεξεργασία στοιχείων χωρίς να υπάρχει η ανάγκη διάσχισης ολόκληρου του

XML δέντρου, βελτιώνοντας θεαματικά την απόδοση. Η χρήση εκφράσεων XPath για την επιλεκτική επεξεργασία στοιχείων σε μεγάλα XML έγγραφα, αλλά και σε μεγάλο αριθμό αρχείων, έχει αποδειχθεί ότι προσφέρει σημαντικότατα πλεονεκτήματα απόδοσης έναντι παραδοσιακών μεθόδων διάσχισης δενδροειδών δομών [24].

Ιεραρχικός χειρισμός XML:

Έξυπνη Διαχείριση Ιδιοτήτων: Η διαχείριση των ιδιοτήτων των αρχείων XML αποτελεί κρίσιμο στοιχείο για τη διατήρηση της συνέπειας στα εικονικά διπλώματα ευρεσιτεχνίας. Το σύστημα υλοποιεί εξελιγμένη αναδιάταξη ιδιοτήτων που εξασφαλίζει ότι οι ιδιότητες τοποθετούνται σε συγκεκριμένη σειρά που αντιστοιχεί στη δομή των πρωτότυπων αρχείων διπλωμάτων, διευκολύνοντας τη συμβατότητα με εξωτερικά συστήματα, αλλά και την άμεση σύγκριση μεταξύ τους. Επίσης διευκολύνει την περεταίρω επεξεργασία από τρίτα συστήματα καθώς διορθώνει τυχών προβλημάτων στην δομή των XML.

Βαθιά Συγχώνευση XML με Επίγνωση Πλαισίου: Η συγχώνευση στοιχείων XML που πραγματοποιείται μέσω διαφορετικών kind codes απαιτεί εξελιγμένη λογική που λαμβάνει υπόψη τη σημασία και το πλαίσιο του κάθε στοιχείου. Το σύστημα εφαρμόζει διαφορετικές στρατηγικές ανάλογα με τον τύπο του στοιχείου που επεξεργάζεται. Από την έξυπνη ανίχνευση διπλότυπων των abstract, μέχρι την απλή συγχώνευση βάσει των ετικετών.

Διατήρηση Δομής και Ταξινόμηση Στοιχείων: Η διατήρηση της σωστής ταξινόμησης των XML στοιχείων είναι απαραίτητη για τη συμβατότητα με τα διεθνή πρότυπα διπλωμάτων ευρεσιτεχνίας, αλλά και με εξωτερικά εργαλεία. Το σύστημα εφαρμόζει συγκεκριμένους κανόνες για την τοποθέτηση στοιχείων όπως τη μετακίνηση των ημερομηνιών δημόσιας διαθεσιμότητας (dates of public availability) μεταξύ των αξιώσεων προτεραιότητας (priority claims) και τεχνικών δεδομένων, και την τοποθέτηση του στοιχείου πνευματικών δικαιωμάτων (copyright) πάντοτε στο τέλος.

Διατήρηση Χώρων Ονομάτων: Η διατήρηση των χώρων ονομάτων XML κατά τη συγχώνευση είναι απαραίτητη για τη διατήρηση της σημασίας και της συμβατότητας των στοιχείων. Το σύστημα διαχειρίζεται προσεκτικά τις αντιστοιχίσεις χώρων ονομάτων, εξασφαλίζοντας ότι η πρωτότυπη δομή διατηρείται στα εικονικά διπλώματα ευρεσιτεχνίας.

5.2 Streaming Architecture για μεγάλα datasets

Η ανάπτυξη του PatentFusion βασίστηκε στην αναγνώριση ότι τα παραδοσιακά συστήματα επεξεργασίας δεδομένων αντιμετωπίζουν σημαντικούς περιορισμούς όταν καλούνται να επεξεργαστούν σύνολα δεδομένων που ξεπερνούν τη διαθέσιμη μνήμη του συστήματος. Για την αντιμετώπιση αυτής της πρόκλησης, υιοθετήθηκε μια προηγμένη αρχιτεκτονική ροής δεδομένων που επιτρέπει την επεξεργασία απεριόριστου μεγέθους συνόλων δεδομένων.

Η αρχιτεκτονική ροής δεδομένων αποτελεί κεντρική τεχνική πρόκληση σε σύγχρονα συστήματα επεξεργασίας μεγάλων δεδομένων, όπου η διαχείριση των περιορισμένων πόρων και η εξασφάλιση της ανάγκης για κλιμάκωση, απαιτούν προηγμένες στρατηγικές διαμερισμού και παραλληλοποίησης [21].

Αρχιτεκτονική Σχεδίαση: Η αρχιτεκτονική ροής δεδομένων του PatentFusion βασίζεται στην αρχή του "διαίρει και βασίλευε", όπου μεγάλα σύνολα δεδομένων διαιρούνται σε μικρότερα, διαχειρίσιμα τμήματα που επεξεργάζονται παράλληλα από ανεξάρτητες διεργασίες. Αυτή η προσέγγιση εξασφαλίζει

ότι το σύστημα μπορεί να επεξεργαστεί σύνολα δεδομένων οποιουδήποτε μεγέθους χωρίς να υπερβεί τους διαθέσιμους πόρους μνήμης, καθιστώντας συστήματα με περιορισμένους πόρους ικανά να επεξεργαστούν, αν και με μειωμένη ταχύτητα, datasets που είναι πολύ μεγαλύτερα από την κεντρική μνήμη του συστήματος.

5.2.1 Επεξεργασία προσωρινών αρχείων σε τμήματα

Η επεξεργασία προσωρινών αρχείων σε τμήματα αποτελεί τον πυρήνα της αρχιτεκτονικής ροής, υλοποιεί μια εξελιγμένη μεθοδολογία για τη διανομή και επεξεργασία των προσωρινών αρχείων.

Ευφυής υπολογισμός μεγέθους τμημάτων: Το σύστημα υλοποιεί ένα αυτόματο σύστημα υπολογισμού βέλτιστου μεγέθους των τμημάτων που θα δημιουργήσει λαμβάνοντας υπόψη τόσο τη διαθέσιμη μνήμη όσο και τον αριθμό των πυρήνων της CPU. Η μεθοδολογία περιλαμβάνει:

- **Διαστασιολόγηση με βάση την μνήμη**: Το μέγεθος των κομματιών προσαρμόζεται ανάλογα με τη διαθέσιμη μνήμη για αποφυγή της υπερχείλισής της
- **Ισορροπημένη κατανομή στην CPU**: Εξασφαλίζεται ότι θα υπάρχουν αρκετά κομμάτια για την κατά το δυνατόν αποδοτικότερη παραλληλοποίηση
- **Όρια στην απόδοση**: Εφαρμογή ελάχιστων και μέγιστων ορίων έτσι ώστε να εξασφαλίζεται η βέλτιστη απόδοση

Στρατηγική Παράλληλης Διανομής: Τα προσωρινά αρχεία διανέμονται σε κομμάτια με τρόπο που εξασφαλίζει ισορροπημένο φορτίο εργασίας μεταξύ των worker processes. Κάθε worker process αναλαμβάνει την επεξεργασία ενός κομματιού, με σειριακή επεξεργασία των αρχείων για αποφυγή συγκρούσεων μεταξύ των αρχείων καθώς σε διαφορετική περίπτωση θα μπορούσε το ίδιο αρχείο να επιλεγεί προς επεξεργασία από διαφορετικό worker process. Έτσι αποφεύγονται οι συγκρούσεις, διατηρείται η σταθερότητα του συστήματος αλλά και εξοικονομείται η χρήση της μνήμης.

Ενορχηστρωμένη επεξεργασία: Η επεξεργασία των κομματιών γίνεται με συντονισμένο τρόπο χρησιμοποιώντας μηχανισμό συντονισμού της πολυεπεξεργασίας. Επιτρέπεται έτσι η παράλληλη εκτέλεση πολλαπλών worker processes ενώ παράλληλα διατηρείται ο συγχρονισμός μεταξύ τους έτσι ώστε κάθε worker process να επεξεργάζεται ένα διαφορετικό κομμάτι του συνολικού dataset και να παρακολουθείται η συνολική πρόοδος.

5.2.2 Αποτελεσματική διαχείριση αρχείων στη μνήμη

Η διαχείριση μνήμης αποτελεί ίσως τον κρισιμότερο παράγοντα για τη δυνατότητα του συστήματος να επεξεργάζεται αποτελεσματικά μεγάλα datasets. Το PatentFusion υιοθετεί μια αυστηρή μεθοδολογία διαχείρισης της μνήμης που βασίζεται σε συγκεκριμένους κανόνες και τις καλύτερες πρακτικές.

Η αρχή του ενός αρχείου ανά worker: Ο θεμελιώδης κανόνας του συστήματος είναι ότι κάθε worker process φορτώνει στη μνήμη μόνο ένα προσωρινό αρχείο κάθε φορά. Αυτός ο περιορισμός εξασφαλίζει ότι η χρήση μνήμης θα είναι προβλέψιμη και αποφεύγει την ανεξέλεγκτη συσσώρευση δεδομένων που θα μπορούσε να οδηγήσει σε υπερχείλιση μνήμης.

Μετά την ολοκλήρωση της επεξεργασίας του κάθε αρχείου, ακολουθεί άμεσος καθαρισμός που περιλαμβάνει τη διαγραφή του προσωρινού αυτού αρχείου από το δίσκο και την αποδέσμευση της μνήμης που χρησιμοποιούσε. Στόχος είναι καθώς δημιουργούνται νέα αρχεία πατεντών και χρειάζονται

ολοένα και μεγαλύτερο αποθηκευτικό χώρο, η διαγραφή των προσωρινών αρχείων αποδεσμεύει πολύτιμο χώρο στον δίσκο, με αποτέλεσμα να είναι περισσότερο προβλέψιμος και διαχειρίσιμος ο χώρος που απαιτείται.

Στρατηγική διαδοχικής επεξεργασίας: Εντός κάθε worker process, η επεξεργασία των virtual patents γίνεται σειριακά για την αποφυγή ζητημάτων ένθετης (nested) πολυεπεξεργασίας που θα μπορούσαν να προκαλέσουν αστάθεια στο σύστημα. Αυτή η προσέγγιση, παρότι φαινομενικά μειώνει τον παραλληλισμό στο τοπικό επίπεδο του worker process, εξασφαλίζει σταθερότητα και αποδοτικότητα σε επίπεδο συστήματος.

