UNIVERSITÁ POLITECNICA DELLE MARCHE

FACOLTÁ DI INGENGERIA

Corso di Laurea triennale in Ingegneria Gestionale

SISTEMA PER LA DISAMBIGUAZIONE AUTOMATICA DEL NOME DEGLI AUTORI PRESENTI IN UN DATABASE BIBLIOMETRICO

SYSTEM FOR THE AUTOMATIC AUTHOR NAME DISAMBIGUATION IN A BIBLIOMETRIC DATABASE

Relatore:

Prof. Domenico Potena

Correlatore:

Prof. Diego D'Adda

Tesi di Laurea di:

Chiara Menna

Anno Accademico 2022/2023

Indice

Elenco tabelle

Elenco figure

Capitolo 1
Introduzione

Il presente progetto di ricerca ha come oggetto l'analisi della procedura di disambiguazione riguardante il nome degli autori presenti in un database bibliometrico, e l'analisi qualitativa dei dati relativi ai docenti universitari italiani, nonché ricercatori, e le loro rispettive pubblicazioni scientifiche.

Con l'articolo 16 della Legge 240 del 2010 [1], per la partecipazione ai concorsi nelle singole università per la qualifica di professore di I o II fascia, è previsto come requisito necessario il possesso dell'Abilitazione Scientifica Nazionale.

Per la verifica di quest'ultimo, si utilizzano degli indicatori, stabiliti dall' articolo 2 del DM del 08/08/2018, n.589 [2], che variano in base al Settore Scientifico Disciplinare (SSD), grazie ai quali si possono valutare i titoli e le ricerche scientifiche di ogni candidato.

In particolare, uno degli indicatori fa riferimento alle pubblicazioni e alle citazioni presenti nella più grande banca dati internazionale "Scopus".

Per questo motivo si è resa necessaria la procedura di disambiguazione, descritta nel Capitolo 2, tramite cui viene associato a ciascun docente presente nel dataset del Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR), l'identificativo Scopus corrispondente.

Si otterranno così informazioni riguardanti le pubblicazioni, le references e i loro i rispettivi autori.

Nel Capitolo 3, verrà riportata la struttura della base di dati, nonché una descrizione delle tabelle. Verranno poi approfonditi i casi di

ambiguità riscontrati, così da poter poi procedere con una quantificazione di essi e una possibile risoluzione.

Il Capitolo 4 sarà dedicato all'analisi qualitativa dei dati relativi alle pubblicazioni e alle references.

E infine, il Capitolo 5 presenterà le conclusioni del progetto di ricerca svolto e una descrizione di quelle che potrebbero essere le prospettive future della base di dati e delle relative criticità dell'algoritmo analizzate.

Capitolo 2
Procedura di disambiguazione

In questo capitolo verranno descritte le informazioni utili all'algoritmo che effettua la procedura di disambiguazione, la quale consentirà di ottenere un identificativo Scopus univoco, nonché la conseguente assegnazione delle informazioni di interesse per ogni docente.

2.1 Identificatore univoco

Vengono di seguito analizzati i dati necessari per l'ottenimento dell'identificatore Scopus al fine dell'identificazione univoca.

2.1.1 Descrizione dei dati

- Tre sono i settori disciplinari scelti, e descritti in Tabella 2.1: ING-IND/17, ING-IND/35 e SECS-P/06; i primi due bibliometrici e l'ultimo non bibliometrico poiché ogni SSD dipende da indicatori differenti e di conseguenza potrebbero esserci comportamenti citazionali diversi.

SSD	Descrizione settori
ING-IND/17	studia le metodologie ed i criteri generali che presiedono alla pianificazione, progettazione, realizzazione e gestione degli impianti industriali
ING-IND/35	studia aspetti progettuali, economici, organizzativi e gestionali in campo ingegneristico
SECS-P/06	settore dell'economia applicata

Tabella 2.1: Descrizione dei settori disciplinari

- I dati relativi ai docenti appartenenti ai tre settori disciplinari scelti sono stati ricavati dal MIUR e sono riportati nella Tabella 2.2.

Dati	Descrizione
Fascia	e.g. ordinario, associato, ricercatore
Nome	cognome e nome
Genere	M o F
Ateneo	ateneo di appartenenza
Facoltà	facoltà dell'ateneo
SSD	settore scientifico-disciplinare
SC	settore concorsuale
Dipartimento	dipartimento di afferenza all'interno della facoltà
d_statale	1 = statale, 0 = non statale
anno	anno di riferimento dei dati

Tabella 2.2: Descrizione dati forniti dal MIUR

- In riferimento alla voce *anno* riportata in Tabella 2.2, si specifica che l'arco temporale preso in considerazione è quello che va dal 2001 al 2021. Vengono poi aggiunti due ulteriori campi, descritti in Tabella 2.3, fondamentali ai fini della disambiguazione.

Definizione	Descrizione
id_originale	id progressivo che consente di individuare in maniera univoca l'autore a partire dal suo nome e cognome
is_duplicate	variabile inserita per verificare che non siano presenti individui con stesso nome e cognome, con lo stesso SSD e nello stesso anno

Tabella 2.3: Descrizione campi aggiuntivi

- Dalle ricerche effettuate su Scopus si ottengono diverse informazioni, quelle prese in considerazione sono riportate in Tabella 2.4.

