UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Scienze Fisiche, Informatiche e Matematiche

CORSO DI LAUREA IN INFORMATICA

**Titolo Provvisorio**

Laureando:

Paolo Crotti

Relatore:

Dott. Mauro Andreolini

Anno Accademico 2018/2019

Sommario

[1 Introduzione 2](#_Toc2529662)

[2 Strumenti e risorse del progetto 3](#_Toc2529663)

[2.1 Google Scholar 3](#_Toc2529668)

[2.2 Microsoft Academic 3](#_Toc2529669)

[1 Bibliografia 4](#_Toc2529670)

# 1 Introduzione

Questo lavoro di tesi si concentra su un aspetto legato alla scrittura degli articoli scientifici: la ricerca dei “related works”, cioè di tutte quelle pubblicazioni correlate a ciò su cui si sta lavorando. Lo scopo è quello di realizzare un programma che aiuti l’autore a informarsi su un dato argomento, permettendogli di individuare in modo automatico gli articoli, i topic e gli autori correlati.

La scrittura di un articolo ha delle difficoltà: è necessario definire l’obbiettivo e le tappe per raggiungerlo, suddividere il corpo in paragrafi chiari e intuitivi, fornire una moltitudine di esempi e similitudini così da favorire la comprensione da parte del lettore, ma soprattutto richiede una visione completa dell’argomento di cui si sta parlando [1]. La ricerca scientifica viene svolta su un argomento ben preciso attorno al quale si possono individuare delle problematiche ben chiare, inoltre la ricerca deve poter aggiungere novità allo stato dell’arte oppure rivedere in una diversa ottica argomenti già affrontati. Il suo scopo è quello di essere utile agli altri aggiungendo qualcosa di nuovo e tutti gli altri lavori fatti sullo stesso argomento dovranno tenerne conto. La ricerca deve inoltre fornire elementi che permettano di metterla in discussione (verifica delle conclusioni raggiunte) e consentire ad altri di continuarne il lavoro.

Per raggiungere tutti questi obbiettivi è necessario scrivere articoli di qualità ed è per cui fondamentale informarsi approfonditamente sull’argomento, prima di iniziare il lavoro. La prima legge della comunicazione di Wittington dice che “quando qualcuno spiega un argomento che non ha ben capito, sarà compreso solo da chi ne sa più di lui” [2]. Quindi se si scrive un articolo senza aver compreso appieno l’argomento di cui si sta parlando, è probabile che il lettore non riuscirà a capirlo.

Quindi avere una profonda conoscenza dell’argomento aiuta a:

1. Non commettere errori che comprometterebbero l’immagine dell’autore e dell’editore.
2. Scegliere gli aspetti più interessanti da trattare, individuati grazie all’esperienza personale maturata col tempo.
3. In generale a scrivere meglio: usare gli esempi e le parole più efficaci e creare un filo logico più chiaro.

Il problema principale nella scrittura degli articoli è la scrittura dello stato dell’arte definito come “il punto cui sono arrivate le ricerche in una determinata disciplina” [3] perché per avere una visione completa dell’argomento sarebbe necessario che l’autore fosse a conoscenza di qualsiasi pubblicazione mai uscita al riguardo. La soluzione proposta consiste nel cercare sui principali motori di ricerca per le pubblicazioni, quindi Google Scholar e Microsoft Academic gli articoli correlati e costruire un grafo di ricerca. In questo grafo, i nodi sono gli articoli mentre gli archi rappresentano le citazioni. L’idea è quella di partire da una o più pubblicazioni, raccoglierne gli articoli correlati e costruire dei grafi da cui estrarre informazioni come gli articoli e gli autori più importanti o i topic più diffusi.

SOLUZIONE PROPOSTA

Difficile scoprire nuovi topic dai vecchi?

Per capire come sono correlati tra loro i topic all’interno del related work creo un secondo grafo in cui mostro i topic collegati tra loro in modo da evidenziare non solo quelli che compaiono più frequentemente (per quello c’è una classifica) ma la gerarchia che esiste tra loro, per esempio il topic “rootkit” compare solo in articoli che ne citano altri di “computer security”, questo mi dice che all’interno di questo ambito di ricerca “rootkit” è (secondario) rispetto a “computer security”.

