



**DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA DEI SISTEMI
DEL TERRITORIO E DELLE COSTRUZIONI**

**RELAZIONE PER IL CONSEGUIMENTO DELLA
LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE**

***Definizione del processo di scrittura di una query per
la ricerca di documenti tecnici***

SINTESI

RELATORI

Prof. Ing. Gualtiero Fantoni

*Dipartimento di Ingegneria Civile e
Industriale, Università di Pisa*

Ing. Giovanni De Santis

Erre Quadro S.r.l.

Dott. Filippo Chiarello

*Dipartimento di Ingegneria Civile e
Industriale, Università di Pisa*

IL CANDIDATO

Antonietta Pagnozzi

a.pagnozzi@outlook.it

Sessione di Laurea Magistrale del 30/09/2020
Anno accademico 2018/1019
Consultazione non consentita

Sommario

La presente tesi si propone come obiettivo quello di spostare l'attenzione delle attività di scrittura della query dall'aspetto tecnico-informatico verso un approccio gestionale, con focus sulla progettazione di un processo che risulta ad ora non strutturato e ingegnerizzato. Tale approccio può garantirne la ripetibilità, la trasparenza della ricerca, la completezza dei risultati che vengono forniti in output e che forniscono l'input per le analisi successive. Inoltre, questo rappresenta il primo passo da compiere per poterlo automatizzare.

Le attività di tesi sono state svolte in collaborazione con Erre Quadro S.r.l, azienda spin-off dell'Università di Pisa, che si occupa, come attività principale, dell'analisi di proprietà intellettuale con focus sulla ricerca brevettuale. Il processo studiato è quello della *patent retrieval*, ovvero il ritrovamento di documenti brevettuali, e fa riferimento agli strumenti utilizzati internamente all'azienda.

Abstract

The goal of my work is to shift the focus of the query from the informatic aspect to an engineering managerial approach. Nowadays query design is mainly based on the experience and intuition of the analyst and it is highly time consuming. The design approach can guarantee the repeatability, the transparency of the research, the comprehensiveness of the results that characterize in output and that provide input for subsequent analysis.

Moreover, this may represent the first step to be taken in order to automate it.

This project work was carried out in collaboration with Erre Quadro S.r.l, a spin-off company of the University of Pisa, leader in intellectual property management and patent research.

The process is defined with a focus on patent retrieval, i.e. the discovery of patent documents, and it refers to the tools used within the company.

1 Introduzione

Il flusso di lavoro di analisi dei dati prevede come macro-attività l'importazione, l'analisi, la comprensione e la comunicazione delle informazioni. Questo genera risultati utili se e solo se in input vengono utilizzati dati che soddisfano determinati requisiti inerenti alla ricerca. Nel campo delle ricerche professionali¹ sono gli esperti *information retrieval* che provvedono a fornire i dati sotto forma di documenti tecnici ricercandoli entro vincoli particolari, con database e strumenti specializzati e con limiti di tempo e budget. Queste ricerche si differenziano in maniera sostanziale dalla ricerca via web sia per gli scopi, sia perché si basano sulle informazioni contenute in documenti tecnici, che sono multi-pagina, multi-modali, multi-lingua, semi-strutturati, ricchi di metadati e composti da un linguaggio prevalentemente tecnico-scientifico². Tra queste, le attività di *patent retrieval*, ovvero di ritrovamento dei documenti brevettuali, sono le più complesse.

Un brevetto è uno strumento della proprietà intellettuale utilizzato per tutelare legalmente l'inventore, fornendogli il diritto di vietare a terzi di produrre, usare, mettere in commercio, vendere o importare l'invenzione³. Non vi è invece alcun limite alla consultazione dei documenti brevettuali in quanto sono interamente pubblici. Per rendere difficile il ritrovamento del singolo documento così da tenere l'invenzione quanto più possibile segreta, viene utilizzato un linguaggio composto da un vocabolario molto complesso, con termini vaghi e astratti ma allo stesso tempo estremamente tecnici e specifici, creando così il problema della ricerca brevettuale linguisticamente impegnativa⁴ che rende particolarmente complesso il compito di *patent retrieval*.

La query di ricerca è il ponte tra l'attuale comprensione del bisogno di informazioni e il sistema di accesso a queste⁵, tipicamente contenute in database di documenti espressi in linguaggio naturale. La query utilizzata è quindi di tipo testuale. A differenza dei database relazionali⁶, non vi sono al momento degli standard per la sintassi di ricerca, che dipende prevalentemente dal particolare database interrogato e dalla modalità con cui i documenti vi sono memorizzati e i dati sono indicizzati.

¹ Tony Russell-Rose, J. C. (2018). Information retrieval in the workplace: A comparison of professional search practices. Springer

² Shalaby, W., & Zadrozny, W. (2019) Patent retrieval: a literature review. Springer

³ Articolo 66, Codice della proprietà industriale (D.lgs. 10 febbraio 2005, n. 30)

⁴ Shalaby, W., & Zadrozny, W. (2019) Patent retrieval: a literature review. Springer

⁵ Hearst, M. A. (2009). Query Specification. In *Search user interfaces* (p. 91-119). Cambridge University Press.

⁶ Database che archiviano le informazioni in tabelle interrelate, che consentono ricerche e aggiornamenti incrociati e utilizzano un linguaggio standardizzato per la ricerca.

2 Conoscenza di base

In letteratura sono analizzate due tipologie di query di tipo testuale: query composta da un intero documento oppure composta da una sequenza di keyword. Nella trattazione si farà riferimento a questa seconda tipologia formata dagli elementi in Tabella 1.

