Rilevamento di plagio

Da Wikipedia, l'enciclopedia libera
Salta alla navigazioneSalta alla ricerca

Questo articolo potrebbe **richiedere la pulizia** per soddisfare gli standard di qualità di Wikipedia . Non è stato specificato alcun motivo di pulizia . Si prega di contribuire a migliorare questo articolo, se possibile. (Dicembre 2010) (Scopri come e quando rimuovere questo messaggio modello)

Il rilevamento di plagio è il processo di individuazione di casi di plagio in un'opera o un documento. L'uso diffuso dei computer e l'avvento di Internet hanno reso più facile plagiare il lavoro degli altri. La maggior parte dei casi di plagio si trovano nel mondo accademico, dove i documenti sono in genere saggi o relazioni. Tuttavia, il plagio può essere trovato praticamente in qualsiasi campo, compresi romanzi, articoli scientifici, disegni d'arte e codice sorgente.

Il rilevamento di plagio può essere manuale o assistito dal software. Il rilevamento manuale richiede uno sforzo notevole e un'eccellente memoria, ed è poco pratico nei casi in cui devono essere confrontati troppi documenti oppure i documenti originali non sono disponibili per il confronto. Il rilevamento assistito da software consente di confrontare vaste raccolte di documenti tra loro, rendendo molto più probabile la rilevazione di successo.

La pratica del plagio mediante l'utilizzo di sufficienti sostituzioni di parole per eludere il software di rilevamento è nota come rogeting . [1]

Contenuto

[mostra]

Rilevazione assistita dal software [modifica]

Il rilevamento di plagio assistito da computer (CaPD) è un'attività di recupero di informazioni (IR) supportata da sistemi IR specializzati, denominati sistemi di rilevamento di plagio (PDS).

Nei documenti di testo [modifica]

I sistemi per il rilevamento del plagio di testo implementano uno dei due approcci di rilevamento generici, uno esterno e l'altro intrinseco. [2] I sistemi di rilevamento esterni confrontano un documento sospetto con una raccolta di riferimento, che è una serie di documenti ritenuti autentici. [3] Sulla base di un modello di documento scelto e di criteri di somiglianza predefiniti, l'attività di rilevamento consiste nel recuperare tutti i documenti che contengono un testo simile a un grado al di sopra della soglia scelta per il testo nel documento sospetto. [4] PDS intrinseco analizza esclusivamente il testo da valutare senza eseguire confronti con documenti esterni. Questo approccio mira a riconoscere i cambiamenti nello stile di scrittura unico di un autore come un indicatore di potenziale plagio.[5] PDS non sono in grado di identificare in modo affidabile il plagio senza giudizio umano. Le somiglianze sono calcolate con l'aiuto di modelli di documento predefiniti e potrebbero rappresentare falsi positivi. [6] [7] [8] [9] [10]

Efficacia delle impostazioni di istruzione superiore [modifica]

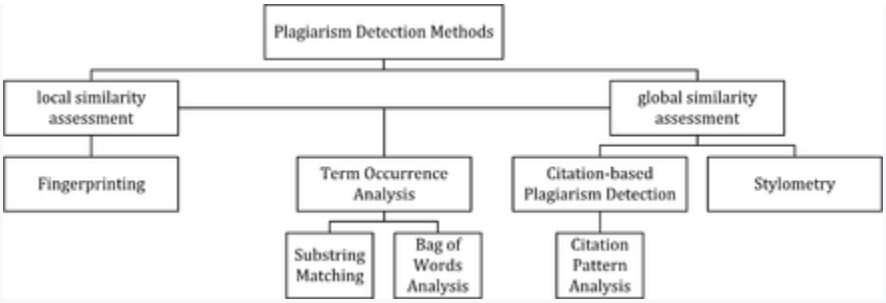
Questa sezione **si basa in gran parte o interamente su un'unica fonte** . La discussione pertinente può essere trovata nella pagina di discussione . Aiutateci a migliorare questo articolo introducendo citazioni su fonti aggiuntive. (Dicembre 2017)