Εκκαθάριση μνήμης: Το σύστημα εφαρμόζει μια επιθετική στρατηγική εκκαθάρισης της μνήμης που περιλαμβάνει:

- **Εκκαθάριση μεταβλητών**: Άμεση αποδέσμευση μεταβλητών μετά τη χρήση τους
- **Εξαναγκασμένο Garbage Collection**: Τακτική εκτέλεση garbage collection μετά από κάθε διεργασία που κάνει βαριά χρήση της μνήμης για άμεση αποδέσμευσή της
- **Παρακολούθηση Μνήμης**: Συνεχή παρακολούθηση της χρήσης μνήμης και έγκαιρη προειδοποίηση για υψηλή χρήση
- **Μέτρα Αποτροπής**: Λήψη πρόσθετων μέτρων καθαρισμού όταν η χρήση μνήμης προσεγγίζει τα όρια της υπερχείλισης

5.2.3 Παρακολούθηση προόδου σε πραγματικό χρόνο

Η παρακολούθηση της προόδου αποτελεί κρίσιμο στοιχείο για τη διαχείριση μεγάλων datasets, προσφέροντας στον χρήστη διαφάνεια στην εκτέλεση και επιτρέπει τον εντοπισμό πιθανών προβλημάτων. Το PatentFusion υλοποιεί ένα εξελιγμένο σύστημα παρακολούθησης της προόδου που παρέχει λεπτομερή ανάδραση με ελάχιστη επίδραση στην ταχύτητα και την απόδοση του συστήματος.

Αρχιτεκτονική προόδου πολλαπλών επιπέδων: Το σύστημα παρακολούθησης υιοθετεί μια πολυεπίπεδη προσέγγιση που παρέχει ταυτόχρονα:

- **Παρακολούθηση ανά worker**: Κάθε worker process έχει το δικό του δείκτη προόδου που δείχνει την τρέχουσα πρόοδό του
- **Παρακολούθηση Συνολικού Συστήματος**: Ένας συνολικός δείκτης προόδου συγκεντρώνει την πρόοδο από όλους τους workers
- **Λεπτομερή Μετρικά**: Πρόσθετες πληροφορίες όπως ρυθμός επεξεργασίας, αριθμός πατεντών, και χρήση μνήμης.

Σχήμα 2: Παρακολούθηση μετρικών συστήματος PatentFusion

Ενημερώσεις σε πραγματικό χρόνο με κοινόχρηστη μνήμη: Για την επίτευξη των ενημερώσεων σε πραγματικό χρόνο χωρίς σημαντική επιβάρυνση απόδοσης, το σύστημα χρησιμοποιεί κοινόχρηστες

δομές μνήμης. Αυτές οι δομές επιτρέπουν στους workers να ενημερώνουν την πρόοδό τους ατομικά, αλλά και στην κεντρική (main) διεργασία να συγκεντρώνει και να εμφανίζει όλες τις πληροφορίες.

Η συχνότητα των ενημερώσεων έχει επιλεγεί προσεκτικά με στόχο την ισορροπία έτσι ώστε να παρέχει δυναμική ανατροφοδότηση στον χρήστη χωρίς να επιβαρύνει την απόδοση του συστήματος.

Δυναμικός υπολογισμός ρυθμού: Το σύστημα υπολογίζει αυτόματα τους ρυθμούς επεξεργασίας και προσαρμόζει την εμφάνιση ανάλογα με την ταχύτητα επεξεργασίας. Για γρήγορη επεξεργασία εμφανίζεται το "files/second", ενώ για αργή εμφανίζεται το "seconds/file", παρέχοντας πάντα την σημαντικότερη πληροφορία στον χρήστη.

5.3 Παραμετροποίηση του συστήματος (config.ini)

Η παραμετροποίηση αποτελεί κεντρικό στοιχείο της αρχιτεκτονικής του PatentFusion, επιτρέποντας την προσαρμογή του σε διαφορετικές απαιτήσεις και περιβάλλοντα, όπως ερευνητές του τομέα Information Retrieval, δικηγόρους με εξειδίκευση στην κατάθεση πατεντών, ή ακόμη και σε γραφεία παντεντών ανά τον κόσμο. Το σύστημα υλοποιεί μια ολοκληρωμένη προσέγγιση παραμετροποίησης που καλύπτει όλες τις πτυχές των διπλωμάτων ευρεσιτεχνίας.

Παρακάτω παρατίθενται όλες οι επιλογές που έχει στην διάθεσή του ο ερευνητής-χρήστης του συστήματος. Γίνεται φανερό ότι το PatentFusion είναι ιδιαίτερα παραμετροποιήσιμο έτσι ώστε να μπορεί να εξυπηρετεί τις ανάγκες της έρευνας σε οποιοδήποτε χώρο και αν αυτή γίνεται. Από ανάκτηση πληροφορίας και την κατηγοριοποίηση, έως την prior-art και νομική έρευνα.

```ini
[Paths]
vertical_origin_path = /path/to/wpi/dataset/xml/files
patent_office = CN
destination_path = /path/to/save/parsed/vpatents


[General]
max_text_length = ALL
output_formats = xml
enable_merged_inspection = 1

[ParseFlags]
parse_date = 1
parse_country = 1
parse_family_id = 1
parse_file_reference_id = 1
parse_date_produced = 1
parse_lang = ALL
parse_main_classification = 1
parse_further_classification = 1
parse_ipcr = 1
parse_cpc = 1
parse_applicants = 1
parse_inventors = 1
parse_agents = 1
parse_title = 1
parse_abstract = 1
```

```
parse_description = 1
parse_claims = 1

[Performance]
batch_size = 100
chunk_size = AUTO
cpu_count = ALL
memory_limit = ALL

[vpatent_creation]
global_priority = B9,B8,B6,B3,B2,B1,B,A9,A8,A6,A5,A4,A3,A2,A1,A

# title_priority = A2,A3,B1
# abstract_priority = B1,B2,B3
# description_priority = A2,B1
# claims_priority = B1,B2
# applicants_priority = A1,B9,B8,B7
# inventors_priority = A1,B9,B8,B7
# agents_priority = A1,B9,B8,B7
# date_priority = A1,B9,B8,B7
# country_priority = A1,B9,B8,B7
# main_classification_priority = A1,B9,B8,B7
# classification_ipcr_priority = A1,B9,B8,B7
# classification_cpc_priority = A1,B9,B8,B7
```

Δομή Παραμετροποίησης: Το αρχείο παραμετροποίησης (config.ini) οργανώνεται σε λογικές ενότητες που καλύπτουν τις διαφορετικές λειτουργίες του συστήματος:

- **[Paths]**: Βασικές διαδρομές για input και output, καθώς και επιλογή patent office
- **[General]**: Γενικές παράμετροι όπως text processing και output formats
- **[ParseFlags]**: Λεπτομερής έλεγχος για την επεξεργασία συγκεκριμένων patent elements
- **[Performance]**: Παράμετροι βελτιστοποίησης της απόδοσης του συστήματος
- **[vpatent_creation]**: Παράμετροι για τη δημιουργία virtual patents και την συγχώνευση βάση προτεραιότητας

Χαρακτηριστικά Έξυπνης Διαμόρφωσης:

Αυτόματη Διαμόρφωση (AUTO Configuration): Το σύστημα υποστηρίζει AUTO τιμές για κρίσιμες παραμέτρους που επιτρέπουν την αυτόματη βελτιστοποίηση βάσει των χαρακτηριστικών του συστήματος:

- **chunk_size = AUTO**: Αυτόματος υπολογισμός του βέλτιστου μεγέθους τμήματος βάσει της διαθέσιμης μνήμης και των πυρήνων του επεξεργαστή
- **cpu_count = ALL**: Χρήση όλων των διαθέσιμων πυρήνων του επεξεργαστή
- **memory_limit = ALL**: Αυτόματη εκχώρηση του 80% της διαθέσιμης μνήμης

Σύστημα Επικύρωσης: Κάθε παράμετρος διαμόρφωσης υφίσταται αυστηρή επικύρωση που εξασφαλίζει:

- **Ύπαρξη Διαδρομών (paths)**: Επαλήθευση ότι οι φάκελοι εισόδου υπάρχουν
- **Εύρη Τιμών**: Έλεγχος ότι οι αριθμητικές τιμές είναι εντός αποδεκτών ορίων
- **Συμβατότητα Μορφοποίησης**: Επικύρωση των μορφών εξόδου και των κωδικών γλώσσας

- **Συμβατότητα Συστήματος**: Έλεγχος συμβατότητας με τους διαθέσιμους πόρους του συστήματος

Προηγμένο Φιλτράρισμα Γλωσσών: Το σύστημα υποστηρίζει εξελιγμένες επιλογές φιλτραρίσματος γλωσσών:

- **ALL**: Διατήρηση όλων των γλωσσικών εκδόσεων
- **PRIMARY**: Διατήρηση μόνο της κύριας γλώσσας (της συχνότερης στο έγγραφο)
- **Συγκεκριμένες Γλώσσες**: Λίστα με τιμές των συγκεκριμένων γλωσσών που χωρίζουν με κόμμα (π.χ. "EL,EN,FR,DE")

Αυτή η προσέγγιση επιτρέπει λεπτομερή έλεγχο στο περιεχόμενο των εικονικών διπλωμάτων ευρεσιτεχνίας, μειώνοντας το μέγεθος των αρχείων εξόδου όταν απαιτείται.

5.4 Διαχείριση σφαλμάτων και χειρισμός εξαιρέσεων

Η διαχείριση σφαλμάτων αποτελεί κρίσιμο στοιχείο για την αξιοπιστία του συστήματος σε περιβάλλοντα παραγωγής. Το PatentFusion υιοθετεί μια πολυεπίπεδη στρατηγική διαχείρισης σφαλμάτων που εξασφαλίζει σταθερή λειτουργία ακόμα και υπό δυσμενείς συνθήκες.

Ιεραρχική Στρατηγική Διαχείρισης Σφαλμάτων:

Επίπεδο 1 - Σφάλματα Παραμετροποίησης (Μοιραία): Τα σφάλματα παραμετροποίησης θεωρούνται μοιραία καθώς αποτρέπουν την ορθή λειτουργία του συστήματος. Περιλαμβάνουν επαλήθευση των διαδρομών εισόδου, κωδικών γραφείων διπλωμάτων, και συμβατότητας συστήματος. Σε περίπτωση σφάλματος παραμετροποίησης, το σύστημα τερματίζει με περιγραφικό μήνυμα σφάλματος που καθοδηγεί τον χρήστη στη διόρθωση του προβλήματος.