Elementi	Keyword	Parola singola
	Multiword	Insieme di keyword, che combinate suggeriscono un argomento. Si utilizzano quando si esprime un concetto mediante una perifrasi ⁷ . Una multiword viene inserita tra “ ”.
Operatori	Booleani	Da inserire tra keyword o multiword. Le operazioni effettuate dal database sono di tipo insiemistico e seguono la logica dell'algebra booleana.
	Posizione	Specificano in quale campo del documento brevettuale vanno ricercate le keyword.
	Prossimità	Specifica il numero n di parole che possono essere contenute tra le keyword che costituiscono la multiword. Se utilizzato vengono ritrovati i documenti contenenti sia le multiword così come scritte in query sia quelle separate da n parole ⁸ .
Funzioni ⁹	Stemma	Il procedimento di stemma viene svolto dall'algoritmo di ricerca utilizzato dal database. Questo prevede la riduzione della keyword alla propria radice e la ricerca di questa. Il software restituisce tutti i documenti che contengono non solo la parola ricercata ma anche quelli che contengono le parole con la stessa radice.
	Carattere speciale *	Posto ad inizio o fine keyword in sostituzione di prefissi o suffissi e svolge il ruolo di jolly, ovvero può assumere qualsiasi valore. Il database restituisce i documenti che contengono le parole composte dalla keyword ricercata con qualsiasi espressione al posto dell'asterisco.

Tabella 1 - Elementi query

Gli elementi inseriti in query influiscono sulla qualità del risultato. Per tale valutazione vengono usate le metriche di *precision*, indice di pulizia, e *recall*, indice di completezza. In riferimento alla Figura 1, per una certa ricerca e per una data query, si distinguono tra i documenti brevettuali totali: A insieme dei documenti rilevanti ma non ritrovati, B quelli rilevanti e ritrovati, C quelli non rilevanti per la ricerca ma ritrovati. Possiamo quindi definire le misure tramite la numerosità di tali insiemi attraverso Equazione 1 ed Equazione 2:

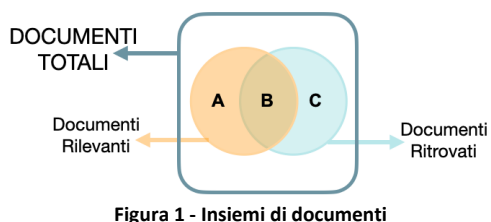


Figura 1 - Insiemi di documenti

$$Precision = \frac{\# \text{ documenti rilevanti in output}}{\# \text{ documenti ritrovati}} = \frac{\#B}{\#B + \#C}$$

Equazione 1 - Formula precision

$$Recall = \frac{\# \text{ documenti rilevanti output}}{\# \text{ documenti rilevanti}} = \frac{\#B}{\#A + \#B}$$

Equazione 2 - Formula recall

⁷ Procedimento espressivo consistente nell'usare, anziché un termine unico, un insieme di parole per mezzo delle quali si spiega una cosa, un concetto o una persona senza utilizzare il termine proprio [Zanichelli]

⁸ Escluse le stop-word, ovvero le parole comuni che non aggiungono informazioni per esempio articoli e pronomi. Ogni database ne ha una propria lista.

⁹ La differenza tra le due sta nel fatto che per lo stemma è l'algoritmo di *stemming* a definire la base della keyword da ricercare, mentre per il carattere speciale l'analista può modulare e decidere dove tagliare. Inoltre, l'asterisco può essere utilizzato anche con il codice delle classi, mentre lo stemma no.

La misura di recall è particolarmente complessa da effettuare in quanto non è nota a priori la numerosità dei documenti rilevanti per una data ricerca, essendo proprio ciò che va ritrovato con la query. Nella trattazione verranno utilizzate le metriche facendo riferimento alla definizione degli insiemi in Figura 1 e al fatto che queste si comportano in maniera inversamente proporzionale: con l'aumentare della numerosità dei documenti in output si ha un miglioramento dell'indice di *recall* rispetto a quello di *precision*¹⁰. I requisiti richiesti in termini di metriche variano a seconda delle tipologie di analisi: analisi dello stato dell'arte tra cui le analisi di anteriorità, *freedom to operate*, validità/invalidità e *design around*, hanno come requisito la completezza del set, ovvero la *recall*; le analisi statistiche, tra cui mappatura brevettuale, *foresight* tecnologico e monitoraggio della tecnologia, hanno come requisito la pulizia del set, cioè *precision*.

La ragione di fondo per cui vengono utilizzati i brevetti per effettuare tali analisi è che questi racchiudono il 90% dell'intero sapere tecnico¹¹ in maniera pubblica e liberamente consultabile. Le informazioni per effettuare le varie analisi sono contenute nelle diverse parti del documento brevettuale composto da dati bibliografici: inventori, assegnatari, giurisdizioni, date, classi, citazioni e da dati testuali: titolo, *abstract*, *claim*, *description*. A seconda della tipologia di analisi saranno necessarie alcune informazioni piuttosto che altre. Si riporta un approfondimento in Appendice A.

Tutti questi elementi vanno considerati in fase di scrittura della query di ricerca. Le modalità con cui attualmente gli analisti svolgono tale fase, dal momento in cui si manifesta il bisogno di informazioni fino ad arrivare ad un set di documenti definito, non è delineato allo stato dell'arte. Lo scopo è quindi quello di tracciare l'insieme delle attività e delle decisioni necessarie per affrontare la ricerca di un set brevettuale. In riferimento alla metodologia Cardillo & Folino¹² sono state svolte le attività tramite due differenti approcci: *bottom-up* e *top-down*.

3 Affiancamento in azienda

L'approccio *bottom-up*, utilizzato per l'attività di affiancamento, ha come obiettivo di raccogliere la maggior quantità di dati a partire dagli esperti del dominio, dati che

¹⁰ Lupu, M., Piroi, F., & Stefanov, V. (2017). Quantifying the Difference: IR Measures. In *Current Challenges in Patent Information*, Second Edition Retrieval (p. 53-56).

¹¹ S. Vieider, I. Lantschner, A. Franzoi. (2012) *Ricerche brevetti e marchi alla portata di tutti!* Camera di commercio di Bolzano Brevetti e marchi.

¹² Cardillo, E., & Folino, A. (2008). Terminologia e documenti per la formalizzazione standardizzata della conoscenza.

successivamente vengono organizzati in un modello¹². Il risultato viene riportato nel diagramma della Tabella 2.