Dati	Descrizione
dc_identifier	identificatore Scopus
Surname	cognome dell'autore
Name	nome dell'autore
subject areas	insieme delle prime tre tematiche dei paper sulla base del numero di pubblicazioni fatte in ogni area
affiliation_id	identificatore dell'ateneo di appartenenza

Tabella 2.4: Descrizione dati Scopus

Il risultato della procedura fornirà un *dc_identifier* assegnato o un risultato "ambiguo" nel caso in cui non ci sia univocità degli identificatori.

2.2 Pubblicazioni e references

Vengono di seguito analizzati i dati necessari per la corretta corrispondenza delle informazioni relative alle pubblicazioni e alle *references* e l'identificativo univoco di ogni docente.

2.2.1 Descrizione dei dati

Per ricavare il codice identificativo di ogni *paper* sono state effettuate delle ricerche su Scopus che, oltre alle informazioni riportate in Tabella 2.4, hanno prodotto anche quelle relative alle pubblicazioni e alle *references*.

Per quelle che sono le finalità, si è scelto di prendere in considerazione solo i campi riportati nella Tabella 2.5, a cui è stato inserito un ulteriore

campo, *is_references*, per distinguere i *paper*s degli autori dalle loro rispettive *references*.

Dati	Descrizione
id documento	codice identificativo di Scopus del *paper*
Titolo	titolo della pubblicazione, se presente
Data	data di pubblicazione del *paper*
pubblicationName	nome della rivista in cui è stato pubblicato il *paper*
scourceType	tipologia di rivista (e.g. journal, conference)
sourceSubType	sotto tipologia della pubblicazione (e.g. articolo, conference *paper*)
n_citation	numero di citazioni complessive
is_reference	0 = non è una *reference*, 1 = è una *reference*

Tabella 2.5: Descrizione dati dei paper Scopus

2.2.2 Informazioni sulle references

Una volta ottenuto l'identificatore univoco per ogni *paper*, si è utilizzato quest'ultimo per ricavare le pubblicazioni inserite nelle *references*.

Verrà così a crearsi una relazione tra un *paper* e la sua bibliografia attraverso l'identificatore del documento associato all'autore, o a più autori, e l'identificatore del documento presente nella *reference*, ottenendo così per ogni *paper* tanti risultati quante sono le pubblicazioni inserite all'interno della bibliografia.

Capitolo 3
Struttura del database e casi di ambiguità

In questo capitolo verranno descritte le tabelle presenti nel database così da capire quali siano le relazioni che ci permettono di ottenere, salvo casi specifici che verranno analizzati, le corrispondenze richieste tra autori, pubblicazioni e citazioni.

3.1 Descrizione delle tabelle

Di seguito l'elenco delle attuali 14 tabelle presenti nel database.

1. *affiliation_scopus*: contiene le informazioni sulle affiliazioni degli autori presenti nella tabella *dati_scopus_per_score*.
2. *ambigui_dc_id_duplicato*: tabella contenente tutti gli autori che negli anni hanno cambiato SSD o che sono stati riportati con nomi o cognomi diversi negli anni, e di conseguenza non sono stati disambiguati;
3. *ambigui_duplicati*: contiene gli autori aventi stesso nome, cognome e SSD nello stesso anno, motivo per cui la procedura non è andata a buon fine.
4. *ambigui_non_disambiguati*: presenta tutti gli autori che non hanno ottenuto un identificatore univoco dalla procedura di disambiguazione.
5. *authorship_1*: relazione tra i *papers* e i rispettivi autori, per ciascun *paper* saranno presenti tante coppie (id_autore, id_documento) quanti sono i suoi autori.
6. *aut_scopus_1*: contiene tutti gli autori dei *papers*, anche quelli appartenenti agli SSD non analizzati.

7. *citations_1*: numero di citazioni di un *paper* per ogni anno a partire dalla sua pubblicazione.
8. *dati_miur*: contenente i dati degli autori, ovvero i docenti, presi dal file MIUR.
9. *dati_scopus_per_score*: risultati Scopus a seguito della ricerca dell'autore attraverso i dati presenti nel file MIUR.
10. *definitivi_finali_completi*: tabella in cui sono presenti tutte le informazioni, anno per anno, di tutti gli individui definitivi, ovvero quelli a cui è stato assegnato correttamente un identificatore univoco.
11. *definitivi_finali_unici*: individui definitivi rappresentati attraverso altri campi.
12. *non_esistenti_finali*: raccoglie tutti i docenti ai quali non può essere assegnato un identificativo Scopus dalla procedura di disambiguazione.
13. *papers_1*: caratteristiche di interesse dei *papers*.
14. *references_1*: relazione esistente tra un *paper* e le sue *references*.

3.2 Casi di ambiguità

In questo paragrafo verranno analizzati tutti i casi di ambiguità presenti nelle tabelle *ambigui_dc_id_duplicato, ambigui_duplicati* e *ambigui_non_disambiguati*, tabelle che meglio esprimono le criticità dell'algoritmo.