È stato condotto uno studio per testare l'efficacia del software di rilevamento di plagio in un contesto di istruzione superiore. Una parte dello studio ha assegnato un gruppo di studenti a scrivere un articolo. Questi studenti sono stati prima istruiti sul plagio e hanno informato che il loro lavoro doveva essere gestito attraverso un sistema di rilevamento dei plagi. Un secondo gruppo di studenti è stato incaricato di scrivere un articolo senza alcuna informazione sul plagio. I ricercatori si

aspettavano di trovare tassi più bassi nel primo gruppo, ma hanno trovato all'incirca lo stesso tasso di plagio in entrambi i gruppi. [11]

Approcci [modifica]

La figura seguente rappresenta una classificazione di tutti gli approcci di rilevamento attualmente in uso per il rilevamento di plagio assistito da computer. Gli approcci sono caratterizzati dal tipo di valutazione di similarità che intraprendono: globale o locale. Gli approcci di valutazione della similarità globale utilizzano le caratteristiche ricavate da parti più ampie del testo o dal documento nel suo complesso per calcolare la somiglianza, mentre i metodi locali esaminano solo i segmenti di testo preselezionati come input.



Classificazione dei metodi di rilevamento del plagio assistito da computer

Fingerprinting [modifica]

L'impronta digitale è attualmente l'approccio più ampiamente applicato al rilevamento di plagio. Questo metodo costituisce una sintesi rappresentativa dei documenti selezionando da essi una serie di sottostringhe multiple (n-grammi). Gli insiemi rappresentano le impronte digitali ei loro elementi sono chiamati minuzie. [12] [13] Un documento sospetto viene controllato per il plagio calcolando la sua impronta digitale e interrogando le minuzie con un indice precompilato di impronte digitali per tutti i documenti di una collezione di riferimento. Minutie corrispondenti a quelle di altri documenti indicano segmenti di testo condivisi e suggeriscono potenziali plagi se superano una soglia di similarità prescelta. [14]Le risorse e il tempo computazionale sono fattori limitanti per il rilevamento delle impronte digitali, motivo per cui questo metodo in genere confronta solo un sottoinsieme di minuzie per accelerare il calcolo e consentire controlli in raccolte molto grandi, come Internet. [12]

Accoppiamento di stringhe [modifica]

La corrispondenza delle stringhe è un approccio prevalente utilizzato nell'informatica. Quando viene applicato al problema del rilevamento di plagio, i documenti vengono confrontati per sovrapposizione di testo letterale. Sono stati proposti numerosi metodi per affrontare questo compito, alcuni dei quali sono stati adattati al rilevamento di plagio esterno. Il controllo di un documento sospetto in questa impostazione richiede il calcolo e l'archiviazione di rappresentazioni comparabili in modo efficiente per tutti i documenti della raccolta di riferimento per confrontarli a coppie. In generale, per questa attività sono stati utilizzati modelli di documento suffisso, quali alberi di suffisso o vettori di suffisso. Ciò nonostante, la corrispondenza della sottostringa rimane costosa dal punto di vista computazionale, il che la rende una soluzione non valida per il controllo di ampie raccolte di documenti. [15] [16] [17]

Sacchetto di parole [modifica]

L'analisi del sacco di parole rappresenta l'adozione del recupero dello spazio vettoriale , un tradizionale concetto di IR, al dominio del rilevamento del plagio. I documenti sono rappresentati come uno o più vettori, ad esempio per parti di documenti diversi, che vengono utilizzati per calcoli di similarità saggia di coppia. Il calcolo della similarità può quindi basarsi sulla misura della similarità del coseno tradizionale o su misure di similarità più sofisticate. [18] [19] [20]

Analisi delle citazioni [modifica]