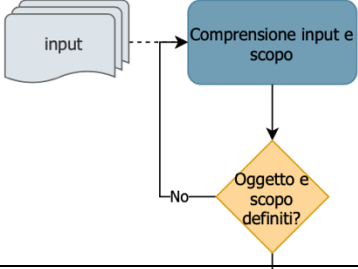
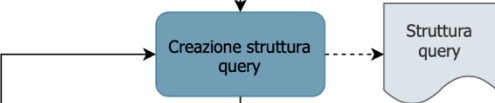
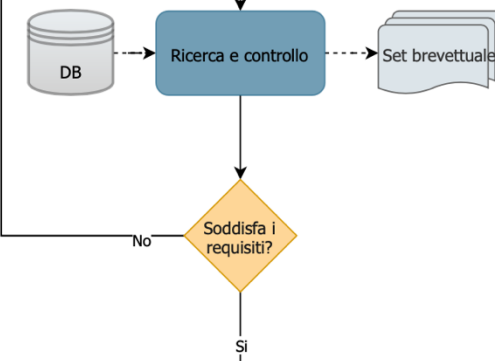
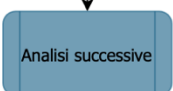
Diagramma di flusso	t ¹³	Descrizione	Metodo / strumento
	18%	Analista prende visione dell'input e comprende l'argomento della ricerca e lo scopo. Il trigger del processo viene fornito da un input di tipo testuale, multimediale o interazione verbale.	Discrezione dell'analista / Ricerche web- Contatto cliente
	4%	Scelta delle entità e degli elementi che costituiranno la query, tra quelli definiti in Tabella 1.	Due tipologie di modelli di progettazione
	64%	L'analista definisce un blocco della struttura da espandere, trova le keyword o sul database aziendale o mediante la ricerca di sinonimi su dizionario, le concatena, le memorizza e passa al blocco successivo. Si ripete maniera iterativa fino a che viene svolta su tutti i blocchi. In Appendice B il workflow relativo.	Foglio note Foglio Excel Knowhow aziendale Dizionari online
	14%	L'analista ricerca la query nel database brevettuale e prende visione dei risultati. Attraverso un controllo qualitativo sui primi risultati comprende se il set brevettuale soddisfa i requisiti per le analisi successive o se è necessario effettuare cambiamenti alla query.	Discrezione dell'analista
		Una volta definito il set di brevetti l'analista procede con le analisi successive.	

Tabella 2 - Processo di scrittura della query

Il processo deriva dallo svolgimento dell'attività di affiancamento dell'analista di Erre Quadro S.r.l. tramite osservazione diretta. Lo strumento utilizzato è il foglio di raccolta dati e un

¹³ Durata media in percentuale della macro-attività, valutata su una durata media di 140 minuti.

sistema di annotazione libera con l'obiettivo di tener traccia del flusso di lavoro, delle modalità delle attività svolte, degli strumenti utilizzati, delle iterazioni effettuate e del tempo impiegato dall'analista. Queste informazioni sono state scomposte per attività, strumenti e tempi e poi sintetizzate in un unico modello riportato in sintesi in Tabella 2.

Nelle analisi successive le attività sono state messe in relazione rispetto a tre grandezze, così da effettuare considerazioni sul processo e definire le modifiche da apportare.

impatto sull'efficienza: quanto la durata dell'attività influisce sulla durata globale;

l'efficacia globale del processo: quanto lo svolgimento del processo impatta sul raggiungimento del risultato;

impatto dell'esperienza dell'analista sui risultati: grado di esperienza necessaria da parte dell'analista per il corretto svolgimento dell'attività.

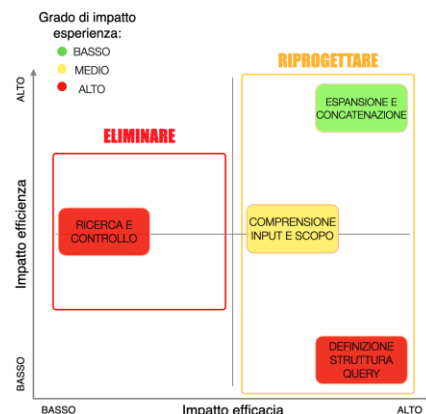
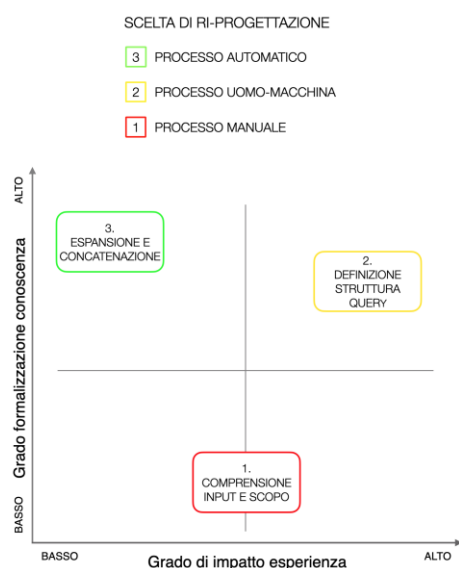


Figura 2 - Grafico e spiegazione efficacia-efficienza-conoscenza

In ottica ISO 9001 trattando il processo come speciale, si decide di eliminare il controllo finale in quanto eventuali "difettosità" risultano evidenti solamente dopo l'utilizzo della query. Di conseguenza sono stati pianificati strumenti e metodi sulle altre attività, basandosi su scelte di riprogettazione, prese in base all'impatto dell'esperienza e al grado di formalizzazione della conoscenza necessaria, così come mostrato in Figura 3.



3. Richiede bassa esperienza formalizzabile del tutto. Per questo si propone di rendere il processo automatico.

2. Richiede attività decisionali da parte dell'analista prese mediante un processo logico che è possibile estrarre e formalizzare. Dato che l'impatto dell'esperienza sui risultati è molto elevato la proposta è quella di strutturare un processo svolto tramite la collaborazione uomo-macchina.