Le tipologie di ambiguità riscontrate sono 6:

- CASO A: autori aventi più di un SSD, questo è il caso in cui l'autore ha un suo identificativo univoco Scopus (*dc_identifier*) che nel database viene associato al nome e al cognome dell'autore tante volte quanti sono i settori scientifico disciplinari

cambiati da lui negli anni. Inoltre, alcuni autori sono presenti sia nella tabella *definitivi_finali_unici* con un SSD sia nella tabella *ambigui_dc_id_identifier* o *ambigui_non_duplicati* con l'altro.

- CASO B: autori che vengono riportati in anni diversi con diverso nome o cognome e vengono quindi considerati come persone diverse.

- CASO C: autori che sono passati da una fascia ad un'altra nello stesso anno, e vengono riportati due volte nella tabella *dati_miur* con entrambe le fasce.

- CASO D: autori non presenti nella tabella *dati_scopus_per_score* nonostante siano stati correttamente inseriti nella tabella *dati_miur*.

- CASO E: omonimi, autori presenti nella tabella *ambigui_non_disambiguati* con lo stesso nome, cognome e SSD, ma diverso *dc_identifier* perché di fatto sono persone diverse che insegnano in atenei diversi negli stessi anni.

- CASO F: autori presenti nella tabella *ambigui_non_disambiguati* con lo stesso nome, cognome e SSD, ma diverso *dc_identifier* perché gli autori hanno più profili Scopus e di conseguenza hanno diversi codici identificati; questo spesso accade perché gli autori non hanno provveduto a riunificare i profili presenti su Scopus, profili che spesso presentano affiliazioni diverse in base al periodo di insegnamento dell'autore.

Per le analisi che seguiranno, il caso E ed il caso F saranno riuniti in un unico caso denominato CASO GENERICO in cui sono presenti tutti gli autori che vengono riportati con stesso nome e cognome, indipendentemente dai casi di omonimia o di stessi autori aventi più profili Scopus.
Verranno di seguito riportati degli esempi relativi ad alcune tipologie.

3.2.1 Caso A

Sono stati confrontati i valori degli autori presenti nella tabella *dati_miur* con quelli presenti nella tabella *dati_scopus_per_score*; è evidente come, nella maggior parte dei casi, si tratti della stessa persona che negli anni ha cambiato SSD, e di conseguenza, si potrebbero riunire tutti gli individui che presentano il problema tenendo conto di questa caratteristica.
La Tabella 3.1 riporta un esempio riguardante l'autore Oronzo Altamura che ha ottenuto lo stesso codice identificativo Scopus nonostante avesse due SSD diversi (ING-INF/01 e ING-INF/05); nella Tabella 3.2 e nella Tabella 3.3 i dati necessari dell'autore per comprendere meglio il caso in esame.

nome	cognome	dc_identifier	ssd
Oronzo	Altamura	7801576645	ING-INF/01
Oronzo	Altamura	7801576645	ING-INF/05

Tabella 3.1: Tabella *ambigui_dc_id_duplicati*

id	nome	cognome	ssd	anno
10680	Oronzo	Altamura	ING-INF/01	2001
1811	Oronzo	Altamura	ING-INF/05	2002
16111	Oronzo	Altamura	ING-INF/05	2003
35597	Oronzo	Altamura	ING-INF/05	2004
29352	Oronzo	Altamura	ING-INF/05	2005
34043	Oronzo	Altamura	ING-INF/05	2006
5905	Oronzo	Altamura	ING-INF/05	2007
18571	Oronzo	Altamura	ING-INF/05	2008
9536	Oronzo	Altamura	ING-INF/05	2009
40965	Oronzo	Altamura	ING-INF/05	2010
31305	Oronzo	Altamura	ING-INF/05	2011
26231	Oronzo	Altamura	ING-INF/05	2012

Tabella 3.2: Tabella *dati_miur*

dc_identifier	surname	name	id_originale	ssd	anno
7801576645	Altamura	Oronzo	10680	ING-INF/01	2001
7801576645	Altamura	Oronzo	1811	ING-INF/05	2002
7801576645	Altamura	Oronzo	16111	ING-INF/05	2003
7801576645	Altamura	Oronzo	35597	ING-INF/05	2004
7801576645	Altamura	Oronzo	29352	ING-INF/05	2005
7801576645	Altamura	Oronzo	34043	ING-INF/05	2006
7801576645	Altamura	Oronzo	5905	ING-INF/05	2007
7801576645	Altamura	Oronzo	18571	ING-INF/05	2008
7801576645	Altamura	Oronzo	9536	ING-INF/05	2009
7801576645	Altamura	Oronzo	40965	ING-INF/05	2010
7801576645	Altamura	Oronzo	31305	ING-INF/05	2011
7801576645	Altamura	Oronzo	26231	ING-INF/05	2012

Tabella 3.3: Tabella *dati_scopus_per_score*

Capiamo quindi come in realtà Oronzo Altamura, che compare una sola volta nella tabella *dati_miur*, ma due volte in quella degli *ambigui_dc_id_duplicato*, sia la stessa persona che ha cambiato negli

anni il settore scientifico disciplinare, motivo per cui l'algoritmo segnala un'ambiguità.

In considerazione del fatto che sul sito Scopus l'autore è presente una sola volta, con quello specifico *dc_identifier*, si può assegnarglielo e quindi procedere con la procedura di disambiguazione.