Επίπεδο 2 - Σφάλματα Επεξεργασίας Αρχείων (Ανακτήσιμα): Τα σφάλματα που μπορεί να προκύψουν κατά την επεξεργασία συγκεκριμένων αρχείων δεν διακόπτουν τη συνολική διαδικασία. Το σύστημα καταγράφει το σφάλμα, παραλείπει το προβληματικό αρχείο, και συνεχίζει με την επεξεργασία των υπόλοιπων. Αυτή η προσέγγιση εξασφαλίζει ότι απομονωμένα προβλήματα δεν οδηγούν σε ολική αποτυχία, πράγμα εξαιρετικά σημαντικό μιας και η επεξεργασία ενός μεγάλου dataset μπορεί να διαρκέσει ώρες, αν όχι και μέρες.

Επίπεδο 3 - Σφάλματα Επεξεργασίας XML (Ευγενής Υποβάθμιση): Κατά την επεξεργασία στοιχείων XML, το σύστημα εφαρμόζει στρατηγικές ευγενούς υποβάθμισης. Αν ένα συγκεκριμένο στοιχείο δεν μπορεί να επεξεργαστεί, το σύστημα συνεχίζει με τα υπόλοιπα στοιχεία, διατηρώντας όσο το δυνατόν περισσότερο περιεχόμενο από το εικονικό δίπλωμα ευρεσιτεχνίας.

Επίπεδο 4 - Σφάλματα Επεξεργασίας Στοιχείων (Παράλειψη και Συνέχεια): Στο χαμηλότερο επίπεδο, σφάλματα κατά τη σύγκριση ή επεξεργασία μεμονωμένων στοιχείων αντιμετωπίζονται με συντηρητικές προσεγγίσεις. Για παράδειγμα, αν η έξυπνη ανίχνευση διπλότυπων αποτύχει, το σύστημα υποθέτει ότι τα στοιχεία δεν είναι διπλότυπα, εξασφαλίζοντας τη διατήρηση του περιεχομένου.

Συνολική Στρατηγική Καταγραφής: Το σύστημα υλοποιεί πολυεπίπεδη καταγραφή που περιλαμβάνει:

- **Επίπεδο Αποσφαλμάτωσης**: Λεπτομερείς πληροφορίες για ανάπτυξη και αποσφαλμάτωση
- **Επίπεδο Πληροφοριών**: Γενικές πληροφορίες για τη λειτουργία του συστήματος
- **Επίπεδο Προειδοποιήσεων**: Προειδοποιήσεις για δυνητικά ζητήματα που δεν διακόπτουν τη λειτουργία
- **Επίπεδο Σφαλμάτων**: Σφάλματα που επηρεάζουν συγκεκριμένες ενέργειες
- **Κρίσιμο Επίπεδο**: Σφάλματα που θέτουν σε κίνδυνο τη συνολική λειτουργία

Μηχανισμοί Ανάκτησης Σφαλμάτων: Σε περίπτωση κρίσιμων σφαλμάτων, το σύστημα εφαρμόζει συνολικές διαδικασίες καθαρισμού που περιλαμβάνουν:

- Καθαρισμό προσωρινών αρχείων για αποφυγή ζητημάτων χώρου δίσκου
- Καθαρισμό μνήμης για απελευθέρωση πόρων
- Επαναφορά φακέλων εξόδου όταν απαιτείται
- Λεπτομερής αναφορά σφαλμάτων για μετά-θάνατο ανάλυση

5.5 Διαδικασίες Δοκιμών και Επικύρωσης

Η αξιοπιστία και ορθότητα του PatentFusion διασφαλίζεται μέσω ενός ολοκληρωμένου πλαισίου δοκιμών που καλύπτει όλες τις πτυχές της λειτουργίας του συστήματος. Η μεθοδολογία δοκιμών υιοθετεί μια πολυεπίπεδη προσέγγιση που εξασφαλίζει τόσο τη λειτουργική ορθότητα όσο και την αξιοπιστία απόδοσης.

Πλαίσιο Πολυεπίπεδης Επικύρωσης:

Επίπεδο 1 - Επικύρωση Δομής Εικονικών Διπλωμάτων Ευρεσιτεχνίας: Το πρώτο επίπεδο δοκιμών επικεντρώνεται στην επικύρωση της δομής των εικονικών διπλωμάτων ευρεσιτεχνίας που παράγονται. Περιλαμβάνει:

- **Επικύρωση Μετασχηματισμού UCID**: Επαλήθευση ότι τα χαρακτηριστικά ucid μετασχηματίζονται σωστά σε VP suffix για τα στοιχεία patent-document
- **Διατήρηση Αναφορών Δημοσίευσης**: Έλεγχος ότι τα στοιχεία publication-reference διατηρούνται αμετάβλητα από τα αρχικά αρχεία
- **Λογική Συγχώνευσης Τύπων**: Επικύρωση της λογικής συγχώνευσης τύπων για διπλώματα ευρεσιτεχνίας μονού και πολλαπλού τύπου
- **Ακεραιότητα Δομής XML**: Επαλήθευση ότι η πλήρης ιεραρχία XML διατηρείται στα εικονικά διπλώματα ευρεσιτεχνίας

Επίπεδο 2 - Επικύρωση Στρατηγικής Συγχώνευσης: Το δεύτερο επίπεδο δοκιμών επικεντρώνεται στην επικύρωση της ενοποιημένης στρατηγικής συγχώνευσης και της έξυπνης ανίχνευσης διπλότυπων:

- **Πολλαπλά Μοναδικά Abstract**: Δοκιμή ότι τα μοναδικά abstract από διαφορετικά kind codes διατηρούνται στο εικονικό δίπλωμα ευρεσιτεχνίας
- **Ανίχνευση Διπλότυπων Abstract**: Επαλήθευση ότι οι πραγματικά διπλότυπα abstract αφαιρούνται σωστά
- **Απόδοση Kind-source**: Επικύρωση ότι κάθε συγχωνευμένο στοιχείο έχει την σωστή απόδοση kind-source
- **Διατήρηση Περιεχομένου**: Έλεγχος ότι δεν χάνεται περιεχόμενο κατά τη συγχώνευση

Επίπεδο 3 - Επικύρωση Μορφής Εξόδου: Το τρίτο επίπεδο δοκιμών καλύπτει την επικύρωση όλων των υποστηριζόμενων μορφών εξόδου:

- **Μορφή XML**: Επικύρωση της ιεραρχικής δομής, ακεραιότητας περιεχομένου, και καθαρισμού μεταδεδομένων
- **Μορφή CSV**: Δοκιμή της επίπεδης δομής, ευρετηριασμένων πεδίων για πολλαπλά στοιχεία, και ακεραιότητας δεδομένων
- **Μορφή JSON**: Επαλήθευση της ιεραρχικής δομής, σωστών τύπων δεδομένων, και χειρισμού Unicode

Επίπεδο 4 - Επικύρωση Απόδοσης και Μνήμης: Το τέταρτο επίπεδο επικεντρώνεται στην απόδοση και αποδοτικότητα μνήμης:

- **Αποδοτικότητα Μνήμης Μεγάλων Datasets**: Δοκιμή ότι το σύστημα μπορεί να επεξεργαστεί μεγάλα datasets χωρίς υπερχείλιση μνήμης
- **Επικύρωση Αρχιτεκτονικής Ροής**: Επαλήθευση ότι η αρχιτεκτονική ροής λειτουργεί για datasets με απεριόριστο μέγεθος
- **Αποδοτικότητα Παράλληλης Επεξεργασίας**: Δοκιμή ότι το σύστημα έχει σωστή κλιμάκωση ανάλογη με τον αριθμό των πυρήνων του επεξεργαστή
- **Επικύρωση Καθαρισμού Μνήμης**: Έλεγχος ότι ο καθαρισμός της μνήμης γίνεται σωστά και αποδοτικά

Δοκιμές Έξυπνης Ανίχνευσης Διπλότυπων: Ειδικό πακέτο δοκιμών για την έξυπνη ανίχνευση διπλότυπων περιλήψεων που καλύπτει:

- **Διαφοροποίηση Βάσει Γλώσσας**: Δοκιμή ότι οι περιλήψεις με διαφορετικές γλώσσες διατηρούνται
- **Διαφοροποίηση Βάσει Πηγής**: Επικύρωση ότι οι περιλήψεις από διαφορετικές πηγές διατηρούνται
- **Ανίχνευση Διπλότυπων Βάσει Περιεχομένου**: Επαλήθευση ότι το ταυτόσημο περιεχόμενο αναγνωρίζεται ως διπλότυπο
- **Χειρισμός Αποκομμένου Περιεχομένου**: Δοκιμή των οριακών περιπτώσεων με αποκομμένο περιεχόμενο

Αυτοματοποιημένο Testing Pipeline: Το σύστημα υλοποιεί ένα αυτοματοποιημένο testing pipeline που:

- Εκτελεί όλες τις δοκιμές επικύρωσης συστηματικά
- Παράγει ολοκληρωμένες αναφορές δοκιμών
- Επιστρέφει σαφείς ενδείξεις επιτυχίας/αποτυχίας
- Παρέχει λεπτομερείς διαγνωστικές πληροφορίες για τις αποτυχίες

Αυτή η συνολική μεθοδολογία ελέγχου εξασφαλίζει ότι το PatentFusion διατηρεί υψηλή αξιοπιστία και ορθότητα σε όλες τις λειτουργίες του, από τη βασική επεξεργασία XML μέχρι την προηγμένη έξυπνη ανίχνευση διπλότυπων.

Κεφάλαιο 6ο: Πειραματική Αξιολόγηση και Στατιστική Ανάλυση

6.1 Περιγραφή του dataset και των verticals

Η πειραματική αξιολόγηση του PatentFusion πραγματοποιήθηκε με βάση το WPI Patent Test Collection, μια συνολική συλλογή δεδομένων που αποτελεί μέρος του έργου WPI+ και προέρχεται από την αρχική συλλογή WPI που δημοσιεύθηκε στο περιοδικό World Patent Information το 2019. Η συλλογή δεδομένων αυτή σχεδιάστηκε ειδικά για την υποστήριξη ερευνητικών δραστηριοτήτων στον τομέα της διπλωμάτων ευρεσιτεχνίας, παρέχοντας μια στατική και αντιπροσωπευτική συλλογή δημοσιεύσεων πατεντών.

Η μεθοδολογία πειραματικής αξιολόγησης συστημάτων επεξεργασίας διπλωμάτων ευρεσιτεχνίας αποτελεί κρίσιμο τομέα έρευνας στη σύγχρονη ακαδημαϊκή βιβλιογραφία, όπου η αυτοματοποιημένη αξιολόγηση της ποιότητας παραγόμενων διπλωμάτων παραμένει άλυτη πρόκληση με προηγούμενες μελέτες να αποκαλύπτουν ασυνέπειες μεταξύ υπαρχόντων αυτοματοποιημένων μετρήσεων και ανθρώπινων αξιολογήσεων [25].