1. Richiede attività di lettura ed elaborazione delle informazioni con un processo di conoscenza basato su esperienza e consapevolezza difficilmente replicabile da una macchina. La scelta per la riprogettazione è quella lasciare l'attività manuale fornendo un metodo.

Figura 3 - Grafico e spiegazione esperienza-formalizzazione conoscenza e scelte di riprogettazione

Affinché sia possibile attuare le scelte di riprogettazione viene estratta e formalizzata la conoscenza alla base del processo di scrittura della query.

3.1 Definizione di un modello di progettazione della query e degli obiettivi

Per fornire una base alle attività successive si rielabora prima la conoscenza disponibile.

Per quanto riguarda i modelli di progettazione sulla base di quelli esistenti e sulla teoria dell'analisi funzionale si perfeziona un unico modello che prevede la scomposizione del sistema da ricercare in cinque entità: Soggetto, Oggetto, Verbo, Caratteristica e Ambito, sufficienti¹⁴ per una definizione completa. Si utilizza logica funzionale in quanto la descrizione del funzionamento di una invenzione deve essere necessariamente fornita in un brevetto per essere approvato¹⁵, quindi la ricerca di parole che indicano il funzionamento all'interno della query porta a livelli di recall elevati¹⁶.

Gli obiettivi dell'attività di *patent retrieval*, sono definiti in base alle tipologie di analisi che si trovano a valle di tale attività, enunciate al paragrafo 2 e all'osservazione dell'analista. Infatti, nella pratica si distinguono tre obiettivi di *patent retrieval*: anteriorità stretta, che precede le analisi di FTO e di invalidità/validità, anteriorità larga, che precede le analisi di anteriorità e di *design around* e ricerca statistica, che precede l'analisi di tipo statistico.

4 Reverse Engineering

Il secondo approccio top-down si focalizza sulla definizione di un modello di expertise al fine di filtrare la conoscenza acquisita¹² ed è applicato sulle attività di reverse engineering effettuate mediante le ricerche svolte dagli analisti di Erre Quadro S.r.l. L'input della attività è un insieme di file, uno per ogni ricerca, che contiene sia la query utilizzata per la ricerca finale sia le intermedie¹⁷. Sono state svolte attività di scomposizione di ogni query divisa secondo le entità e gli scopi descritti al 3.1 e ulteriormente divisa secondo gli elementi della Tabella 1.

Quindi sono state effettuate delle analisi su query finali e sulle iterazioni, guardando i cambiamenti della struttura e ricostruendo le decisioni prese dall'analista. Da qui è stato possibile definire una tabella che rappresenta l'impatto sulle metriche degli elementi della query, riportata per completezza in Appendice C, e da questa è stato generato un modello delle decisioni a forma matriciale, riportato nella Figura 4, che contiene la base della conoscenza applicata dall'analista in fase di strutturazione della query.

¹⁴ Ma non necessarie, ovvero in fase di strutturazione della query non devono obbligatoriamente essere presenti tutte le entità.

¹⁵ Una invenzione per essere brevettata deve dimostrare di soddisfare i requisiti di novità, step inventivo e applicabilità industriale. Questo viene dimostrato attraverso la descrizione nel dettaglio dell'invenzione, del funzionamento e della sua applicazione. Per questo sono rilevabili tali informazioni nei brevetti.

¹⁶ Bonaccorsi, A., Fantoni, G., Aprea, R., & Gabelloni, D. (2019). Functional Patent Classification.

¹⁷ Come è possibile vedere nel diagramma di flusso l'attività di scrittura della query si svolge in maniera iterativa. Tendenzialmente l'analista scrive varie query prima di decidere quale utilizzare. Queste vengono tutte memorizzate nel file Excel della ricerca.

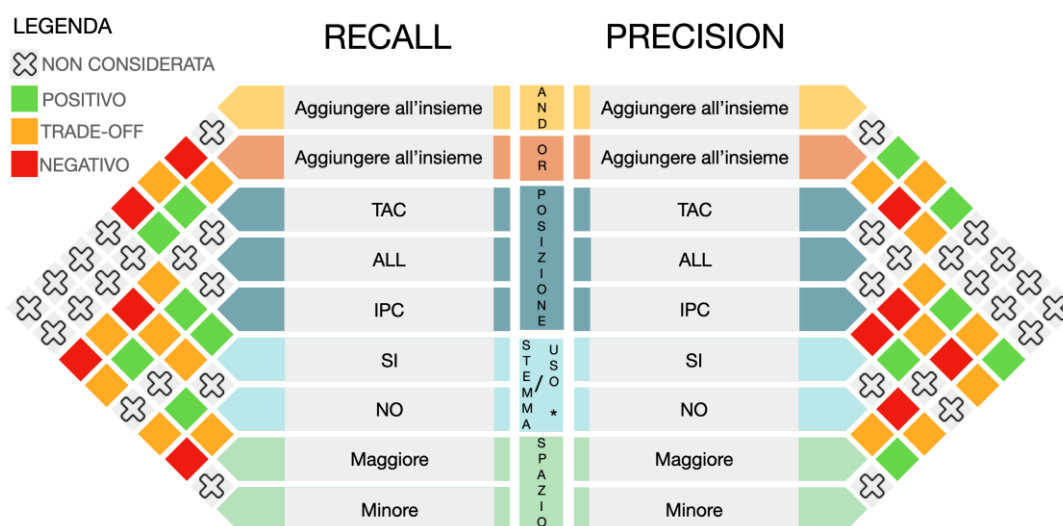


Figura 4 - Matrici delle decisioni

Tramite queste matrici l'analista può essere supportato nella decisione della struttura. Mediante un approccio visivo si comprende l'impatto che una data decisione ha rispetto alle metriche. La parte sinistra contiene la matrice relativa alla metrica di *recall* e a destra la *precision*. Per ogni coppia di elementi da definire della query¹⁸ viene valutato se l'impatto sulla metrica è negativo, positivo o è una scelta di *trade-off*. Questa si verifica nel caso in cui uno dei due elementi avesse un impatto negativo, l'altro un impatto positivo. L'obiettivo è chiaramente quello di raggiungere un compromesso globale tra la metrica di *precision* e *recall* compatibilmente con lo scopo della ricerca.