NOTA: questo tipo di grafo lo posso costruire solo quando la ricerca viene svolta sul search engine “Microsoft Academic” perché qui sono presenti i topic (più avanti dovrò spiegare come sono strutturati gli articoli in MA e GS, magari con della documentazione) che sono invece assenti su Google Scholar.

Per costruirlo è necessario navigare tra le pagine web di Google Scholar ed estrarre le poche informazioni sugli articoli

METRICHE UTILIZZATE

Quanti articoli per firstN? OBB: anche se ne aumento il numero, non aggiungo articoli rilevanti che quindi appartengono al related work.

Quanti thread? COME: ripeto più volte la stessa ricerca con numero di thread diverso e trovo la massima efficienza (buttaci anche il grafo) (argmax(E)), caso mai mettici anche la tabella dello SpeedUp.

# 2 Strumenti e risorse del progetto



## Google Scholar

Google Scholarè un motore di ricerca accessibile liberamente che tramite parole chiave specifiche consente di individuare vari tipi di pubblicazioni come articoli scientifici, tesi di laurea e libri. Consente inoltre di reperire articoli da una vasta gamma di case editrici che si rivolgono al mondo dello studio e della ricerca da associazioni scientifiche e professionali e università, oltre che nella galassia di articoli scientifici e culturali distribuiti nel Web.

Uno dei suoi pregi è quello di fornire, tramite una ricerca Web, l’accesso gratuito a una versione full text dell’articolo senza il permesso del giornale che lo ha pubblicato. Questo perché cerca copie degli articoli indicizzati anche su siti considerati non sicuri. Scholar non fornisce una lista dei periodici da cui prende gli articoli perché si limita a fare una lista di tutte le pubblicazioni dalla sua ricerca per il Web. Oltre alla ricerca classica in cui cerca all’interno del proprio indice le parole inserite dall’utente, permette anche una Ricerca Avanzata. Qui è possibile specificare le parole che devono comparire nel titolo e/o nel corpo, gli autori, gli editori e la data di pubblicazione. [4] Tra le informazioni dei risultati compare il numero degli articoli che hanno citato il risultato, questi articoli possono essere recuperati dal link pertinente. Altro link rilevante è quello che porta agli articoli correlati cioè quelli che sono

Come Scholar ordina i risultati? Il numero di citazioni è il parametro tenuto maggiormente in considerazione nell’algoritmo di ranking di Scholar; infatti articoli citati spesso compaiono con molta più frequenza tra i risultati delle ricerche. Di conseguenza Scholar sembra essere più indicato per la ricerca di articoli che seguono la corrente principale piuttosto che quelli che propongono una visione alternativa. [5]

## Microsoft Academic

Capitolo 3

Funzionalità e implementazione del prototipo

Sviluppi futuri

Incrociare i risultati dei documenti rilevanti tra Google Scholar e Microsoft Academic, confrontando i titoli degli articoli in modo intelligente (perché a volte i titoli differiscono tra i due motori di ricerca). Una soluzione potrebbe essere confrontare anche gli autori (che però non sono completamente presenti su Scholar) e la data (che non riesco a ricavare da Academic). Come ulteriore problema le date sono spesso diverse tra i due e lo stesso articolo è presente più volte con date diverse.

# Bibliografia

|  |  |
| --- | --- |
| [1] | «Mestiere di Scrivere,» [Online]. Available: http://www.mestierediscrivere.com/articolo/articolotecnico.html. |
| [2] | U. d. V. D. d. C. e. Civiltà. [Online]. Available: http://www.dcuci.univr.it/documenti/OccorrenzaIns/matdid/matdid007359.pdf. |
| [3] | «Wikipedia: Stato dell'Arte,» [Online]. Available: https://it.wikipedia.org/wiki/Stato\_dell%27arte. |
| [4] | «Comparison Search Engine,» [Online]. Available: https://dspace3-labs.atmire.com/bitstream/handle/123456789/7634/338.pdf. |