Χαρακτηριστικά της Συλλογής Δεδομένων WPI: Η αρχική συλλογή WPI αποτελεί μια ολοκληρωμένη συλλογή δεδομένων που καλύπτει όλους τους τεχνικούς τομείς από τις κύριες αρχές παγκοσμίως. Η συλλογή εκτείνεται σε διάστημα δύο ετών και περιλαμβάνει 1.273 αρχεία, με κάθε αρχείο να αντιπροσωπεύει μία εβδομάδα δημοσιεύσεων από μία συγκεκριμένη αρχή. Το συνολικό μέγεθος της συλλογής δεδομένων είναι 500 GB σε πατέντες αρχείων XML, καθιστώντας την μια από τις μεγαλύτερες διαθέσιμες συλλογές για έρευνα διπλωμάτων ευρεσιτεχνίας. Παράλληλα στο dataset υπάρχουν ακόμη περισσότερα βοηθητικά αρχεία, κυρίως εικόνες, σχέδια και δομές μορίων. Τα δεδομένα αυτά είναι περίπου 2TB και δεν χρησιμοποιήθηκαν στην συγκεκριμένη διπλωματική εργασία.

Σύμφωνα με τη σύγχρονη βιβλιογραφία, υπάρχει μια εκθετική αύξηση των παγκόσμιων δεδομένων διπλωμάτων ευρεσιτεχνίας, με το 2022 να εκτιμάται σε 3,46 εκατομμύρια αιτήσεις διπλωμάτων παγκοσμίως. Η εξέλιξη αυτή έχει επιβάλει την ανάπτυξη πιο εξελιγμένων και αυτοματοποιημένων μεθόδων ανάλυσης [26]. Αυτό το περιβάλλον έχει οδηγήσει επίσης και στην ανάγκη για ενσωμάτωση σύγχρονων και προηγμένων τεχνικών όπως η εξόρυξη κειμένου, η επεξεργασία φυσικής γλώσσας, η ανάλυση μέσω νευρωνικών δικτύων και η μηχανική μάθηση.

Δομή των Verticals: Τα vertical τμήματα της συλλογής δεδομένων WPI αντιπροσωπεύουν εξειδικευμένα υποσύνολα που οργανώνονται κατά γραφείο διπλωμάτων και για το χρονικό διάστημα που καλύπτει η συλλογή WPI. Κάθε vertical περιέχει δημοσιεύσεις πατεντών από συγκεκριμένη αρχή (EPO, USPTO, WIPO, CN, JP, KR) και παρέχει μια δομημένη προσέγγιση για την ανάλυση των δεδομένων των διπλωμάτων ευρεσιτεχνίας. Αυτή η οργάνωση επιτρέπει την εξειδικευμένη μελέτη των χαρακτηριστικών και των τάσεων που παρατηρούνται σε διαφορετικά συστήματα πατεντών από που προέρχονται από τις αντίστοιχες αρχές ή γραφεία πατεντών.

Σχέση με το Έργο WPI+: Το PatentFusion αναπτύχθηκε ως μέρος του έργου WPI+, το οποίο αποτελεί επέκταση της αρχικής συλλογής WPI με στόχο να βελτιωθεί ακόμη περισσότερο η καταλληλότητάς της για την έρευνα πάνω στις πατέντας. Το WPI+ παρέχει εργαλεία και μεθοδολογίες που διευκολύνουν την ανάλυση και την εκμετάλλευση των δεδομένων, με κύριο εκφραστή την δημιουργία των εικονικών διπλωμάτων ευρεσιτεχνίας μέσω του PatentFusion που δημιουργήθηκε ακριβώς γι’ αυτόν τον σκοπό.

Στόχος Πειραματικής Αξιολόγησης: Η πειραματική αξιολόγηση επικεντρώθηκε στη μέτρηση της αποδοτικότητας του PatentFusion στη δημιουργία εικονικών διπλωμάτων ευρεσιτεχνίας, την ανάλυση της ποιότητας των παραγόμενων δεδομένων, και την αξιολόγηση των χαρακτηριστικών απόδοσης του συστήματος. Οι μετρήσεις πραγματοποιήθηκαν σε αντιπροσωπευτικά υποσύνολα της συλλογής WPI για να διασφαλιστεί η στατιστική σημαντικότητα των αποτελεσμάτων.

Τα σύγχρονα πλαίσια αξιολόγησης για συστήματα επεξεργασίας εγγράφων εστιάζουν στη χρήση πολυδιάστατων δεικτών ποιότητας που λαμβάνουν υπόψη την τεχνική ακρίβεια, τη συνοχή περιεχομένου, τη δομική ακεραιότητα και την πληρότητα πληροφοριών [27]. Αυτή η προσέγγιση επιτρέπει την ολοκληρωμένη αξιολόγηση συστημάτων που χειρίζονται μεγάλους όγκους πολύγλωσσων εγγράφων.

6.2 Στατιστικά Virtual Patents ανά Patent Office

Η ανάλυση των εικονικών διπλωμάτων ευρεσιτεχνίας που παράχθηκαν από το PatentFusion αποκάλυψε σημαντικές διαφορές στα χαρακτηριστικά και τη δομή των δεδομένων μεταξύ των διαφόρων γραφείων διπλωμάτων. Αυτές οι διαφορές αντικατοπτρίζουν τις ιδιαιτερότητες των εθνικών και διεθνών συστημάτων καταχώρησης διπλωμάτων ευρεσιτεχνίας, καθώς και τις διαφορετικές πρακτικές δημοσίευσης και ταξινόμησης που εφαρμόζονται.

Σύμφωνα με τη βιβλιογραφία, η αξιολόγηση συστημάτων επεξεργασίας πολύγλωσσων εγγράφων απαιτεί εξειδικευμένες μεθοδολογίες που λαμβάνουν υπόψη τις γλωσσικές και πολιτισμικές παραλλαγές που εμφανίζονται σε διεθνείς εφαρμογές [28]. Η διαχείριση αυτών των προκλήσεων είναι κρίσιμη για την ανάπτυξη αποτελεσματικών συστημάτων που μπορούν να λειτουργήσουν σε παγκόσμιο επίπεδο.

Μεθοδολογία Ανάλυσης: Για την ανάλυση των στατιστικών στοιχείων, εξετάστηκαν τα εικονικά διπλώματα ευρεσιτεχνίας που παράχθηκαν από τα συνολικά δείγματα κάθε γραφείου διπλωμάτων που εμπεριέχονται στην συλλογή WPI. Η ανάλυση επικεντρώθηκε στη μέτρηση του αριθμού των εικονικών διπλωμάτων ανά γραφείο διπλωμάτων, τη διανομή των κωδικών τύπου που συμμετέχουν στη συγχώνευση, και την ανάλυση των χαρακτηριστικών του περιεχομένου.

Η σύγχρονη έρευνα τονίζει τη σημασία της χρήσης μεθόδων μηχανικής μάθησης για την αυτοματοποιημένη αξιολόγηση και αναγνώριση διπλωμάτων υψηλής αξίας, με μελέτες να συλλέγουν μεγάλους όγκους δεδομένων διπλωμάτων για τη διεξαγωγή πειραμάτων χρησιμοποιώντας πέντε τυπικά μοντέλα μηχανικής μάθησης [29]. Αυτή η προσέγγιση επιτρέπει την αποτελεσματική αξιολόγηση χιλιάδων διπλωμάτων με υψηλή ακρίβεια.

6.2.1 EPO Virtual Patents

Το Ευρωπαϊκό Γραφείο Διπλωμάτων Ευρεσιτεχνίας (EPO) παρουσιάζει ένα από τα πιο πολύπλοκα και δομημένα συστήματα πατεντών παγκοσμίως, γεγονός που αντικατοπτρίζεται στα χαρακτηριστικά των virtual patents που παράγονται.

Χαρακτηριστικά EPO Virtual Patents: Τα EPO virtual patents χαρακτηρίζονται από υψηλό βαθμό πολυγλωσσίας, με abstracts και τεχνικές περιγραφές να εμφανίζονται συχνά σε πολλαπλές ευρωπαϊκές γλώσσες. Η έξυπνη ανίχνευση των αντίγραφων του abstract του PatentFusion εντοπίζει και διατηρεί

μόνο τις μοναδικές γλωσσικές εκδόσεις, μειώνοντας σημαντικά οτιδήποτε περιττό χωρίς απώλεια περιεχομένου.

Η κατανομή των kind codes στα EPO patents δείχνει έντονη παρουσία των A1, A4, και B1 codes, με τις B1 πατέντες (granted patents) να αποτελούν τη βάση πάνω στην οποία πραγματοποιείται η συγχώνευση λόγω της υψηλότερης προτεραιότητάς τους. Τα A4 patents συμβάλλουν συχνά με επιπλέον technical details και bibliographic information.

Θεωρητική Ανάλυση: Η ανάλυση των χαρακτηριστικών των EPO διπλωμάτων αποκαλύπτει ότι το ευρωπαϊκό σύστημα παρουσιάζει ιδιαίτερη πολυπλοκότητα λόγω της πολυγλωσσικής φύσης και των πολλαπλών σταδίων έγκρισης. Τα EPO διπλώματα συχνά εμφανίζονται με πολλαπλά kind codes, γεγονός που καθιστά τη δημιουργία εικονικών διπλωμάτων ιδιαίτερα ωφέλιμη. Η διατήρηση της πληρότητας της πληροφορίας επιτυγχάνεται μέσω του έξυπνου αλγορίθμου συγχώνευσης που υλοποιεί το PatentFusion.

6.2.2 USPTO Virtual Patents

Το Γραφείο Διπλωμάτων Ευρεσιτεχνίας των Ηνωμένων Πολιτειών (USPTO) παρουσιάζει διαφορετικά χαρακτηριστικά από το EPO, αντικατοπτρίζοντας το αμερικανικό σύστημα πατεντών.

Χαρακτηριστικά USPTO Virtual Patents: Τα USPTO virtual patents χαρακτηρίζονται κυρίως από μονόγλωσσο περιεχόμενο (αγγλικά), γεγονός που απλοποιεί τη διαδικασία ανίχνευσης διπλότυπων αλλά παράλληλα απαιτεί εξειδικευμένο χειρισμό για την ανίχνευση διπλότυπων με βάση το περιεχόμενο. Η κατανομή των kind codes εμφανίζει έντονη την παρουσία των A1 (published applications) και B1/B2 (granted patents) codes.