Nella Tabella 3.4 è presente il risultato della procedura.

nome	cognome	dc_identifier	ssd	
Oronzo	Altamura	7801576645	ING-INF/01	ING-INF/05

Tabella 3.4: Risultato della procedura

3.2.2 Caso B

Si può inoltre analizzare il caso in cui la procedura di disambiguazione non è andata a buon fine, non a causa della presenza di più SSD per lo stesso autore, ma per la presenza del nome o del cognome dell'autore scritto in diverso modo nei vari anni nella tabella *dati_miur*.

Nella Tabella 3.5 viene riportato il caso di Federica Vigano' che ha un solo SSD, ma le è stato assegnato lo stesso *dc_identifier* nonostante avesse il cognome diverso; capiamo però, grazie all'analisi di alcuni dei campi delle tabelle del database riportate nella Tabella 3.6 e nella Tabella 3.7, che si tratta della stessa persona.

nome	cognome	dc_identifier	ssd
Federica	Vigano'	56506986900	SECS-P/06
Federica	Vigano	56506986900	SECS-P/06

Tabella 3.5: Tabella *ambigui_dc_id_duplicati*

id	nome	cognome	ssd	anno
52043	Federica	Vigano'	SECS-P/06	2011
52200	Federica	Vigano'	SECS-P/06	2012
52350	Federica	Vigano'	SECS-P/06	2013
52490	Federica	Vigano'	SECS-P/06	2014
52641	Federica	Vigano'	SECS-P/06	2015
52800	Federica	Vigano'	SECS-P/06	2016
52964	Federica	Vigano'	SECS-P/06	2017
53134	Federica	Vigano	SECS-P/06	2018
53310	Federica	Vigano'	SECS-P/06	2019

Tabella 3.6: Tabella *dati_miur*

dc_identifier	surname	name	id_originale	ssd	anno
56506986900	Viganò	Federica	52043	SECS-P/06	2011
56506986900	Viganò	Federica	52200	SECS-P/06	2012
56506986900	Viganò	Federica	52350	SECS-P/06	2013
56506986900	Viganò	Federica	52490	SECS-P/06	2014
56506986900	Viganò	Federica	52641	SECS-P/06	2015
56506986900	Viganò	Federica	52800	SECS-P/06	2016
56506986900	Viganò	Federica	52964	SECS-P/06	2017
56506986900	Viganò	Federica	53134	SECS-P/06	2018
56506986900	Viganò	Federica	53310	SECS-P/06	2019

Tabella 3.7: Tabella *dati_scopus_per_score*

Nella Tabella 3.8 è presente il risultato della procedura.

nome	cognome	dc_identifier	ssd
Federica	Vigano'	56506986900	SECS-P/06

Tabella 3.8: Risultato della procedura

3.2.3 Caso F

Diverso è il caso degli autori aventi stesso nome, cognome e SSD, ma diverso *dc_identifier*.

Grazie all'*affiliation_id*, e all'ateneo riportato nella tabella *dati_miur*, è stato possibile disambiguare la maggior parte degli autori, associando loro in maniera univoca il rispettivo *dc_identifier*.

Inoltre, nel caso in cui risultasse esserci lo stesso *affiliation_id*, si è tenuto conto di ulteriori dati:

- l'*institution history* di Scopus che riporta tutti gli atenei in cui l'autore ha insegnato;
- il numero di pubblicazioni presente su Scopus, grazie al quale si evince che lo stesso autore ha a lui associato più profili Scopus, che andrebbero riunificati dalla specifica persona;
- le *subject_area*, fondamentali per verificare la compatibilità degli insegnamenti con l'SSD specifico.

Sulla base di queste informazioni è stato possibile associare il *dc_identifier* corretto, e di conseguenza disambiguare gli autori.

Un esempio è quello dell'autore Alessandro Micarelli che presenta due *dc_identifier*, entrambi relativi ad autori che hanno come ateneo associato quello di "Roma Tre". Per capire quindi se si trattasse di due persone diverse, sono state analizzate le *subject_area* e il numero di pubblicazioni, arrivando alla conclusione che la persona fosse la stessa.

La Tabella 3.9, 3.10 e 3.11 dettagliano la situazione dell'autore Alessandro Micarelli.

nome	cognome	dc_identifier	ssd
Alessandro	Micarelli	57210939886	ING-INF/05
Alessandro	Micarelli	52564169500	ING-INF/05

Tabella 3.9: Tabella *ambigui_non_disambiguati*

id	nome	cognome	ateneo	ssd	anno
10604	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2001
2227	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2002
16526	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2003
36031	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2004
29808	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2005
34525	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2006
6411	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2007
19106	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2008
10071	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2009
41500	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2010
31845	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2011
26792	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2012
1512	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2013
37071	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2014
14461	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2015
23753	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2016
4335	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2017
5611	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2018
42785	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2019
11946	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2020
40065	Alessandro	Micarelli	ROMA TRE	ING-INF/05	2021

Tabella 3.10: Tabella *dati_miur*

dc_identifier	surname	name	ateneo	subject_area	pubblicazioni
52564169500	Micarelli	Alessandro	ROMA TRE	COMP:15, MATH:5, ENGI:1	21
57210939886	Micarelli	Alessandro	ROMA TRE	COMP:95, MATH:32, ENGI:11	138

Tabella 3.11: Tabella *dati_scopus_per_score*

Nella Tabella 3.12 è presente il risultato della procedura.

nome	cognome	dc_identifier	ssd
Alessandro	Micarelli	57210939886	ING-INF/05

Tabella 3.12: Risultato della procedura

3.2 Quantificazione dei dati e possibili soluzioni

Verranno adesso riportate le percentuali delle casistiche analizzate, nonché le possibili soluzioni da adottare per individuarle e risolverle.

