UNIVERSITÀ DI BOLOGNA

TESI DI LAUREA IN INFORMATICA

Real-Time Collaborative Editing: modelli e strumenti per l'analisi della collaborazione in spazio e tempo

---

Autore:
Luca ALTARIVA

Relatore:
Angelo DI IORIO

7 dicembre 2018

UNIVERSITÀ DI BOLOGNA

Abstract

Scuola di Scienze
Informatica

Real-Time Collaborative Editing: modelli e strumenti per l'analisi della collaborazione in spazio e tempo

di Luca ALTARIVA

Gli editor collaborativi in tempo reale (RTCE) sono strumenti vastamente utilizzati sul web. Tuttavia, ad oggi esistono pochissimi studi che analizzano i pattern di utilizzo di questi strumenti. Questo lavoro si propone di espandere uno di questi studi, utilizzando un modello precedentemente definito per l'analisi dell'uso degli RTCE a partire dal loro registro delle modifiche, per crearne uno piú fine. Infine, verrà mostrato un esempio di come questo nuovo modello possa essere implementato e applicato per effettuare l'analisi di un grande numero di documenti di testo.

Ringraziamenti

Un ringraziamento speciale al mio relatore, il professore Angelo Di Iorio, per tutto l'aiuto fornito e per l'incredibile disponibilità dimostrata durante la realizzazione di questa tesi.

Un altro ringraziamento è dedicato a tutta la mia famiglia, per essere stata al mio fianco per ventidue incredibili anni.

Un ulteriore ringraziamento a tutti i miei amici e colleghi, per avermi sempre sopportato e sostenuto.

Ringrazio infine tutte le persone buone che ho incontrato lungo la strada, per avere reso questo mondo un posto migliore.

Grazie,
Luca Altariva

Indice

Elenco delle figure

Dedicata ad un folle, di cedere inesperto,
e alla nave con cui è affondato.

Capitolo 1

Introduzione

Un editor collaborativo in tempo reale è un programma che permette a piú persone di modificare un documento in simultanea. Esempi di tali programmi sono Etherpad [7] e Google Docs [9], che permettono a piú utenti di lavorare ad un documento di testo contemporaneamente, o anche in diversi momenti.

In teoria, questi strumenti rendono possibile una collaborazione molto piú efficiente, fornendo degli strumenti di lavoro molto potenti per la stesura simultanea di un documento di testo.

Tuttavia, è necessario chiedersi se questa possibilità di modifica simultanea venga effettivamente sfruttata. È forse possibile che questi programmi vengano utilizzati per usufruire di altre funzionalità, come la sincronizzazione in cloud delle modifiche o per il mantenimento di una cronologia delle versioni del documento, e che la possibilità di modifica simultanea di un documento venga posta in secondo piano? Per rispondere a questa domanda sono stati effettuati diversi studi [15, 3, 2, 5]. In particolare, D'Angelo et al. [6], propongono un modello concettuale per classificare le operazioni di modifica effettuate da un utente come "collaborative" o "non-collaborative" a seconda della sua distanza - in posizione (spazio), tempo o entrambi - da modifiche effettuate da altri autori. Questo modello viene usato per analizzare le cronologie delle modifiche di 14000 documenti, per concludere che:

1. Metà dei documenti hanno un singolo autore e quindi nessuna collaborazione.
2. La collaborazione su parti vicine di un documento avviene spesso, ma solo asincronicamente, con gli autori che ci lavorano sopra a turni.
3. La collaborazione simultanea su parti vicine di un documento avviene molto raramente.

Questa tesi si propone di:

- Creare un modello che vada ad espandere quello precedente.
- Introdurre un nuovo concetto per la misura della collaborazione.
- Mostrare un esempio di come il nuovo modello possa essere usato per estrarre dei dati in maniera bottom-up.

Per cominciare, si tratterà di vari studi legati agli editor collaborativi in tempo reale (capitolo 2). A seguire, si parlerà del nuovo modello partendo dall'introdurre quello precedente e arrivando ai nuovi concetti (capitolo 3), per poi mostrare un esempio della sua applicazione (capitolo 4). Per concludere, si discuterà di una sua implementazione (capitolo 5) che verrà poi utilizzata per l'estrazione e l'analisi di diversi dati, di cui si tratta nel capitolo 6.