5 Ri-definizione del processo

Attraverso l'ultima attività di formazione al lavoro sono state svolte in maniera autonoma e completa diverse ricerche di *patent retrieval*. In questa fase è stato delineato un processo di scrittura della query che segue il flusso di lavoro descritto in Tabella 2 ma con strumenti e metodi ridefiniti. Si riporta in Tabella 8 l'intero processo mediante la rappresentazione degli *asset* operativi per un processo di Data Science¹⁹. Lo strumento prevede la strutturazione del processo in: dati, materia prima del processo; software, di interfaccia tra macchine e persone; modello ponte tra software e metodi; metodi, utilizzati dalla componente umana per la gestione del processo²⁰.

¹⁸ Gli operatori di posizione nella matrice prendono in considerazione quelli utilizzati nella pratica delle ricerche, effettuate in azienda. TAC sta per titolo-Abstract-Claim e significa che la ricerca avviene solo in quegli specifici campi brevettuali, ALL sta per tutti i campi, IPC sta per la classe, deve essere seguito dal codice della classe da ricercare. Per spazio si intende l'opzione prossimità e si considera solo per multiword.

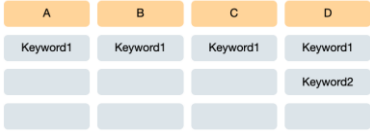
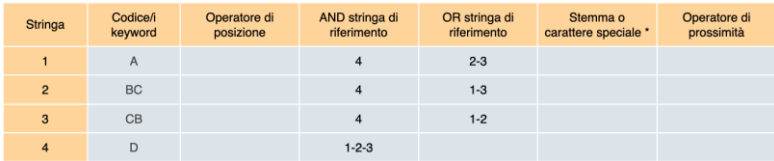
¹⁹ Lo strumento è stato proposto dal Dott. Gianluca Ruggi nella Tesi di laurea magistrale presentata in sessione di Laurea Magistrale il giorno 29/04/2020, con relatori Prof. Ing. Gualtiero Fantoni e Dott. Filippo Chiarello, dal titolo "Metodi di Progettazione per la Data Science".

²⁰ Non devono necessariamente essere presenti tutte e quattro le componenti, in Tabella 8 non si inseriscono le righe relative agli elementi mancanti.

1. Comprensione input e scopo																	
Obiettivo	Dati	Input alla ricerca															
	Software	Piattaforma di interfaccia															
	Modelli	Obiettivo di ricerca Anteriorità larga Anteriorità stretta Ricerca statistica Tabella 3 – Obiettivo della ricerca															
	Metodi	L’analista prende visione dell’input e definisce lo scopo della ricerca segnandolo nella tabella.															
Comprensione input	Dati	Input alla ricerca															
	Metodi	Check list Per quale motivo l’invenzione viene realizzata? Come può essere utilizzata l’invenzione? Qual è la funzione principale dell’invenzione? Descrivere la funzione con un verbo ed un oggetto o soggetto Qual è la caratteristica principale invenzione? Qual è l’ambito della ricerca dell’invenzione? Figura 5 - Check list Dai dati in input l’analista appunta gli elementi seguendo le domande della check list. Fornire una risposta a tali domande guida l’analista alla comprensione della ricerca da effettuare. Più le risposte sono espresse con una frase semplice composta da elementi essenziali come verbo, soggetto e complemento oggetto più l’analista avrà ben chiaro il focus della ricerca da effettuare.															
Definizione degli elementi	Dati	Risposte check list															
	Software	Piattaforma di interfaccia															
	Modelli	<table><tr><th>Soggetto</th><th>Oggetto</th><th>Verbo</th><th>Caratteristica</th><th>Ambito</th></tr><tr><td></td><td>Espressione1_Oggetto</td><td></td><td></td><td>Espressione1_Ambito</td></tr><tr><td></td><td>Espressione2_Oggetto</td><td></td><td></td><td></td></tr></table> Tabella 4 - Definizione delle entità²¹ In tabella si memorizzano le espressioni per ogni entità, possono essere presenti nessuna, una o più espressioni e possono essere parola singola o perifrasi⁷.	Soggetto	Oggetto	Verbo	Caratteristica	Ambito		Espressione1_Oggetto			Espressione1_Ambito		Espressione2_Oggetto			
	Soggetto	Oggetto	Verbo	Caratteristica	Ambito												
	Espressione1_Oggetto			Espressione1_Ambito													
	Espressione2_Oggetto																
Metodi	Procedura: l’analista seguendo il modello di progettazione definito al paragrafo 3.1 decide quale entità definire e per ognuna fornisce una o più espressioni. In questa fase considera per ogni espressione la presenza di perifrasi da aggiungere in tabella o la presenza di polisemie ²² da dover disambiguare attraverso l’aggiunta di espressioni che precisano il significato.																
2. Progettazione della struttura della query																	
Definizione keyword	Dati	Tabella 4 - Definizione delle entità [compilata]															
	Software	Piattaforma di interfaccia															
	Modelli	<table><tr><th></th><th>Keyword</th><th>Codice keyword</th></tr><tr><td>Espressione1_Oggetto</td><td>Keyword1_Espressione1_Oggetto</td><td>A</td></tr><tr><td>Espressione2_Oggetto</td><td>Keyword1_Espressione2.1_Oggetto</td><td>B</td></tr><tr><td>Espressione2_Oggetto</td><td>Keyword1_Espressione2.2_Oggetto</td><td>C</td></tr><tr><td>Espressione1_Ambito</td><td>Keyword1_Espressione1_Ambito Keyword2_Espressione1_Ambito</td><td>D</td></tr></table> Tabella 5 - Associazione Espressione-keyword-elemento identificativo In tabella si memorizzano per ogni espressione una keyword a cui viene poi associato un codice keyword identificativo, alfabetico nell’esempio.		Keyword	Codice keyword	Espressione1_Oggetto	Keyword1_Espressione1_Oggetto	A	Espressione2_Oggetto	Keyword1_Espressione2.1_Oggetto	B	Espressione2_Oggetto	Keyword1_Espressione2.2_Oggetto	C	Espressione1_Ambito	Keyword1_Espressione1_Ambito Keyword2_Espressione1_Ambito	D
	Keyword	Codice keyword															
Espressione1_Oggetto	Keyword1_Espressione1_Oggetto	A															
Espressione2_Oggetto	Keyword1_Espressione2.1_Oggetto	B															
Espressione2_Oggetto	Keyword1_Espressione2.2_Oggetto	C															
Espressione1_Ambito	Keyword1_Espressione1_Ambito Keyword2_Espressione1_Ambito	D															