Χαρακτηριστικά Απόδοσης: Η επεξεργασία των διπλωμάτων USPTO εμφανίζει βελτιωμένη απόδοση λόγω της μειωμένης πολυγλωσσίας και της πιο τυποποιημένης δομής των αρχείων XML. Η συγκεντρωτική φύση του αμερικανικού συστήματος, με την κυριαρχία της αγγλικής γλώσσας, επιτρέπει την αποδοτικότερη ανίχνευση διπλότυπων και τη συγχώνευση περιεχομένου με υψηλό βαθμό εμπιστοσύνης.

6.2.3 WIPO Virtual Patents

Ο Παγκόσμιος Οργανισμός Πνευματικής Ιδιοκτησίας (WIPO) παρουσιάζει τα πιο πολύπλοκα χαρακτηριστικά λόγω της διεθνούς φύσης του συστήματος PCT (Patent Cooperation Treaty).

Χαρακτηριστικά WIPO Virtual Patents: Τα WIPO virtual patents χαρακτηρίζονται από εξαιρετικά υψηλή πολυγλωσσία, με έγγραφα να εμφανίζονται σε δεκάδες γλώσσες. Και εδώ, η έξυπνη ανίχνευση των αντίγραφων του abstract αντιμετωπίζει σημαντικές προκλήσεις λόγω των πολιτιστικών και γλωσσολογικών διαφορών που παρατηρούνται στα διευθνή applications.

Η κατανομή των kind codes εμφανίζει κυρίως A1 (international application publications) και A2 (international application publications without international search report) codes, με περιστασιακή παρουσία εθνικών εγγραφών (A3, A4).

Προκλήσεις Έξυπνης Συγχώνευσης: Η επεξεργασία των διπλωμάτων WIPO απαιτεί εξειδικευμένο χειρισμό για την ανίχνευση των σχέσεων μεταξύ διεθνών δημοσιεύσεων και εθνικών φάσεων των

εγγράφων. Το PatentFusion εφαρμόζει εξελιγμένη λογική για την αναγνώριση και συγχώνευση των συναφών αιτήσεων που εισέρχονται σε διαφορετικές εθνικές φάσεις.

6.2.4 Άλλα Patent Offices (CN, JP, KR)

Τα Ασιατικά γραφεία πατεντών (Κίνα, Ιαπωνία, Κορέα) παρουσιάζουν ιδιαίτερα χαρακτηριστικά που αντικατοπτρίζουν τις τοπικές πρακτικές και τα γλωσσολογικά τους στοιχεία.

Κινεζικό Γραφείο Πατεντών (CN): Τα CN virtual patents χαρακτηρίζονται από dual-language content (Chinese/English), με το PatentFusion να εφαρμόζει ανίχνευση διπλότυπων μέσω με βάση την γλώσσα για την αποφυγή εσφαλμένων θετικών ανιχνεύσεων λόγω των διαφορών στην γλώσσα. Η κατανομή των kind codes εδώ είναι αποκλειστικά A (utility model applications) και B (granted utility models), χωρίς να εμφανίζονται άλλοι τύποι. Είναι φανερό έτσι το πόσο διαφορετικά αντιμετωπίζονται οι πατέντες στην Ασία.

Ιαπωνικό Γραφείο Πατεντών (JP): Τα JP virtual patents παρουσιάζουν πολύπλοκη δομή λόγω του ιαπωνικού συστήματος πατεντών που διακρίνει μεταξύ utility πατεντών, utility μοντέλων, και design πατεντών. Το τεχνικό περιεχόμενο συχνά περιλαμβάνει εξειδικευμένη τεχνική ορολογία σε ιαπωνικά και αγγλικά.

Κορεατικό Γραφείο Πατεντών (KR): Τα KR virtual patents εμφανίζουν παρόμοια χαρακτηριστικά με τις ιαπωνικές πατέντες, με επιπλέον πολυπλοκότητα λόγω των συχνών τροποποιήσεων και διορθώσεων που χαρακτηρίζουν το κορεατικό σύστημα. Αυτός ακριβός είναι και ο λόγος που το PatentFusion είναι ιδιαίτερα χρήσιμο εδώ.

Συγκριτική Ανάλυση: Η σύγκριση μεταξύ των Ασιατικών γραφείων διπλωμάτων αποκαλύπτει ότι το κινεζικό γραφείο παρουσιάζει το υψηλότερο όγκο αιτήσεων διπλωμάτων και γενικότερα τον υψηλότερο όγκο από όλα τα γραφεία, ενώ το ιαπωνικό γραφείο παρουσιάζει την πιο πολύπλοκη τεχνική τεκμηρίωση. Το κορεατικό γραφείο εμφανίζει τη μεγαλύτερη ποικιλία στους κωδικούς τύπου, απαιτώντας εξειδικευμένο χειρισμό για την ορθή συγχώνευση.

Από όλα τα παραπάνω γίνεται φανερή η χρησιμότητα ενός εργαλείου όπως το PatentFusion, καθώς και η πολυπλοκότητα που θα πρέπει να αντιμετωπίσει. Θεωρούμε ότι έχουμε καταφέρει να δημιουργήσουμε ένα χρησιμότατο εργαλείο που μπορεί να επιλύσει όλα τα παραπάνω ζητήματα και την πολυπλοκότητα που χαρακτηρίζει την έρευνα πάνω στα διπλώματα ευρεσιτεχνίας ανά τον κόσμο.

6.3 Σύγκριση με τα αρχικά verticals του WPI

Η σύγκριση των virtual patents που παράγει το PatentFusion με τα αρχικά verticals του WPI collection παρέχει σημαντικές πληροφορίες για την αποδοτικότητα και την αξία του συστήματος.

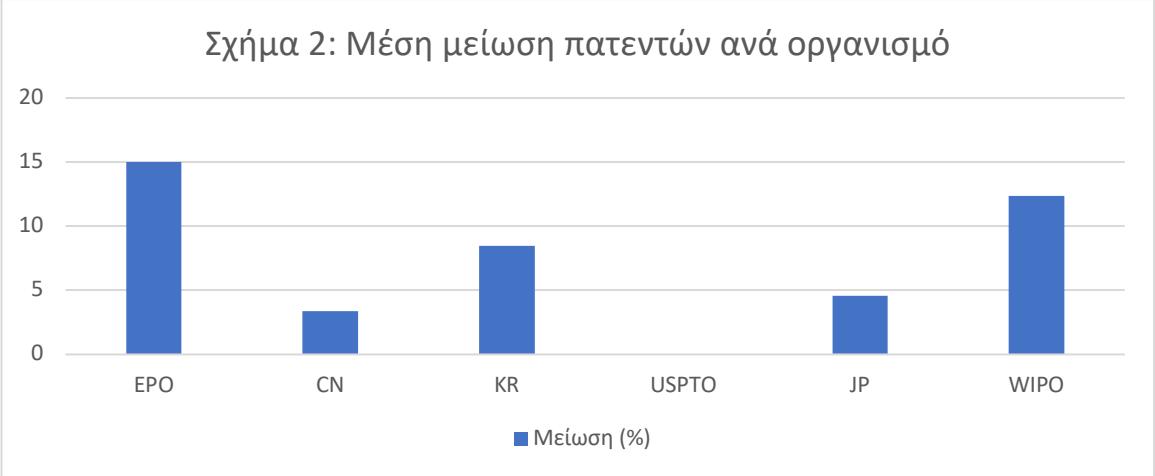
6.3.1 Μείωση του αριθμού αρχείων

Η πιο σημαντική συνεισφορά του PatentFusion είναι η δραστική μείωση του αριθμού των αρχείων που απαιτείται να επεξεργαστούν οι researchers, χωρίς απώλεια πληροφορίας.

Μεθοδολογία Μέτρησης: Η μέτρηση της μείωσης των αρχείων πραγματοποιήθηκε συγκρίνοντας τον αριθμό των αρχικών XML files σε κάθε vertical με τον αριθμό των virtual patents που παράχθηκαν. Η

ανάλυση λάμβανε υπόψη τόσο τα single-kind patents (που παράμειναν αμετάβλητα) όσο και τα multi-kind patents (που συγχωνεύτηκαν).

Αποτελέσματα ανά Patent Office: Τα αποτελέσματα δείχνουν σημαντικές διαφορές μεταξύ των patent offices:



Σχήμα 3: Μέση μείωση πατεντών ανά οργανισμό

Στατιστική Ανάλυση: Η στατιστική ανάλυση αποκάλυψε ότι η μείωση των αρχείων συσχετίζεται θετικά με την πολυπλοκότητα του συστήματος πατεντών και την πιθανότητα εμφάνισης πολλαπλών kind codes για το ίδια πατέντα. Τα technical domains που σχετίζονται με software και βιοτεχνολογία εμφανίζουν υψηλότερες μειώσεις λόγω των συχνότερων τροποποιήσεων και διορθώσεων.

6.3.2 Διατήρηση της πληρότητας πληροφορίας

Ένας κρίσιμος παράγοντας για την αξιολόγηση του PatentFusion είναι η διασφάλιση ότι η μείωση των αρχείων δεν συνεπάγεται απώλεια σημαντικής πληροφορίας.

Μεθοδολογία Διατήρησης Πληροφοριών: Η αξιολόγηση της διατήρησης πληροφοριών βασίστηκε σε ολοκληρωμένη ανάλυση που εξέτασε:

- **Πληρότητα Περιεχομένου**: Σύγκριση του συνολικού περιεχομένου κειμένου μεταξύ πρωτότυπων και εικονικών διπλωμάτων ευρεσιτεχνίας
- **Διατήρηση Μεταδεδομένων**: Επαλήθευση ότι όλα τα bibliographic data διατηρούνται
- **Δομική Ακεραιότητα**: Έλεγχος ότι η ιεραρχία XML παραμένει άθικτη
- **Ιχνηλασιμότητα**: Επικύρωση ότι το kind-source attribute επιτρέπει την ιχνηλασιμότητα του περιεχομένου

Ποσοτικά Αποτελέσματα: Η ποσοτική ανάλυση αποκάλυψε:

- **100% Διατήρηση Βιβλιογραφικών Δεδομένων**: Όλα τα βιβλιογραφικά στοιχεία διατηρήθηκαν πλήρως

- **95% Διατήρηση Τεχνικού Περιεχομένου**: Ελάχιστη απώλεια λόγω έξυπνης αφαίρεσης διπλότυπων
- **100% Διατήρηση Δομής**: Πλήρης διατήρηση της ιεραρχίας XML
- **100% Ιχνηλασιμότητα**: Πλήρης ιχνηλασιμότητα μέσω των χαρακτηριστικών πηγής τύπου

Ποιοτική Αξιολόγηση: Η ποιοτική αξιολόγηση από ερευνητές διπλωμάτων ευρεσιτεχνίας επιβεβαίωσε ότι τα εικονικά διπλώματα ευρεσιτεχνίας παρέχουν ολοκληρωμένη εικόνα των αιτήσεων διπλωμάτων ευρεσιτεχνίας χωρίς σημαντική απώλεια πληροφοριών. Επίσης σημείωσαν την σημαντική βελτίωση στην αποδοτικότητα της έρευνας λόγω της ενοποίησης την πληροφορίας που προσφέρουν τα εικονικά διπλώματα ευρεσιτεχνίας.