Si ricorda che il totale degli autori che sono stati sottoposti alla procedura di disambiguazione sono 4841, e le tipologie di ambiguità sono state riscontrate solo nelle tabelle *ambigui_dc_id_duplicato, ambigui_duplicati* e *ambigui_non_disambiguati* (totale righe tabelle 1453).

Riassumendo, le tipologie di ambiguità riscontrate sono 6 e vengono graficate in Figura 3.1:

- CASO A: 62 autori su 4841 autori;
- CASO B: 3 autori su 4841 autori;
- CASO C: 4 autori su 4841 autori;
- CASO D: 3 autori su 4841 autori;
- CASO GENERICO (E+F): 392 autori su 4841 autori.

Per un totale di 464 casi.

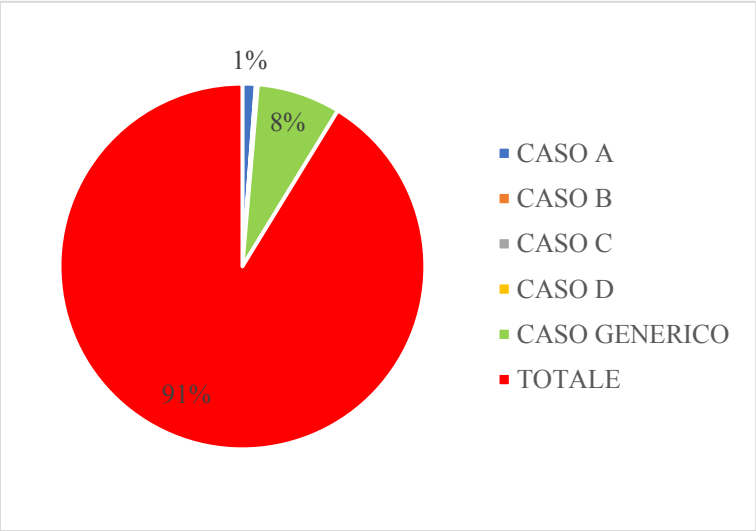

Figura 3.1: Casi di ambiguità totali

Inoltre, analizzando più nello specifico le tabelle del database in cui sono presenti i casi sopra elencati, si nota che:

- nella tabella *ambigui_dc_id_duplicato* sono presenti CASO A (48/62), CASO B (3/3) e CASO D (2/3);

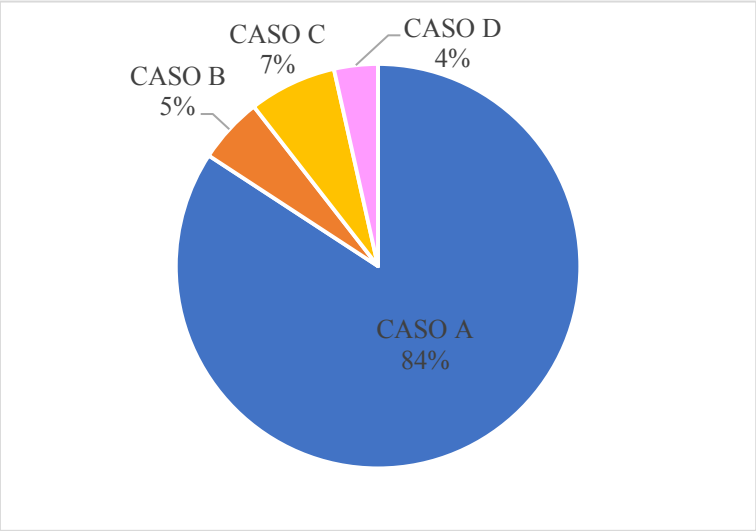

Figura 3.2: Casi di ambiguità tabella *ambigui_dc_id_duplicato*

- nella tabella *ambigui_duplicati* sono presenti CASO C (4/4), CASO D (1/3), CASO GENERICO (3/392).

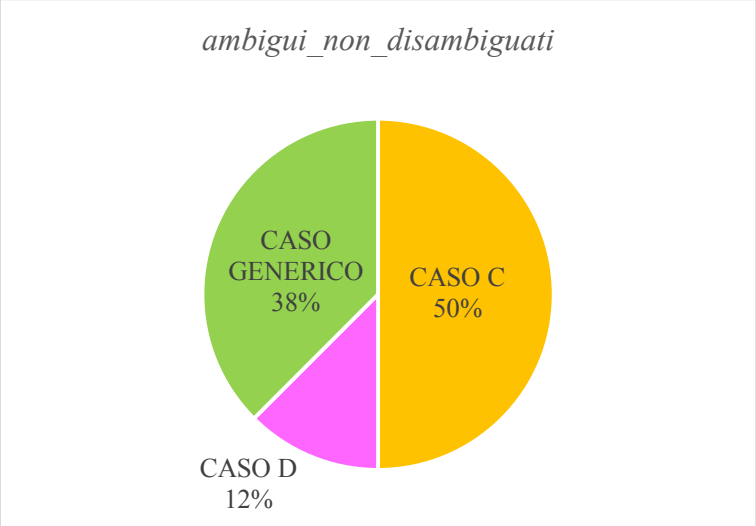

Figura 3.3: Casi di ambiguità tabella *ambigui_duplicati*

- nella tabella *ambigui_non_disambiguati* sono presenti CASO A (14/62), CASO GENERICO (389/392).