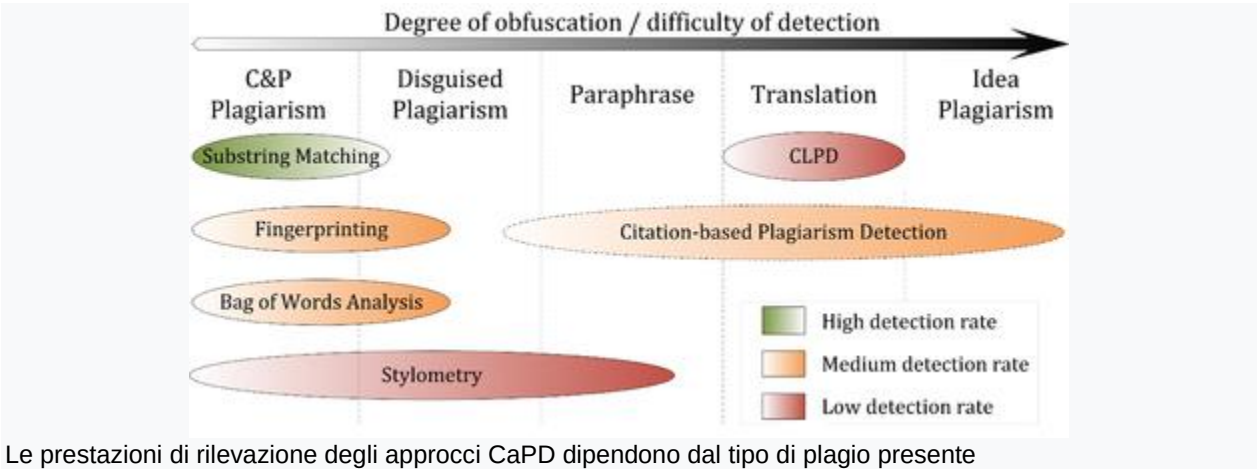
Il rilevamento di plagio basato su citazioni (CbPD) [21] si basa sull'analisi delle citazioni ed è l'unico approccio al rilevamento di plagio che non si basa sulla somiglianza testuale. [23]CbPD esamina le citazioni e le informazioni di riferimento nei testi per identificare modelli simili nelle sequenze di citazioni. In quanto tale, questo approccio è adatto per testi scientifici o altri documenti accademici che contengono citazioni. L'analisi delle citazioni per individuare il plagio è un concetto relativamente giovane. Non è stato adottato dal software commerciale, ma esiste un primo prototipo di un sistema di rilevamento del plagio basato su citazioni. [23]L'ordine simile e la prossimità delle citazioni nei documenti esaminati sono i criteri principali utilizzati per calcolare le somiglianze del modello di citazione. I modelli di citazione rappresentano sequenze non esclusive contenenti citazioni condivise dai documenti confrontati. [22] [24] I fattori, incluso il numero assoluto o la frazione relativa delle citazioni condivise nel modello, nonché la probabilità che le citazioni coincidano in un documento sono anche considerati per quantificare il grado di somiglianza dei modelli. [22] [24] [25] [26]

Stylometry [modifica]

La stitometria include metodi statistici per quantificare lo stile di scrittura unico di un autore [27] [28] ed è usata principalmente per l'attribuzione dell'autore o CaPD intrinseco. Costruendo e confrontando modelli stilometrici per segmenti di testo diversi, è possibile rilevare passaggi stilisticamente diversi dagli altri, quindi potenzialmente plagiati. [5]

Performance [modifica]

Valutazioni comparative dei sistemi di rilevamento dei plagi [3] [29] [30] [31] [32] [33] indicano che le loro prestazioni dipendono dal tipo di plagio presente (vedi figura). Ad eccezione dell'analisi del modello di citazione, tutti gli approcci di rilevamento si basano sulla somiglianza testuale. È quindi sintomatico che l'accuratezza del rilevamento diminuisca e più casi di plagio vengano offuscati.