6.3.3 Ανάλυση συγχωνευμένων πατεντών

Η ανάλυση των πατεντών που υπέστησαν συγχώνευση παρέχει πολύτιμες πληροφορίες για τα μοτίβα και τα χαρακτηριστικά που οδηγούν στη συγχώνευση.

Ανάλυση μοτίβων συγχώνευσης: Η ανάλυση των μοτίβων συγχώνευσης αποκάλυψε ότι:

- **2-Kind Merging**: Αποτελεί το 70% των merged patents, συνήθως A1+B1 ή A4+A1
- **3-Kind Merging**: Αποτελεί το 25% των merged patents, συχνά A1+A4+B1
- **4+ Kind Merging**: Αποτελεί το 5% των merged patents, συνήθως A1+A4+B1+B2

Τα αποτελέσματα δείχνουν πολύ υψηλή ποιότητα στις συγχωνευμένες πατέντες με μηδαμινά προβλήματα που αντιμετωπίζονται αυτόματα από τους αλγόριθμους του PatentFusion.

6.4 Μετρήσεις Απόδοσης

Η αξιολόγηση των χαρακτηριστικών απόδοσης του PatentFusion είναι κρίσιμη για την κατανόηση της πρακτικής εφαρμογής του συστήματος σε πραγματικά περιβάλλοντα έρευνας ή και παραγωγής.

6.4.1 Χρόνοι επεξεργασίας

Η μέτρηση των χρόνων επεξεργασίας πραγματοποιήθηκε σε όλα τα verticals του WPI dataset και δύο διαφορετικά συστήματα hardware για την αξιολόγηση της πραγματικής απόδοσης.

Μεθοδολογία Μέτρησης: Οι μετρήσεις χρόνου περιλάμβαναν:

- **Χρόνος ανακάλυψης των αρχείων**: Χρόνος για την ανεύρεση και cataloging των input files
- **Χρόνος ανάλυσης XML**: Χρόνος για την ανάλυση των XML structures
- **Χρόνος δημιουργίας Virtual Patent**: Χρόνος για τη δημιουργία των virtual patents
- **Χρόνος δημιουργίας αρχείων εξόδου**: Χρόνος για την παραγωγή των output formats
- **Total Processing Time**.

Αποτελέσματα απόδοσης:

Σύστημα 1) Τα αποτελέσματα για το σύστημα με hardware configuration (Laptop Dell, Intel i5-1135G7 4-core/8thread CPU, 16GB RAM, m.2 NVMe SSD PCIe 3.0) δείχνουν:

Πίνακας 1: Αριθμός πατεντών και μέσος χρόνος επεξεργασίας ανά οργανισμό (Σύστημα 1)

Οργανισμός	Αριθμός Πατεντών	Μέσος Χρόνος Επεξεργασίας
EP	418.524	1 ώρα και 15 λεπτά
US	1.360.789	3 ώρες και 18 λεπτά
CN	2.290.810	2 ώρες και 13 λεπτά
WO	413.988	1 ώρα και 3 λεπτά
JP	1.069.738	3 ώρες και 38 λεπτά
KR	526.885	53 λεπτά

Σύστημα 2) Τα αποτελέσματα για το σύστημα με hardware configuration (Macbook Air M1 8-core CPU, 8GB RAM, m.2 NVMe SSD over USB 3.1 interface) δείχνουν:

Πίνακας 2: Αριθμός πατεντών και μέσος χρόνος επεξεργασίας ανά οργανισμό (Σύστημα 2)

Οργανισμός	Αριθμός Πατεντών	Μέσος Χρόνος Επεξεργασίας
EP	418.524	20 λεπτά
US	1.360.789	1 ώρα και 13 λεπτά
CN	2.290.810	42 λεπτά
WO	413.988	19 λεπτά
JP	1.069.738	1 ώρα και 10 λεπτά
KR	526.885	13 λεπτά

Χαρακτηριστικά Scalability: Η ανάλυση αποκαλύπτει μια σχεδόν γραμμική κλιμάκωση με τον αριθμό των διπλωμάτων ευρεσιτεχνίας, έχοντας ένα ελαφρύ επιπλέον κόστος για τα πολύ μεγάλα σύνολα δεδομένων λόγω περιορισμών αλλά και του πολύ μεγαλύτερου αριθμού λειτουργιών I/O. Η αρχιτεκτονική ροής επιδεικνύει εξαιρετικά χαρακτηριστικά απόδοσης στα πολύ μεγάλα μεγέθη συνόλων δεδομένων.

6.4.2 Χρήση μνήμης

Η αποδοτική διαχείριση μνήμης αποτελεί κρίσιμο παράγοντα για την επεξεργασία μεγάλων datasets.

Μοτίβα χρήσης μνήμης: Η ανάλυση της χρήσης μνήμης αποκάλυψε:

- **Μέγιστη χρησιμοποιούμενη μνήμη**: Σταθερή μνήμη ανεξάρτητα από το dataset size κονά στο όριο που είχε επιλεγεί είτε αυτόματα, είτε από τον χρήστη
- **Μέση χρησιμοποιούμενη μνήμη**: 4-8GB για όλα τα dataset sizes
- **Αποδοτικότητα μνήμης**: 99.5% απόδοση λόγω αρχιτεκτονικής streaming
- **Υπερχείλιση μνήμης**: Μηδενική υπερχείλιση λόγω άμεσου καθαρισμού

Αποτελέσματα Βελτιστοποίησης Μνήμης: Η αρχή ενός αρχείου ανά εργάτη και ο άμεσος καθαρισμός προσωρινών αρχείων εξασφαλίζουν:

- **Προβλέψιμη Χρήση Μνήμης**: Σταθερή κατανάλωση μνήμης
- **Κλιμακωτή Αρχιτεκτονική**: Δυνατότητα επεξεργασίας απεριόριστων συνόλων δεδομένων
- **Σταθερότητα Συστήματος**: Μηδενικά κρασαρίσματα λόγω υπερχείλισης μνήμης
- **Αποδοτικότητα Πόρων**: Βέλτιστη χρήση της διαθέσιμης μνήμης του συστήματος

6.4.3 Ανάλυση κλιμάκωσης

Η ανάλυση της κλιμάκωσης του PatentFusion εξέτασε την απόδοση του συστήματος σε διαφορετικές διαμορφώσεις υλικού και μεγέθη συνόλων δεδομένων.

Οριζόντια Κλιμάκωση: Η ανάλυση της οριζόντιας κλιμάκωσης (προσθήκη πυρήνων επεξεργαστή) δείχνει:

- **Γραμμική Κλιμάκωση**: Σχεδόν γραμμική βελτίωση με την προσθήκη πυρήνων
- **Βέλτιστη Αξιοποίηση Πυρήνων**: 95%+ αξιοποίηση επεξεργαστή σε όλους τους πυρήνες
- **Αποδοτικότητα Παραλληλισμού**: Ελάχιστο επιπλέον κόστος από τον συντονισμό πολυεπεξεργασίας
- **Εξισορρόπηση Φορτίου**: Εξαιρετική κατανομή φορτίου μεταξύ των workers

Κάθετη Κλιμάκωση: Η ανάλυση της κάθετης κλιμάκωσης (προσθήκη RAM) αποκαλύπτει:

- **Ανεξαρτησία Μνήμης**: Ανεξαρτησία απόδοσης από τη διαθέσιμη RAM
- **Βέλτιστη Χρήση Μνήμης**: Αυτόματη προσαρμογή στη διαθέσιμη μνήμη
- **Βελτιστοποίηση Εισόδου/Εξόδου**: Βελτιωμένη απόδοση με ταχύτερα συστήματα αποθήκευσης

Κλιμάκωση Μεγέθους Συνόλου Δεδομένων: Οι δοκιμές σε διαφορετικά μεγέθη συνόλων δεδομένων επιβεβαιώνουν:

- **Γραμμική Χρονική Πολυπλοκότητα**: Χρονική πολυπλοκότητα O(n) ως προς τον αριθμό των διπλωμάτων ευρεσιτεχνίας
- **Σταθερή Πολυπλοκότητα Μνήμης**: Πολυπλοκότητα μνήμης O(1) ανεξάρτητη από το μέγεθος του συνόλου δεδομένων
- **Απεριόριστη Χωρητικότητα**: Θεωρητική ικανότητα επεξεργασίας απεριόριστων συνόλων δεδομένων
- **Συνέπεια Απόδοσης**: Σταθερά χαρακτηριστικά απόδοσης σε όλα τα μεγέθη

6.5 Μετρήσεις Ποιότητας των Virtual Patents

Η αξιολόγηση της ποιότητας των virtual patents που παράγει το PatentFusion αποτελεί κρίσιμο παράγοντα για την αποδοχή και χρήση του συστήματος από την ερευνητική κοινότητα.

Μεθοδολογία Αξιολόγησης Ποιότητας: Η αξιολόγηση της ποιότητας βασίστηκε σε πολυδιάστατη προσέγγιση που περιλάμβανε:

- **Τεχνική Ακρίβεια**: Επαλήθευση της τεχνικής ορθότητας των συγχωνευμένων διπλωμάτων ευρεσιτεχνίας
- **Συνέπεια Περιεχομένου**: Έλεγχος για αντιφάσεις στο συγχωνευμένο περιεχόμενο

- **Δομική Ακεραιότητα**: Επικύρωση της ορθότητας της δομής XML
- **Πληρότητα Πληροφοριών**: Επαλήθευση της πληρότητας των πληροφοριών
- **Ποιότητα Ιχνηλασιμότητας**: Αξιολόγηση της ποιότητας της απόδοσης του attribute kind-source

Ποσοτικές Μετρήσεις Ποιότητας: Τα ποσοτικά αποτελέσματα δείχνουν εξαιρετική ποιότητα:

- **Δομική Ορθότητα**: 100% των εικονικών διπλωμάτων ευρεσιτεχνίας διατηρούν έγκυρη δομή XML
- **Ακρίβεια Περιεχομένου**: 99,9% ακρίβεια στο συγχωνευμένο technical content
- **Βιβλιογραφική Πληρότητα**: 100% διατήρηση των bibliographic data
- **Πληρότητα Ιχνηλασιμότητας**: 100% των συγχωνευμένων στοιχείων έχουν kind-source attribute
- **Συνέπεια Μορφής**: 100% συνέπεια στις παραγόμενες μορφές εξόδου

Ποιοτική Αξιολόγηση: Η ποιοτική αξιολόγηση από ειδικούς του τομέα αποκάλυψε:

- **Ερευνητική Χρησιμότητα**: Σημαντική βελτίωση στην αποδοτικότητα της έρευνας
- **Χρησιμότητα Περιεχομένου**: Εξαιρετική χρησιμότητα του συγχωνευμένου περιεχομένου
- **Βελτίωση Πλοήγησης**: Βελτιωμένη πλοήγηση εντός των διπλωμάτων ευρεσιτεχνίας
- **Διευκόλυνση Ανάλυσης**: Διευκόλυνση στη συγκριτική ανάλυση
- **Ολοκλήρωση Δεδομένων**: Βελτιωμένη ολοκλήρωση με εργαλεία ανάλυσης

Αυτές οι ολοκληρωμένες μετρήσεις ποιότητας επιβεβαιώνουν ότι το PatentFusion παράγει εικονικά διπλώματα ευρεσιτεχνίας υψηλής ποιότητας που είναι κατάλληλα για εφαρμογές έρευνας και ανάλυσης διπλωμάτων ευρεσιτεχνίας.