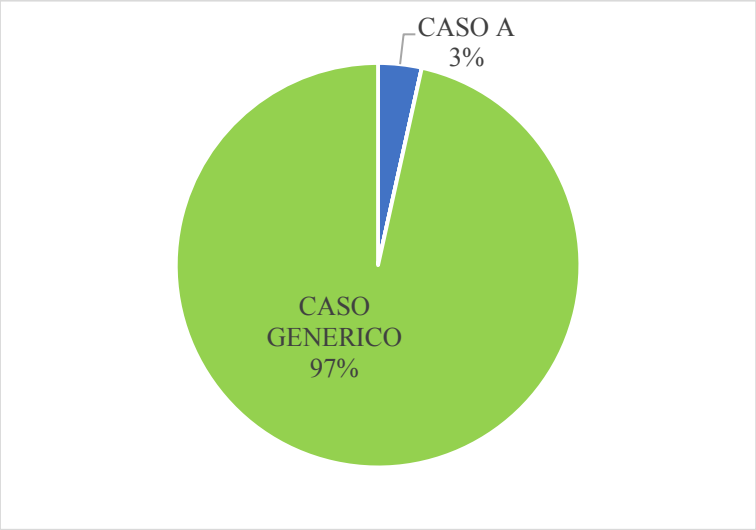

Figura 3.3: Casi di ambiguità tabella *ambigui_non_disambiguati*

Per individuare le diverse tipologie sono stati analizzati specifici parametri che verranno di seguito espressi tramite query SQL:

- CASO A:

```
SELECT nome, cognome, dc_identifier, SUM(num_settori)
FROM (SELECT nome, cognome, dc_identifier, 1 AS num_settori
    FROM `definitivi_finali_unici`
    UNION ALL
    SELECT nome, cognome, dc_identifier, 1 AS num_settori
    FROM ambigui_dc_id_duplicato) t1
GROUP BY nome, cognome, dc_identifier
HAVING SUM(num_settori)>1;
```

1. Per verificare quali autori risultano essere nella tabella *ambigui_dc_id_duplicato* e anche nella tabella *definitivi_finali_unici*:

```
SELECT *
FROM definitivi_finali_unici AS a, ambigui_dc_id_duplicato AS b
WHERE a.nome=b.nome AND a.cognome=b.cognome AND
a.dc_identifier=b.dc_identifier;
```

2. Per verificare quali autori risultano essere nella tabella *ambigui_non_disambiguati* e anche nella tabella *definitivi_finali_unici*:

```
SELECT a.nome, a.cognome, a.dc_identifier, a.ssd, b.nome, b.cognome, b.dc_identifier, b.ssd
FROM ambigui_non_disambiguati as a, definitivi_finali_unici as b
WHERE a.nome=b.nome AND a.cognome=b.cognome
GROUP BY a.nome, a.cognome;
```

3. E per verificare quali autori risultano essere nella tabella *ambigui_dc_id_duplicato* ma non nella tabella *definitivi_finali_unici*:

```
SELECT nome, cognome, dc_identifier, COUNT(dc_identifier) AS num_settori
FROM ambigui_dc_id_duplicato
GROUP BY nome, cognome, dc_identifier
HAVING COUNT(dc_identifier)>1;
```

- CASO B:

```
SELECT a.nome, a.cognome, a.dc_identifier, a.ssd, GROUP_CONCAT(a.nome, a.cognome)
FROM ambigui_dc_id_duplicato AS a, ambigui_dc_id_duplicato AS b
WHERE a.dc_identifier=b.dc_identifier AND (a.nome<>b.nome || a.cognome<>b.cognome)
GROUP BY a.dc_identifier;
```

- CASO C:

```
SELECT nome, cognome, ssd, anno, GROUP CONCAT(fascia)
FROM `dati_miur`
GROUP BY nome, cognome, anno, ssd, ateneo
HAVING COUNT(*)>1;
```

- CASO D: individuato manualmente.

- CASO GENERICO (E+F): si considera il totale delle persone con stesso nome e cognome nella tabella *ambigui_non_disambiguati*.

  ```
  SELECT a.nome, a.cognome
  FROM ambigui_non_disambiguati AS a, ambigui_non_disambiguati AS b
  WHERE a.nome=b.nome AND a.cognome=b.cognome
  GROUP BY a.nome, a.cognome;
  ```

Infine, con lo scopo di automatizzare la procedura, per ogni possibile caso, si è pensato di procedere nei seguenti modi:

- CASO A:

  Partendo dalla tabella *dati_miur*, può essere fatta una verifica sugli ssd presenti negli anni, e nel caso in cui risultino essercene diversi, contare quanti sono e registrarne il valore in un campo denominato "numero_settori" che potrebbe essere aggiunto nella tabella *definitivi_finali_unici*.