²¹ Viene compilata a titolo di esempio. Nelle tabelle successive si procede con il medesimo esempio

²² Proprietà di un segno linguistico di avere più significati. [Zanichelli]. Rientrano quindi nella definizione parole che acquisiscono significato diverso a seconda del contesto in cui vengono utilizzate.

Inizializzazione keyword	Metodi	Per ogni espressione di ogni entità l'analista inserisce una o più keyword con cui iniziare il processo di ricerca dei sinonimi nell'espansione dei termini. Nel caso l'espressione sia una perifrasi allora si divide in tante espressioni quante sono le parole che la compongono e per ognuna di queste si definisce una keyword a cui verrà assegnato un proprio codice.
	Dati	Tabella 5 - Associazione Espressione-keyword-elemento identificativo [compilata]
	Software	Piattaforma di interfaccia
Definizione struttura	Modelli	 Tabella 6 - Vettori keyword Viene trasposta la parte di Tabella 5 relativa a codice keyword e keyword in modo da memorizzare, in forma vettoriale, le keyword da espandere mediante sinonimi che poi vanno concatenate per generare la query.
	Dati	Tabella 3 – Obiettivo della ricerca Tabella 5 - Associazione Espressione-keyword-elemento identificativo [compilata] Tabella 6 - Vettori keyword
	Software	Piattaforma di interfaccia
Definizione struttura	Modelli	 Tabella 7 - Definizione della struttura della query La definizione della struttura della query nella tabella regola le operazioni da effettuare per scrivere la query con le colonne della Tabella 6 una volta espansi i termini. Si utilizzano come riferimento i codici keyword precedentemente definiti e la sintassi degli operatori di posizione, stemma e prossimità devono essere quelli utilizzati per il particolare database. Quindi va definito in primis quale codice keyword inserire nella relativa stringa. Se si prevede di dover concatenare due parole per generare multiword questi vanno inseriti insieme nella stessa riga della tabella. Per ogni stringa l'analista decide l'operatore di posizione, se utilizzare o no la funzione di stemma o inserire il carattere speciale * e l'operatore di prossimità, nel caso di multiword.
	Metodi	In questa fase l'analista deve svolgere un processo decisionale che può essere supportato dalle - Matrici delle decisioni. Per prima cosa devono essere consultate le regole e nel caso in cui la ricerca effettuata rientra tra quelle delle regole definite. Se no l'analista mediante la propria conoscenza e il supporto delle matrici delle decisioni compila la tabella. La tabella prevede di considerare prima gli operatori di AND e OR in relazione agli operatori di posizione e quindi di utilizzare le prime 6 combinazioni delle matrici. Poi una volta definiti si considerano gli operatori di posizione in relazione alle altre scelte.
3. Espansione		
	Dati	Tabella 6 - Vettori keyword [compilato]

		Database aziendale interno
	Software	Linguaggio di programmazione
	Modelli	Algoritmo di espansione, per ogni vettore keyword della tabella in input l'algoritmo trova le keyword sinonimo e le aggiunge alla tabella.
4. Concatenazione		
	Dati	Tabella 6 - Vettori keyword [espanso] Tabella 7 - Definizione della struttura della query
	Software	Linguaggio di programmazione
	Modelli	Algoritmo di concatenazione, per ogni stringa definita nella tabella di struttura della query l'algoritmo prende in input in vettore keyword relativo, effettua le operazioni di concatenazione delle keyword vettore o combinazione delle keyword se sono presenti in input due codici keyword, per generare multiword. Si riporta in Appendice D un esempio di codice.

Tabella 8 - Processo di scrittura della query per la ricerca dei documenti brevettuali

5.1 Considerazioni finali

Il lavoro svolto ha avuto come risultato finale la definizione delle attività, della conoscenza di base e degli *asset* necessari per la strutturazione e l'automazione del processo di scrittura di una query per i documenti tecnici, in particolare quelli brevettuali.

Cambiando l'approccio al *patent retrieval* da quello prevalentemente informatico a quello gestionale è stato possibile suddividere il problema in due sotto-problemi, quello di strutturazione e quello di espansione della query. Il primo viene risolto nella presente trattazione mediante la definizione dei metodi e degli strumenti del processo, il secondo richiede, invece, uno studio futuro dal punto di vista tecnico-informatico, ma notevolmente semplificato rispetto al problema iniziale. Infatti, tramite il processo gestionale di strutturazione della query qui definito, il problema di espansione viene ridotto a ricerca di parole sinonimo, non risentendo della linguistica impegnativa relativa all'analisi brevettuale, risolta mediante la prima fase di strutturazione della query. Così facendo si sono minimizzate le distanze con il campo di *information retrieval* che presenta una letteratura più matura che ha raggiunto risultati, in termini di metriche, superiori. Oltre a questo, gli sviluppi futuri da realizzare per mettere in atto tale processo in ottica di automazione sono segnati in rosso nella Tabella 8 e riguardano l'organizzazione di un database interno compatibile con la struttura vettoriale, la creazione di un'interfaccia di supporto per l'analista che prevede la memorizzazione delle tabelle e la codifica delle matrici delle decisioni in albero delle decisioni, da utilizzare nella fase di definizione della struttura e la programmazione degli algoritmi per la concatenazione e per l'espansione delle keyword.