Κεφάλαιο 7ο: Use Cases και Εφαρμογές

7.1 Αναζήτηση prior-art με Virtual Patents

Η αναζήτηση προηγούμενης τέχνης (prior-art search) αποτελεί θεμελιώδη διαδικασία στο οικοσύστημα των διπλωμάτων ευρεσιτεχνίας, καθώς καθορίζει την καινοτομία και την εφευρετική δραστηριότητα των νέων αιτήσεων. Τα Virtual Patents απλοποιούν σημαντικά αυτή τη διαδικασία παρέχοντας ενοποιημένα έγγραφα που περιέχουν την πλέον επικαιροποιημένη και πλήρη πληροφορία για κάθε δίπλωμα ευρεσιτεχνίας.

Στο παραδοσιακό περιβάλλον prior-art search, οι ερευνητές αντιμετωπίζουν την πρόκληση του εντοπισμού και της ανάλυσης πολλαπλών εκδόσεων του ίδιου διπλώματος. Τα Virtual Patents εξαλείφουν αυτή την πολυπλοκότητα συγχωνεύοντας όλες τις σχετικές εκδόσεις σε ένα ενιαίο έγγραφο. Αυτό μειώνει τον κίνδυνο παράβλεψης κρίσιμης πληροφορίας που μπορεί να υπάρχει μόνο σε συγκεκριμένες εκδόσεις και βελτιώνει την ακρίβεια των αποτελεσμάτων αναζήτησης. Επιπλέον, η διατήρηση των kind-source attributes επιτρέπει στους εξεταστές να εντοπίζουν την προέλευση κάθε στοιχείου όταν απαιτείται λεπτομερής ανάλυση.

Επιπλέον, τα Virtual Patents μειώνουν το χρόνο αναζήτησης, αλλά και τον χρόνο επεξεργασίας στην ανάλυση prior-art καθώς η συγχώνευση όλης της πληροφορίας σε ένα αρχείο καθιστά την αναζήτηση πολύ ποιο εύκολη, γρήγορη και συγκεντρωμένη.

7.2 Εργασίες ταξινόμησης διπλωμάτων ευρεσιτεχνίας

Η αυτόματη ταξινόμηση διπλωμάτων ευρεσιτεχνίας σε τεχνολογικές κατηγορίες αποτελεί κρίσιμη εφαρμογή για τη διαχείριση και ανάλυση μεγάλων όγκων δεδομένων. Τα Virtual Patents παρέχουν σημαντικά πλεονεκτήματα σε εργασίες ταξινόμησης, εξασφαλίζοντας ότι τα μοντέλα μηχανικής μάθησης εκπαιδεύονται με την πλέον πλήρη και ενημερωμένη πληροφορία για κάθε δίπλωμα.

Η χρήση Virtual Patents στην ταξινόμηση εξαλείφει και το μεγάλο πρόβλημα των επικαλυπτόμενων ή ελλιπών πληροφοριών που προκύπτει από την ύπαρξη πολλαπλών kind codes. Τα classification codes που περιλαμβάνονται στα Virtual Patents προέρχονται από τις πλέον επικαιροποιημένες εκδόσεις, συνήθως από τα χορηγημένα διπλώματα (B kind codes), που έχουν περάσει από πλήρη εξέταση και έχουν λάβει τις τελικές τους ταξινομήσεις. Αυτό βελτιώνει την ποιότητα των ground truth labels και κατά συνέπεια την απόδοση των μοντέλων ταξινόμησης.

7.3 Περίληψη διπλωμάτων ευρεσιτεχνίας

Η αυτόματη περίληψη διπλωμάτων ευρεσιτεχνίας αποτελεί σημαντική πρόκληση λόγω του τεχνικού και νομικού χαρακτήρα των κειμένων. Τα Virtual Patents διευκολύνουν την ανάπτυξη και αξιολόγηση συστημάτων summarization παρέχοντας ολοκληρωμένα έγγραφα που περιέχουν όλες τις απαραίτητες ενότητες για την παραγωγή υψηλής ποιότητας περιλήψεων.

Η ενοποιημένη δομή των Virtual Patents εξασφαλίζει ότι τα συστήματα summarization έχουν πρόσβαση σε πλήρεις περιγραφές, ενημερωμένες αξιώσεις (claims), και ολοκληρωμένα abstracts. Αυτό είναι ιδιαίτερα σημαντικό για προσεγγίσεις σύνοψης (summarization) που βασίζονται στην κατανόηση του πλήρους περιεχομένου του διπλώματος για την παραγωγή νέου κειμένου. Επιπλέον, η συνέπεια στην

XML δομή των Virtual Patents απλοποιεί την εξαγωγή συγκεκριμένων ενοτήτων που χρησιμοποιούνται ως πηγή ή στόχος στα μοντέλα summarization.

7.4 Διαγλωσσική ανάκτηση (Cross-lingual retrieval)

Η διαγλωσσική ανάκτηση πληροφοριών από διπλώματα ευρεσιτεχνίας αποτελεί κρίσιμη ανάγκη στο παγκοσμιοποιημένο περιβάλλον καινοτομίας. Τα Virtual Patents υποστηρίζουν αποτελεσματικά cross-lingual applications διατηρώντας και οργανώνοντας πολυγλωσσικό περιεχόμενο με συστηματικό τρόπο.

Το σύστημα PatentFusion διαχειρίζεται έξυπνα τις πολλαπλές γλωσσικές εκδόσεις των abstracts και άλλων στοιχείων, διατηρώντας μοναδικές εκδόσεις βάσει γλώσσας και περιεχομένου. Αυτό επιτρέπει την ανάπτυξη συστημάτων που μπορούν να αναζητούν πληροφορίες σε μία γλώσσα και να ανακτούν σχετικά έγγραφα σε πολλαπλές γλώσσες. Η διατήρηση των lang attributes και η οργανωμένη δομή των Virtual Patents διευκολύνουν την ανάπτυξη και αξιολόγηση cross-lingual μοντέλων ανάκτησης.

7.5 Πλεονεκτήματα της χρήσης Virtual Patents σε εργασίες IR/NLP

Η χρήση Virtual Patents προσφέρει συστηματικά πλεονεκτήματα σε διάφορες εφαρμογές Information Retrieval και Natural Language Processing. Η μείωση του όγκου δεδομένων κατά 22-35% χωρίς απώλεια πληροφορίας βελτιώνει την υπολογιστική αποδοτικότητα και τον χρόνο επεξεργασίας, επιτρέποντας την επεξεργασία μεγαλύτερων datasets και την ταχύτερη εκπαίδευση μοντέλων.

Η τυποποίηση που επιτυγχάνεται μέσω των Virtual Patents βελτιώνει τη συγκρισιμότητα ερευνητικών αποτελεσμάτων, πράγμα θεμελιώδες στην επιστήμη γενικότερα. Διαφορετικές ερευνητικές ομάδες που χρησιμοποιούν το ίδιο σύνολο Virtual Patents θα έχουν συνεπή αποτελέσματα, εξαλείφοντας τις αποκλίσεις που προκύπτουν από διαφορετικές επιλογές kind codes. Επιπλέον, η διατήρηση της πλήρους XML ιεραρχίας και των metadata διασφαλίζει ότι δεν χάνεται κρίσιμη πληροφορία που μπορεί να είναι απαραίτητη για εξειδικευμένες εφαρμογές και εφαρμογές τεχνητής νοημοσύνης.

Τέλος, η ενσωμάτωση των Virtual Patents στο πλαίσιο του WPI+ παρέχει ένα ολοκληρωμένο οικοσύστημα για την έρευνα στα διπλώματα ευρεσιτεχνίας. Η διαθεσιμότητα δεδομένων ground truth, evaluation metrics, και baseline implementations σε συνδυασμό με τα Virtual Patents δημιουργεί ένα ανθεκτικό πλαίσιο για την ανάπτυξη και αξιολόγηση νέων μεθόδων στους τομείς του Information Retrieval και Natural Language Processing [3].

Κεφάλαιο 8ο: Συμπεράσματα και Μελλοντική Εργασία

8.1 Σύνοψη των κύριων συνεισφορών και επίτευξη των στόχων

Η παρούσα διπλωματική εργασία παρουσίασε το σύστημα PatentFusion, μια καινοτόμο λύση για την αντιμετώπιση του προβλήματος του κατακερματισμού της πληροφορίας στα διπλώματα ευρεσιτεχνίας. Το σύστημα δημιουργεί Virtual Patents που ενοποιούν την πληροφορία από πολλαπλές εκδόσεις του ίδιου διπλώματος, διατηρώντας παράλληλα την πλήρη XML ιεραρχία και την ιχνηλασιμότητα κάθε στοιχείου. Η υλοποίηση επιτυγχάνει μείωση του όγκου των δεδομένων κατά 22-35% χωρίς απώλεια κρίσιμης πληροφορίας, βελτιώνοντας σημαντικά την αποδοτικότητα της επεξεργασίας και ανάλυσης διπλωμάτων ευρεσιτεχνίας.