  Per avere a disposizione anche il tipo di settore cambiato, si potrebbe pensare di lasciare la tabella *definitivi_finali_unici* con i soli campi "nome", "cognome", "dc_identifier" e "numero_settori", in maniera tale da avere il totale degli autori con solo l'indicazione del numero di settori.

  Creare una tabella *definitivi_unico_settore* con i campi "nome", "cognome", "dc_identifier" e "ssd", il cui valore del settore verrà estratto dalla tabella *dati_miur*, e creare una tabella *definitivi_piu_settori* con i campi "nome", "cognome", "dc_identifier", "ssd1", "ssd2" e "ssd3", i cui valori saranno sempre ripresi dalla tabella *dati_miur*.

Essendo che il numero di autori con più ssd è molto basso, e tenendo conto del fatto che al massimo, dalle analisi fatte, i settori cambiati sono tre, non risulterà essere un problema aggiungere un campo "ssd3" che verrà riempito solo per un numero limitato di persone, generando quindi poi campi nulli.

- CASO B:

  Partendo dalla tabella dati_miur, può essere confrontata la coppia nome-cognome di un autore, tra tutti gli anni presenti per l'autore, la coppia che avrà una percentuale maggiore di presenza, sarà quella finale, e quindi anche i valori "nome" e "cognome" che ne derivano.

- CASO C:

  Si potrebbe creare una tabella *fasce* in cui ogni fascia ha un proprio peso. Di conseguenza si potrebbero poi verificare nella tabella *dati_miur* i valori, per ogni anno, del campo "fascia" e controllare che ce ne sia solo uno, e se per lo stesso anno il peso della fasciaX risulta essere inferiore del peso della fasciaY, tenere come valore solo quello con peso maggiore.

- CASO D:

  Essendo in numero molto limitato rispetto al totale il problema risiede nel momento dell'inserimento dati, e non nell'algoritmo.

- CASO GENERICO:

  Innanzitutto, bisogna verificare se nella tabella *dati_miur*, nello stesso anno, risultano più persone con lo stesso nome, cognome

e ssd, ma diverso ateneo di appartenenza; in questo modo verranno identificati gli eventuali casi di omonimia tra gli autori presenti nel dataset del MIUR.

Dal momento che gli omonimi possono presentarsi anche a livello internazionale, e di conseguenza potrebbe essere un ostacolo alla corretta disambiguazione dell'autore, bisogna confrontare i campi "nome" e "cognome" della tabella *dati_miur* con quelli presenti su Scopus.

Ottenuti così tutti i probabili omonimi, si deve porre attenzione sull'ateneo di appartenenza: per prima cosa bisogna escludere tutti quegli autori derivanti da Scopus che hanno il campo "affiliation_id" nullo poiché non forniscono alcuna informazione, successivamente, una volta associato il campo "ateneo" della tabella *dati_miur* con il rispettivo campo "id" della tabella *affiliation_id*, si può procedere con il confronto degli atenei tra la tabella *dati_miur* e quella *dati_scopus_per_score*,

Le possibilità diventano due:

1. gli atenei sono diversi, si conferma così il caso di omonimia;
2. gli atenei sono uguali, in questo caso per verificare se è un caso di omonimia oppure è il caso dello stesso autore con più profili Scopus, dovranno essere confrontate le *subject_area*, per verificare la loro corrispondenza con il settore dell'autore, e il numero di pubblicazioni: se le *subject_area* non corrispondono si tratta di omonimi, in caso contrario si considera corretto l'identificativo dell'autore associato a quello che presenta maggiori pubblicazioni.

Da considerare il caso in cui l'autore con più profili sia associato a diversi atenei, in quel caso sarà rilevante l'analisi dell'*institution history* di Scopus che riporta tutti gli atenei in cui l'autore ha insegnato.

Capitolo 4

Analisi qualitativa

Per verificare la qualità del database, considerando un campione di venti persone, sono stati confrontanti i dati relativi alle pubblicazioni e alle *references* derivanti dalla procedura con quelli presenti su Scopus.
Per effettuare i confronti, sono stati esportati dapprima i valori di tutte le pubblicazioni associate ad un preciso autore dal database, ottenuti attraverso una query SQL, e successivamente esportando quelli di interesse da Scopus.
Attraverso un foglio *excel* si è proceduto per il confronto.

4.1 Pubblicazioni

Nello specifico delle pubblicazioni, i dati confrontati sono stati:

- titolo;
- pubblicationName;
- sourceSubType;
- ISSN;
- numero degli autori.

La query SQL utilizzata per il confronto riguardante le pubblicazioni coinvolge le tabelle *definitivi_finali_unici*, *authorship_1* e *papers_1*, che hanno permesso di associare ad ogni autore le rispettive pubblicazioni.
Dalle verifiche effettuate è emerso che, per la maggioranza degli autori presi in esame, risulta esserci una corrispondenza quasi esatta tra le pubblicazioni presenti su Scopus e quelle del database.