Appendice A

Attraverso un processo di sintesi e di rielaborazione delle informazioni contenute nell'articolo di review dello stato dell'arte²³ e lo stato dell'arte di Erre Quadro S.r.l, viene elaborata la Tabella 9 dove si evidenzia il campo brevettuale dal quale trarre informazioni a seconda dell'analisi da svolgere.

Tipologie di analisi		Campo del brevetto	Inventori	Assegnatari	Giurisdizioni	Date	Classi	Citazioni	Titolo	Abstract	Claim	Description
Analisi dello stato dell' arte	Anteriorità Verificare l'esistenza di una data soluzione								✓	✓	✓	✓
	Freedom to Operate Verificare che una data soluzione non entri in contraffazione con un brevetto esistente										✓	
	Validità/ Invalidità Difendere una domanda di brevetto o contestare un brevetto concorrente ²⁴								✓	✓	✓	✓
	Design Around Trovare nuovi spazi di brevettazione e soluzioni tecniche ²⁵								✓	✓	✓	✓
Analisi statistiche	Mappatura brevettuale Creazione di un overview della tecnica relativamente ad un problema tecnico		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Foresight tecnologico Fotografare lo stato della tecnica di un determinato settore ¹²		✓	✓	✓	✓	✓	✓				
	Monitoraggio della tecnologia Identificazione del portafoglio brevetti di un competitor o di una tecnologia		✓	✓	✓	✓	✓	✓				

Tabella 9 - Relazione tra informazioni - campi brevettuali e scopi della ricerca

Appendice B

Per completezza si riporta il diagramma di flusso seguito dall'analista per lo svolgimento della espansione e concatenazione dei termini della query.

²³ Shalaby, W., & Zadrozny, W. (2019). Patent retrieval: a literature review. Springer-Verlag London .

²⁴ Bonino, D. (2009). Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. Springer

²⁵ Lin, V. Patent Trademark Blog . Tratto da <http://www.patenttrademarkblog.com/how-to-design-around-patents/>