Capitolo 2

Studi sugli editor collaborativi in tempo reale

Prima di [6], il tema dell'uso di diversi editor collaborativi in tempo reale (*RTCE, Real-Time Collaborative Editor*) è stato trattato da diversi autori.

Molte ricerche si sono concentrate sull'intervistare gruppi di persone riguardo allo stato del loro rapporto con gli RTCE. Per esempio, Chu e Kenney [3] hanno intervistato 22 studenti universitari che avevano utilizzato sia Google Docs che MediaWiki per un corso, chiedendo quali fossero i loro pareri. Tutti gli studenti hanno risposto di apprezzare l'interfaccia offerta da Google Docs, ma non lo hanno considerato uno strumento strettamente superiore nell'ambito collaborativo.

Un'analisi più ampia è stata condotta da Brodhal et al. [2] intervistando 166 studenti, la cui risposta è stata principalmente scettica verso gli RTCE. La maggior parte di essi, infatti, ha riferito di pensare che l'uso di strumenti collaborativi non abbia contribuito ad aumentare la qualità della collaborazione. Nel 70% dei casi le funzionalità degli strumenti sono addirittura risultate insoddisfacenti. Secondo gli studenti infatti, aggiungere commenti al contenuto sarebbe stato più utile che la possibilità di scrivere collaborativamente.

Un altro lavoro simile è quello di Wang et al., che in [5] descrivono i risultati di un questionario inviato a 30 persone del campo sia industriale che accademico, nel quale venivano fatte domande sull'uso quotidiano degli RTCE. Tra le risposte, il consenso era che le funzionalità degli strumenti fossero preziose per la scrittura collaborativa, ma venne anche osservato come l'adozione su larga scala di questi strumenti fosse ostacolata da problemi di natura sociale, personale o di privacy.

Inoltre, in [14] degli esperti nell'analisi delle abitudini di scrittura con RTCE discutono sui vantaggi e sulle difficoltà dell'uso di Google Docs. Insieme, identificano diversi pattern di collaborazione e alcune problematiche, sia tecniche che collegate alla pianificazione delle attività tra collaboratori, che dovrebbero essere risolte per massimizzare la resa degli RTCE.

Altre ricerche si sono concentrate sull'analisi delle prestazioni e sulla differenza dei vari protocolli e algoritmi utilizzati negli RTCE [13, 4, 12, 8, 1]. A differenza di questi studi, questo lavoro si vuole concentrare su come le persone collaborino a prescindere dalle implementazioni che vengono utilizzate.

Esistono poche analisi quantitative dei registri delle modifiche degli RTCE effettuate per studiare il comportamento degli utenti. La più recente di queste è uno studio di Yunting Sun et al. [15], dove sono stati analizzati i registri dell'attività di tutti gli impiegati di Google per verificare se ad un incremento dell'utilizzo di editor collaborativi corrispondesse un incremento della collaborazione. È stata quindi definita una tecnica per visualizzare la cronologia delle modifiche di un documento per poi essere in grado di analizzarla. La loro conclusione è stata che negli ultimi anni sia la percentuale degli impiegati che collaborano su Google Docs ogni mese

che quella degli impiegati che collaborano con più di due persone è salita, andando addirittura a raddoppiare nel secondo caso.

In maniera simile al modello proposto in [6], Sun ha definito delle finestre temporali di 15 minuti, utilizzate in modo leggermente diverso, ma comunque con lo scopo di riconoscere le modifiche effettuate quasi simultaneamente. Tuttavia, la ricerca si limitava a considerare l'aspetto temporale della collaborazione, senza considerare la posizione delle modifiche come invece succede in [6]. Altre due ricerche sull'analisi quantitativa sono quelle effettuate da Birnholtz et al. [10] e Olson et al. [11]. Il primo di questi lavori riguarda un esperimento in laboratorio con 150 studenti, dove è stato osservato il loro uso di Google Docs - sia con che senza la chat integrata, sia per la collaborazione sincrona che per quella asincrona - per cercare di capire quanto le persone comunicassero durante la collaborazione su un documento di testo e le eventuali relazioni presenti tra la collaborazione nel lavoro e le relazioni sociali. Il test è consistito sia nell'uso di questionari che nell'analisi di vari registri, tra cui quello delle modifiche, e ha avuto, tra le varie conclusioni, che quest'ultime aumentano con la comunicazione, e che effettuarne troppe può avere un effetto negativo sulla collaborazione.