Le prestazioni di rilevazione degli approcci CaPD dipendono dal tipo di plagio presente

Le copie letterali, il plagio di copia e incolla (c&i) o i casi di plagio discretamente mascherati possono essere rilevati con elevata accuratezza dal PDS esterno corrente se la fonte è accessibile al software. Soprattutto le procedure di corrispondenza delle sottostringhe ottengono buone prestazioni per il plagio di c&i, dal momento che comunemente usano modelli di documento senza perdita di dati, come alberi di suffissi. Le prestazioni dei sistemi che utilizzano l'impronta digitale o l'analisi del pacco di parole nella rilevazione delle copie dipendono dalla perdita di informazioni causata dal modello di documento utilizzato. Applicando strategie di pezzi flessibili e strategie di selezione, essi sono meglio in grado di rilevare forme moderate di plagio mascherato rispetto alle procedure di corrispondenza di sottostringa.

Il rilevamento intrinseco del plagio tramite l'uso della stetoscopia può superare in qualche misura i confini della somiglianza testuale confrontando la somiglianza linguistica. Dato che le differenze

stilistiche tra segmenti plagiati e originali sono significative e possono essere identificate in modo affidabile, la stilizzazione può aiutare a identificare il plagio mascherato e parafrasato . È probabile che i confronti stilometrici falliscano nei casi in cui i segmenti sono fortemente parafrasati al punto in cui essi assomigliano più strettamente allo stile di scrittura personale del plagiatore o se un testo è stato compilato da più autori. I risultati dei concorsi internazionali sul rilevamento di plagio nel 2009, 2010 e 2011, [3] [32] [33] e gli esperimenti condotti da Stein,[34] indicano che l'analisi stilometrica sembra funzionare in modo affidabile solo per lunghezze di documenti di diverse migliaia o decine di migliaia di parole, il che limita l'applicabilità del metodo alle impostazioni di CaPD.

Una quantità crescente di ricerca viene eseguita su metodi e sistemi in grado di rilevare plagi tradotti. Attualmente, il rilevamento del plagio tra lingue diverse (CLPD) non è visto come una tecnologia matura [35] e i rispettivi sistemi non sono stati in grado di ottenere risultati di rilevamento soddisfacenti nella pratica. [31]

Il rilevamento del plagio basato su citazioni utilizzando l'analisi del modello di citazione è in grado di identificare parafrasi più forti e traduzioni con percentuali di successo più elevate rispetto ad altri approcci di rilevamento, poiché è indipendente dalle caratteristiche testuali. [22] [25] Tuttavia, poiché l'analisi del modello di citazione dipende dalla disponibilità di sufficienti informazioni sulla citazione, è limitata ai testi accademici. Resta inferiore agli approcci testuali nella rilevazione di passaggi plagiari più brevi, che sono tipici per i casi di plagio di copia e incolla o di agita e incolla; quest'ultimo si riferisce alla miscelazione di frammenti leggermente alterati provenienti da fonti diverse. [36]

Software [modifica]

La progettazione di software di rilevamento di plagio da utilizzare con documenti di testo è caratterizzata da una serie di fattori: [citazione necessaria]

Fattore	Descrizione e alternative
Scopo della ricerca	Nell'internet pubblico, utilizzando motori di ricerca / database istituzionali / database locali, database specifici del sistema. [citazione necessaria]
Tempo di analisi	Ritardo tra il momento in cui un documento viene inviato e il momento in cui i risultati sono resi disponibili. [citazione necessaria]
Capacità del documento / elaborazione in lotti	Numero di documenti che il sistema può elaborare per unità di tempo. [citazione necessaria]
Controllo di intensità	Con quale frequenza e per quali tipi di frammenti di documento (paragrafi, frasi, sequenze di parole a lunghezza fissa) il sistema esegue una ricerca su risorse esterne, come i motori di ricerca.
Confronto tra tipi di algoritmo	Algoritmi che definiscono il modo che il sistema utilizza per confrontare i documenti l'uno con l'altro. [citazione necessaria]