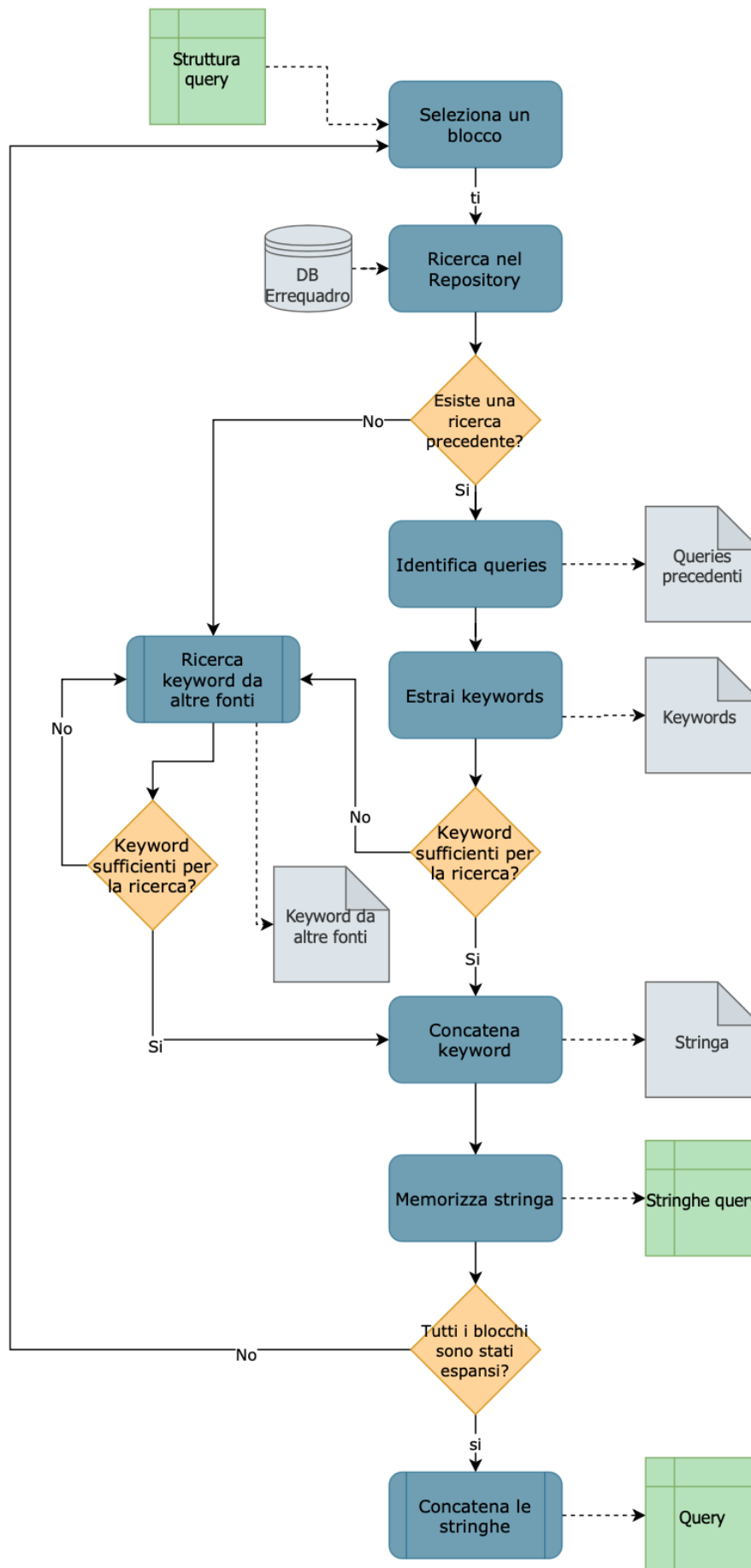


Figura 6 - Processo di espansione e concatenazione

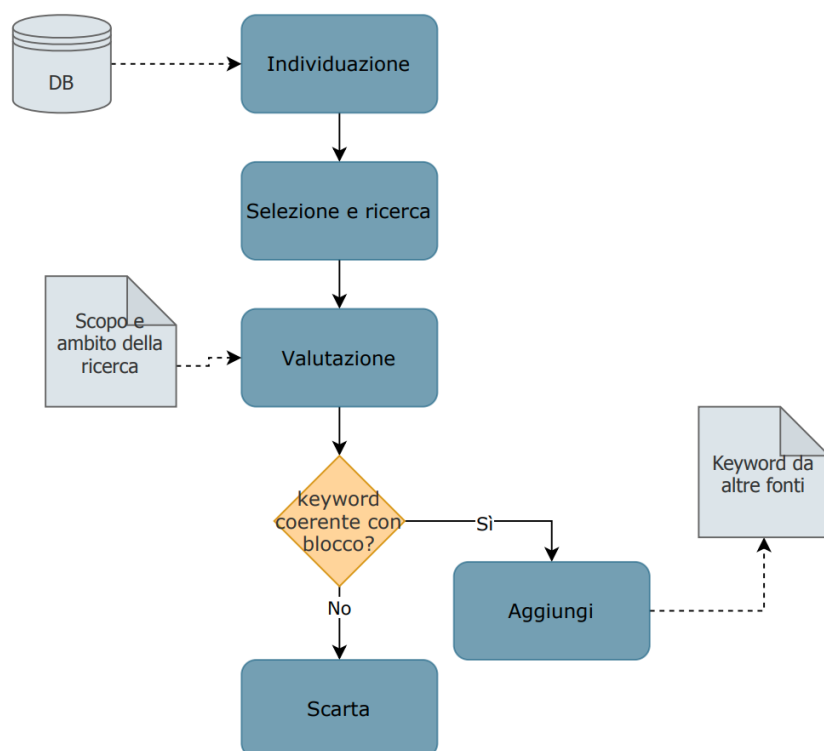


Figura 7 - Processo di ricerca da altre fonti

Appendice C

La tabella di impatto degli elementi fa riferimento alla Figura 1 - Insiemi di documenti

Legenda						
A	Insieme di documenti falsi negativi	↑	Viene aggiunto o inserito	⊖	Numerosità diminuisce	Simboli in rosso: effetto dell'aggiunta
B	Insieme di documenti veri positivi	↓	Viene eliminato o non inserito	⊕	Numerosità aumenta	
C	Insieme di documenti falsi positivi	TAC	Titolo-Abstract-Claim	⊙	Numerosità rimane invariata	
B+C	Insieme di documenti in output	ALL	Tutti i campi	⊕⊖	Numerosità aumenta o rimane invariata	Simboli in blu: effetto della eliminazione
A+B	Insieme di documenti rilevanti	IPC	Classi	⊖⊙	Numerosità diminuisce o rimane invariata	

	Operatori di posizione			Keyword in AND o AND NOT		Keyword in OR		Stemma delle parole o carattere *		Spazio di prossimità	
	TAC	ALL	IPC	↑	↓	↑	↓	Si	No	↑	↓
#A											
#B											
#C											
#B+#C											
#A+#B											
Precision											
Recall											

Tabella 10 - Relazione tra numerosità dei documenti e elementi query

Appendice D

Di seguito si fornisce un esempio di codice, scritto in linguaggio Tidyverse mediante il software Rstudio utilizzato per effettuare la concatenazione delle parole o la combinazione tra le keyword per generare multiword e inserire l'operatore di prossimità. Nel codice presentato la scelta del vettore e dell'operatore di prossimità non è automatizzata ma viene inserito come input.

```
#-----INIZIALIZZARE KEYWORD-----
key1 <- Tabella 5.3.2$A
key1
key2 <- Tabella 5.3.2$B

# si converte in data frame
key1 <- as.data.frame(key1)
key2 <- as.data.frame(key2)

# si eliminano i caratteri nulli
key1 <- key1%>%
  filter(!is.na(key1)) #mettere il nome della colonna

key2 <- key2%>%
  filter(!is.na(key2))

#convertito da factor a character
key1 <- key1%>%
  mutate(key1=as.character(key1))
```



```

key2 <- key2%>%
  mutate(key2=as.character(key2))

#-----CONCATENARE STRINGA SINGOLA DI UN SINGOLO VETTORE-----

#Si definisce la stringa dal vettore singolo e si stampa a schermo o si
memorizza in tabella
key1_string <- key1 %>%
  summarize(text = str_c(key1, collapse = " ")) %>%
  pull(text)
view(key1_string)
print(key1_string)

write.table(key1_string1, file="string1.csv", quote=F, sep=";",
dec=",", na="", row.names=T, col.names=T)

#-----COMBINAZIONI DI MULTIWORDS-----
#ciclo due volte per creare e combinazioni
key1_vect <- key1 %>%
  pull(key1)
key2_vect <- key2 %>%
  pull(key2)

#creo una matrice con tutte le combinazioni in un verso
combination <- NA
for(j in 1:length(key2_vect)){
  for(i in 1:length(key1_vect)){
    combination[(j-1)*length(key1_vect)+i]<- str_c(key2_vect[j],
key1_vect[i], sep=" ")}}
#view(combination)

combination <- as.data.frame(combination)

combination <- combination%>%
  mutate(combination=as.character(combination))

#lo porto in un'unica stringa
string1 <- combination %>%
  summarize(text = str_c(combination, collapse = '" ')) %>%
  pull(text)

#view(string1)
string1 <- as.data.frame(string1)
string1 <- string1%>%
  mutate(string1=as.character(string1))

write.table(string1, file="string1.csv", quote=F, sep=";", dec=",",
na="", row.names=T, col.names=T)

#-----INERIRE LO SPAZIO DEFINITO DALL'UTENTE-----
string1_space <- string1 %>%
  mutate(str_replace_all(string1,'" ','~2 '))

#esportare in excel
write.table(string1_space, file="string1_space.csv", quote=F, sep=";",
dec=",", na="", row.names=T, col.names=T)

```