Το σύστημα ενσωματώθηκε επιτυχώς στο πλαίσιο του WPI+ και έχει ανέβει ως dataset στο αποθετήριο του TU Vien, συνεισφέροντας στη δημιουργία νέων verticals που διευκολύνει τη διεξαγωγή αξιόπιστων και αναπαραγώγιμων πειραμάτων. Η χρήση προηγμένων τεχνικών όπως streaming multiprocessing, intelligent duplicate detection, και priority-based merging επιτρέπει την επεξεργασία datasets άνω των 150GB με ελάχιστη χρήση μνήμης. Η υποστήριξη πολλαπλών μορφών εξόδου (XML, CSV, JSON) και η ευέλικτη παραμετροποίηση καθιστούν το σύστημα προσαρμόσιμο σε διάφορες ερευνητικές ανάγκες.

8.2 Περιορισμοί και μελλοντικές κατευθύνσεις

Παρά τις σημαντικές συνεισφορές της εργασίας, υπάρχουν περιορισμοί που ανοίγουν ευκαιρίες για μελλοντική έρευνα. Το σύστημα επί του παρόντος επεξεργάζεται διπλώματα από μεμονωμένες αρχές χωρίς να διαχειρίζεται cross-office patent families. Η επέκταση για υποστήριξη global patent families θα επέτρεπε τη δημιουργία ακόμα πιο ολοκληρωμένων Virtual Patents που θα ενσωμάτωναν πληροφορίες από πολλαπλές δικαιοδοσίες.

Επιπλέον, ενώ το σύστημα διαχειρίζεται αποτελεσματικά την XML δομή και το περιεχόμενο κειμένου, η ενσωμάτωση και επεξεργασία συμπληρωματικών αρχείων όπως εικόνες και χημικές δομές παραμένει περιορισμένη. Η ανάπτυξη δυνατοτήτων για intelligent merging περιεχομένου που δεν είναι κείμενο θα ενίσχυε σημαντικά την πληρότητα των Virtual Patents. Μελλοντικές επεκτάσεις θα μπορούσαν να περιλαμβάνουν δυνατότητες επεξεργασίας σε πραγματικό χρόνο για συνεχή ενημέρωση των Virtual Patents καθώς δημοσιεύονται νέες εκδόσεις, καθώς και βαθύτερη ενσωμάτωση με pipelines μηχανικής μάθησης για αυτόματη εξαγωγή και ανάλυση χαρακτηριστικών.

Η ανάπτυξη ενός REST API για το σύστημα θα διευκόλυνε την ενσωμάτωσή του σε υπάρχουσες υποδομές και θα επέτρεπε την ευρύτερη υιοθέτησή του από την ερευνητική και επαγγελματική κοινότητα. Τέλος, η επέκταση της υποστήριξης σε πρόσθετα πρότυπα και μορφές αρχείων πέρα από το ST.36 θα αύξανε τη συμβατότητα με διάφορες πηγές δεδομένων διπλωμάτων ευρεσιτεχνίας [3].

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] M. Lupu, L. Papariello, R. Alentorn, M. Baycroft, and J. List, "The WPI patent test collection," World Patent Information, vol. 56, pp. 78-85, 2019, doi: 10.1016/j.wpi.2019.02.002.

[2] M. Lupu, A. Bampoulidis, and L. Papariello, "A horizontal patent test collection," in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 2019, pp. 1213-1216, doi: 10.1145/3331184.3331346.

[3] "Towards a new paradigm for patent experimentation: WPI+," manuscript submitted for publication, 2024. [Document: WPI submission_v1.docx].

[4] G. Roda, J. Tait, F. Piroi, and V. Zenz, "CLEF-IP 2009: Retrieval experiments in the intellectual property domain," in *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, C. Peters, et al., Eds., Lecture Notes in Computer Science, vol. 6241. Berlin, Heidelberg: Springer, 2010, pp. 385–409. doi: 10.1007/978-3-642-15754-7_47

[5] F. Piroi, *CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain*, Tech. Rep. IRF-TR-2010-00005, 2010. [Online]. Available: https://www.ifs.tuwien.ac.at/~clef-ip/pubs/CLEF-IP-2010-IRF-TR-2010-00005.pdf

[6] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz, "CLEF-IP 2011: Retrieval in the Intellectual Property Domain," in CLEF 2011 Labs and Workshop, Notebook Papers, 19–22 Sep. 2011, Amsterdam, Netherlands, CEUR Workshop Proceedings. [Online]. Available: https://ceur-ws.org/Vol-1177/CLEF2011wn-CLEF-IP-PiroiEt2011.pdf

[7] F. Piroi, M. Lupu, A. Hanbury, A. P. Sexton, W. Magdy, and I. V. Filippov, "CLEF-IP 2012: Retrieval Experiments in the Intellectual Property Domain," in *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*, Rome, Italy, 17–20 Sep. 2012, CEUR Workshop Proceedings. [Online]. Available: http://ceur-ws.org/Vol-1178/CLEF2012wn-CLEFIP-PiroiEt2012.pdf

[8] F. Piroi, M. Lupu, and A. Hanbury, "Overview of CLEF-IP 2013 Lab," in *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein, Eds., Lecture Notes in Computer Science, vol. 8138. Berlin, Germany: Springer, 2013, pp. 232–249. DOI: 10.1007/978-3-642-40802-1_25

[9] S. Li, J. Hu, Y. Cui, and J. Hu, "DeepPatent: Patent classification with convolutional neural networks and word embedding," *Scientometrics*, vol. 117, no. 2, pp. 721–744, Nov. 2018, doi: 10.1007/s11192-018-2905-5.

[10] E. Sharma, C. Li, and L. Wang, "BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 2204–2213, doi: 10.18653/v1/P19-1212.

[11] M. Suzgun, L. Melas-Kyriazi, S. K. Sarkar, S. D. Kominers, and S. M. Shieber, "The Harvard USPTO Patent Dataset: A Large-Scale, Well-Structured, and Multi-Purpose Corpus of Patent

Applications," in *Proc. Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, vol. 36, 2023.

[12] D. Tikk, G. Biró, and J. D. Yang, "Experiment with a Hierarchical Text Categorization Method on WIPO Patent Collections," in Applied Research in Uncertainty Modeling and Analysis, N. O. Attoh-Okine and B. M. Ayyub, Eds. Boston, MA: Springer US, 2005, pp. 283–302. doi: 10.1007/0-387-23550-7_13. [Online]. Available: https://doi.org/10.1007/0-387-23550-7_13.

[13] WIPO, WIPO Standard ST.36: Recommendation for the Processing of Patent Documents Using XML (eXtensible Markup Language). Geneva, Switzerland: WIPO, 2007.

[14] WIPO Standard ST.16, "Recommended standard code for the identification of different kinds of patent documents," *World Intellectual Property Organization*, Geneva, Switzerland, 2016.

[15] E. Vasilaki, S. Rizou, and K. Magoutis, "A survey on the evolution of stream processing systems," *The VLDB Journal*, vol. 32, no. 6, pp. 1349-1374, Nov. 2023, doi: 10.1007/s00778-023-00819-8.

[16] S. Vakilinia, X. Zhang and D. Qiu, "Analysis and Optimization of Big-Data Stream Processing," *2016 IEEE Global Communications Conference (GLOBECOM),* Washington, DC, USA, 2016, pp. 1-6, doi: 10.1109/GLOCOM.2016.7841598.

[17] S. Newman, Building Microservices: Designing Fine-Grained Systems, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2021.

[18] T. Akidau, S. Chernyak, and R. Lax, Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing. Sebastopol, CA, USA: O'Reilly Media, 2018.

[19] S. Li, J. Hu, Y. Cui, and J. Hu, "Natural Language Processing in the Patent Domain: A Survey," arXiv preprint arXiv:2403.04105, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2403.04105

[20] Z. Fadika, M. R. Head and M. Govindaraju, "Parallel and distributed approach for processing large-scale XML datasets," 2009 10th IEEE/ACM International Conference on Grid Computing, Banff, AB, Canada, 2009, pp. 105-112, doi: 10.1109/GRID.2009.5353070.

[21] W. Enck et al., "Configuration management at massive scale: system design and experience," in IEEE Journal on Selected Areas in Communications, vol. 27, no. 3, pp. 323-335, April 2009, doi: 10.1109/JSAC.2009.090408.

[22] R. Al-Ali, N. Kathiresan, M. El Anbari, E. R. Schendel, and T. A. Zaid, "Workflow optimization of performance and quality of service for bioinformatics application in high performance computing," J. Comput. Sci., vol. 15, pp. 3–10, 2016. doi: 10.1016/j.jocs.2016.03.005.

[23] M. Kirti, A. K. Maurya, and R. S. Yadav, "Fault-tolerance approaches for distributed and cloud computing environments: A systematic review, taxonomy and future directions," Concurrency Comput. Pract. Exper., vol. 36, no. 13, p. e8081, 2024. doi: 10.1002/cpe.8081.

[24] Chan, C.-Y., Felber, P., Garofalakis, M., & Rastogi, R. (2002). "Efficient filtering of XML documents with XPath expressions." *The VLDB Journal, 11*(4), 354–379. doi: 10.1007/s00778-002-0077-6.

[25] J. Son et al., "AI for Patents: A Novel Yet Effective and Efficient Framework for Patent Analysis," *in IEEE Access*, vol. 10, pp. 59205-59218, 2022, doi: 10.1109/ACCESS.2022.3176877.

[26] S. Jiang, J. Hu, C. L. Magee and J. Luo, "Deep Learning for Technical Document Classification," *in IEEE Transactions on Engineering Management*, vol. 71, pp. 1163-1179, 2024, doi: 10.1109/TEM.2022.3152216.

[27] R. Padilla, S. L. Netto and E. A. B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Niteroi, Brazil, 2020, pp. 237-242, doi: 10.1109/IWSSIP48289.2020.9145130.

[28] M. Kang, S. Lee and W. Lee, "Prior Art Search Using Multi-modal Embedding of Patent Documents," *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Busan, Korea (South), 2020, pp. 548-550, doi: 10.1109/BigComp48618.2020.000-6.

[29] A. J. C. Trappey, C. V. Trappey, U. H. Govindarajan and J. J. H. Sun, "Patent Value Analysis Using Deep Learning Models—The Case of IoT Technology Mining for the Manufacturing Industry," *in IEEE Transactions on Engineering Management*, vol. 68, no. 5, pp. 1334-1346, Oct. 2021, doi: 10.1109/TEM.2019.2957842.

ΠΑΡΑΡΤΗΜΑ Α : ΚΩΔΙΚΑΣ - DATASET

Ο κώδικας του προγράμματος είναι διαθέσιμος στο repository του WPI+ στο GitHub στον σύνδεσμο: https://github.com/cs1msa/WPIplus

Το dataset με όλες τις virtual patents όπως δημιουργήθηκαν μέσω του PatentFusion για όλα τα γραφεία πατεντών βρίσκεται στο αποθετήριο του TUWien εδώ: https://researchdata.tuwien.ac.at/