Più recentemente, Olson et al. [11] hanno analizzato un grande numero di registri di Google Docs per studiare come diversi team di studenti scrivono collaborativamente, raccogliendo 96 documenti scritti nell'arco di 3 anni. L'articolo introduce la nozione di slice, definito come un'istantanea del documento, e di session, definito come un insieme di slices. Collegando gli slices a dei timestamp, è stato possibile unirli in delle sessioni, per poi andare a cercare nelle cronologie delle modifiche dei documenti quante di queste fossero collaborative. L'esperimento ha concluso che la qualità dei documenti è aumentata insieme alla collaborazione, e che il numero di modifiche simultanee è collegata al rapporto sociale tra chi le effettua.

Come questi due studi, questo lavoro si concentra sull'analisi quantitativa dei registri di un editor collaborativo in tempo reale, tuttavia differisce da questi in diversi aspetti principali:

- L'insieme dei documenti che vengono presi in considerazione è molto più elevato, e proviene da una fonte completamente anonima e incontrollata. Pertanto, i documenti non sono stati prodotti ai fini di questa ricerca e inoltre non vengono effettuate analisi sulle composizioni dei loro autori.

- Questo lavoro si concentra sulla pura vicinanza delle modifiche - da un punto di vista spaziale e temporale - invece che su aspetti di comunicazione e relazionali.

- La piattaforma di riferimento è Etherpad, un editor collaborativo in tempo reale simile ma non uguale a Google Docs.

Mentre gli studi quantitativi dei registri delle modifiche risultano scarsi nel campo degli RTCE, questi sono molto più comuni in altri contesti, come quello delle wiki. In questo campo, sono stati proposti diversi modelli e i ricercatori sono riusciti ad analizzare a fondo i pattern d'uso delle varie wiki, tra cui principalmente Wikipedia.

Capitolo 3

Il modello di misura della collaborazione

3.1 Introduzione

In questo capitolo si tratterà della base teorica utilizzata per essere in grado di fornire una misura della collaborazione in un documento di testo realizzato in un editor collaborativo in tempo reale. La parte iniziale consisterà nella discussione del metodo introdotto da D'Angelo et al. in [6], dove viene utilizzato per introdurre delle metriche di valutazione della collaborazione in un documento. Queste metriche sono state applicate sui registri delle modifiche (log) di Etherpad, per cercare quali di queste modifiche siano collaborative, per poi studiare i risultati.

Per cercare la collaborazione nelle modifiche di Etherpad, introducono e definiscono tre diversi concetti:

1. **Collaborazione spaziale**, quando una modifica lavora su un'area di testo abbastanza vicina ad un'altra di un altro autore.

2. **Collaborazione temporale**, quando una modifica è quasi simultanea a quella di un altro autore.

3. **Collaborazione spazio-temporale**, quando una modifica è quasi simultanea e su un'area di testo abbastanza vicina ad un'altra di un altro autore.

Per definire quando due aree di testo siano "abbastanza vicine" o quando due modifiche siano "quasi simultanee" si definisce una finestra spaziale o temporale che indica il limite entro il quale avviene la collaborazione. Per esempio, due modifiche effettuate a 13 secondi l'una dall'altra saranno collaborative con una finestra temporale di 15 secondi ma non con una di 10.

Tuttavia, in questo approccio, queste finestre sono scelte arbitrariamente. Inoltre, quando viene analizzato un documento di testo, l'unica informazione che viene estratta è, per ognuna delle modifiche effettuate, se queste siano collaborative con un'altra modifica qualsiasi di un altro autore o meno.

Questo lavoro si propone di espandere il metodo precedente partendo da quel modello di analisi della collaborazione per proporne uno più fine, che si concentra non solo sull'analisi top-down dei documenti, ma permette anche un approccio bottom-up, che può fare emergere risultati interessanti a partire dai dati stessi. Nel modello che viene proposto non si cerca più se sia presente collaborazione tra due modifiche nello stesso documento, si assume invece che ci sia e ci si domanda in che quantità. In questo modo, è possibile analizzare ogni modifica andando a misurare la sua distanza collaborativa rispetto a una qualsiasi modifica di ogni altro autore (Fig. 3.1).