Precisione e richiamo	Numero di documenti correttamente contrassegnati come plagiati comparati al numero totale di documenti contrassegnati e al numero totale di documenti effettivamente plagiati. Alta precisione significa che sono stati trovati pochi falsi positivi e un alto richiamo significa che pochi falsi negativi sono stati lasciati inosservati. [citazione necessaria]

La maggior parte dei sistemi di rilevamento di plagio su larga scala utilizza database interni di grandi dimensioni (oltre ad altre risorse) che crescono con ogni documento aggiuntivo inviato per l'analisi. Tuttavia, questa funzione è considerata da alcuni come una violazione del copyright degli studenti . [citazione necessaria]

Nel codice sorgente [modifica]

Anche il plagio nel codice sorgente del computer è frequente e richiede strumenti diversi da quelli utilizzati per i confronti di testo nel documento. Una ricerca significativa è stata dedicata al plagio accademico del codice sorgente. [37]

Un aspetto distintivo del plagio del codice sorgente è che non esistono mulini di saggio, come quelli che si possono trovare nel plagio tradizionale. Poiché la maggior parte degli incarichi di programmazione prevede che gli studenti scrivano programmi con requisiti molto specifici, è molto difficile trovare programmi esistenti che li soddisfano già tali requisiti. Poiché l'integrazione di codice esterno è spesso più difficile rispetto alla scrittura da zero, la maggior parte degli studenti che plagiano scelgono di farlo dai propri pari.

Secondo Roy e Cordy, [38] gli algoritmi di rilevamento della similarità del codice sorgente possono essere classificati come basati su

- Stringhe: cerca corrispondenze testuali esatte di segmenti, ad esempio esecuzioni di cinque parole. Veloce, ma può essere confuso rinominando gli identificatori.

- Simboli - come con le stringhe, ma usando un lexer per convertire prima il programma in simboli. Questo elimina gli spazi bianchi, i commenti e i nomi degli identificatori, rendendo il sistema più robusto a semplici sostituzioni di testo. La maggior parte dei sistemi di rilevamento del plagio accademico funziona a questo livello, utilizzando diversi algoritmi per misurare la somiglianza tra sequenze di simboli.

- Alberi parsee: costruisce e confronta gli alberi parse. Ciò consente di rilevare somiglianze ad un livello superiore. Ad esempio, il confronto tra alberi può normalizzare le istruzioni condizionali e rilevare costrutti equivalenti simili tra loro.

- Program Dependency Graphs (PDGs): un PDG acquisisce il flusso effettivo di controllo in un programma e consente di localizzare equivalenze di livello più elevato, con una maggiore dispendio di complessità e tempi di calcolo.

- Metriche: le metriche acquisiscono "punteggi" di segmenti di codice in base a determinati criteri; ad esempio, "il numero di cicli e condizionali" o "il numero di diverse variabili utilizzate". Le metriche sono semplici da calcolare e possono essere confrontate rapidamente, ma possono anche portare a falsi positivi: due frammenti con gli stessi punteggi su un insieme di metriche possono fare cose completamente diverse.

- Approcci ibridi: ad esempio, gli alberi di analisi + gli alberi di suffisso possono combinare la capacità di rilevamento degli alberi di analisi con la velocità offerta dagli alberi di suffisso, un tipo di struttura di dati di corrispondenza delle stringhe.

La classificazione precedente è stata sviluppata per il refactoring del codice e non per il rilevamento di plagio accademico (un obiettivo importante del refactoring è di evitare il codice duplicato, indicato come cloni di codice in letteratura). Gli approcci di cui abbiamo parlato sopra sono efficaci contro diversi livelli di somiglianza; la similarità di basso livello si riferisce al testo identico, mentre la somiglianza di alto livello può essere dovuta a specifiche simili. In un contesto accademico, quando tutti gli studenti sono tenuti a codificare secondo le stesse specifiche, il codice funzionalmente equivalente (con somiglianza di alto livello) è del tutto previsto, e solo la somiglianza di basso livello è considerata una prova di imbroglio.