A fare la differenza sono due casi ricorrenti:

1. i titoli che spesso risultano avere degli errori, riguardanti principalmente la presenza di caratteri speciali, lettere greche, pedici (nel caso di formule chimiche per esempio) e apici (nel caso di numeri ordinali). Oppure, la traduzione in lingua italiana, presente nel titolo originale, che a volte non viene riportata nel database;
2. la mancanza delle pubblicazioni effettuate nell'ultimo anno e mezzo.

Casi eccezionali sono quelli in cui ad un valore presente su Scopus, che sia, per esempio, un ISSN o un sourceSubType, corrisponda un valore nullo nel database.

La Tabella 4.1 riporta esempi di alcune delle pubblicazioni aventi il titolo errato.

Autore	Titolo database	Titolo Scopus
Caterina Ciminelli	"Effects of thermal annealing on the optical characteristics of *K⁺-Na⁺* waveguides"	"Effects of thermal annealing on the optical characteristics of *K+-Na+* waveguides"
Lorenzo Capineri	A large-area PVDF pyroelectric sensor for *CO<inf>2</inf>* laser beam alignment	A large-area PVDF pyroelectric sensor for *CO2* laser beam alignment
Maria Prudenziati	Switching effect in ?-rhombohedral boron	Switching effect in *β*-rhombohedral boron
Lauretta Rubini	Government support and R*&*D investment effectiveness in Chinese SMEs: A complex relationship	Government support and R&D investment effectiveness in Chinese SMEs: A complex relationship

Tabella 4.1: Esempi pubblicazioni con titolo errato

4.2 References

La stessa verifica è stata effettuata per le *references*, tenendo conto del loro totale per pubblicazione.

Il risultato è stato che, anche in questo caso, manca l'aggiornamento delle ultime citazioni. Inoltre, in alcuni casi, è stato riscontrato un disallineamento tra quelle che sono le citazioni presenti negli anni nel database e quelle presenti su Scopus.

La query SQL utilizzata per il confronto coinvolge la tabella *citations_1*, grazie alla quale si sono ottenute non solo il totale delle citazioni per documento, ma anche la verifica del fatto che le citazioni mancanti facessero riferimento all'ultimo anno e mezzo.

Nella Tabella 4.2 è riportato un esempio riguardante le citazioni del documento *"Building Bridges in Global Virtual Teams: The Role of Multicultural Brokers in Overcoming the Negative Effects of Identity Threats on Knowledge Sharing Across Subgroups*" dell'autrice Elisa Mattarelli; si nota come le citazioni mancanti siano quelle più recenti.

	SCOPUS	**DATABASE**
<2019	3	3
2019	4	4
2020	8	8
2021	10	11
2022	13	14
2023	4	0
>2023	0	0
Totale	42	40

Tabella 4.2: Esempi di citazioni non corrispondenti

Capitolo 5
Conclusioni

Molti sono stati gli aspetti analizzati del database bibliografico in esame.

Innanzitutto, è stato fondamentale partire dalle ragioni per cui si è ritenuta necessaria la sua costruzione, entrando poi nel vivo di quella che è la procedura di disambiguazione e dei parametri presi in considerazione per la sua realizzazione.

In merito alla struttura della base di dati, sì può sottolineare come le tabelle siano state organizzate in maniera tale da suddividere le diverse tipologie di informazioni, facilitando così l'individuazione dei casi di ambiguità derivanti dalla procedura, e ben raggruppati in tre tabelle diverse.

Dallo studio approfondito delle ambiguità riscontrate, è emerso che risultano esserci 6 casistiche che in numero corrispondono ad un totale di 464 autori su 4841 totali.

I casi sono stati analizzati separatamente, e per ognuno di loro è stato descritto il modo in cui può essere individuato, nonché una possibile soluzione da utilizzare per rendere più efficiente ed efficace la procedura di disambiguazione da applicare ad eventuali nuovi autori.

Nel complesso, essendo che gli autori che presentano delle particolari situazioni corrispondo a circa il 10% del totale, è evidente come essi rappresentino una minoranza. Tuttavia, si ritiene opportuno risolvere quelle che sono le problematiche più frequenti e, dalle analisi effettuate, si evince come ci siano tutti i parametri necessari per raffinare la procedura di disambiguazione.

Per ultimo, attraverso un'analisi approfondita dei dati, in particolare delle pubblicazioni e delle citazioni, si può notare che vi è una coerenza tra quelli presenti nel database e quelli presenti su Scopus.
Si può concludere quindi che il database analizzato risulta essere una valida fonte di informazioni, e la procedura di disambiguazione relativa al nome degli autori potrà essere perfezionata andando ad automatizzare ciò che manualmente ha permesso di individuare e risolvere i casi di ambiguità.

Bibliografia

[1] LEGGE 30 dicembre 2010, n. 240 (in G.U. n. 10 del 14 gennaio 2011- Suppl. Ord. n.11 - in vigore dal 29 gennaio 2011) - Norme in materia di organizzazione delle università, di personale accademico e reclutamento, nonché delega al Governo per incentivare la qualità e l'efficienza del sistema universitario.

[2] DM 08/08/2018 n.589 Definizione valori - soglia degli indicatori di impatto della produzione scientifica.