Modello precedente	Modello nuovo
È collaborativo con finestra X	Ha distanza J dall'autore A, K da B e L da C
Non è collaborativo con finestra X	Ha distanza L dall'autore A, M da B e N da C

FIGURA 3.1: I due modelli messi a confronto

Vengono quindi introdotti tre nuovi concetti:

1. **Distanza collaborativa spaziale**, misura quanto una modifica sia vicino a una di altri autori.
2. **Distanza collaborativa temporale assoluta**, misura quanto tempo passa tra una modifica e una di altri autori.
3. **Distanza collaborativa temporale relativa**, misura quante modifiche avvengono tra una modifica e una di altri autori.

Si faccia caso a come siano presenti due nozioni di distanza collaborativa temporale: quella assoluta, volta a rispecchiare la definizione originale di collaborazione temporale, e quella relativa, che introduce una nozione completamente nuova rispetto al modello precedente, e che verrà utilizzata per misurare la quantità di collaborazione in relazione a quanto lavoro viene effettuato sul documento che viene esaminato. Si noti inoltre la mancanza di una distanza collaborativa spazio-temporale introdotta invece nel modello precedente. Questa assenza sarà giustificata più avanti a seguito di tutte le altre definizioni.

Definito il modello, questo potrà essere utilizzato per misurare la distanza collaborativa di ogni modifica effettuata su un documento rispetto ad ogni autore che ci abbia lavorato sopra. In questo modo, si ottiene una matrice modifica per autore che potrà essere utilizzata per ulteriori analisi sia top-down, con lo stesso stampo del lavoro precedente, che bottom-up, che può quindi derivare le proprie conclusioni a partire dai dati stessi. Questo lavoro si limiterà ad esaminare le matrici ottenute per l'analisi delle finestre scelte nell'articolo precedente, per proporne di nuove, e per effettuare qualche osservazione preliminare sui valori ottenuti.

3.2 Nozioni preliminari

3.2.1 Etherpad

Per descrivere il modello utilizzato per questa analisi, è necessario innanzitutto soffermarsi su alcuni dettagli che riguardano l'editor di testo da cui provengono i file che vengono analizzati.

Etherpad [7] è un editor collaborativo in tempo reale basato sul web e implementato utilizzando Node.js. Permette a più utenti di lavorare in contemporanea su un documento di testo, definito pad, utilizzando la sua interfaccia tramite un browser web.

Ogni pad ha un registro di tutte le operazioni - definite edit - effettuate su di esso, contenente ogni edit compiuto da tutti gli utenti su quel pad. I file che sono

stati analizzati - che da qui in avanti verranno chiamati file pad - consistono in una successione di oggetti JSON ognuno dei quali rappresenta un edit, e sono stati ottenuti dal lavoro precedente elaborando dei dati provenienti dal database di etherpad. Ogni file pad contiene tutti e soli gli edit di un particolare pad.

Gli oggetti JSON hanno la seguente struttura:

```
{
    "Pad":"nomePad",
    "Revision":"1.0",
    "Author":"nomeAutore",
    "Timestamp":1429102096175,
    "opCode":"INS",
    "preDocLength":0,
    "postDocLength":1,
    "touchedCharsLength":1,
    "preInterval":[1,1],
    "postInterval":[1,2]
}
```

Ogni edit rappresentato ha quindi diversi attributi. Pad e Author rappresentano rispettivamente il nome del Pad su cui è stato effettuato l'edit e dell'autore che lo ha causato. touchedCharsLenght rappresenta quanti caratteri vengono coinvolti. preInterval e postInterval rappresentano l'intervallo di caratteri su cui si va ad agire prima e dopo l'edit. Infine, opCode rappresenta l'operazione effettuata, e può essere di tre tipi diversi:

- INS per l'inserimento di una stringa lunga touchedCharsLength in preInterval.
- DEL per cancellare la stringa contenuta in preInterval.
- UPD per cambiare il contenuto di una stringa in preInterval.

3.2.2 Notazioni

Per effettuare una descrizione del modello, è inoltre necessario introdurre qualche notazione:

- **Intervallo** p[inizio, fine] indica la sequenza di caratteri nel pad p, a partire dalla posizione inizio (che parte da 0) e fino alla posizione fine.
- **Lunghezza** lenght(p) rappresenta il numero di caratteri in p.
- **Timestamp** time(e) è il timestamp dell'edit e.
- **Autore** author(e) è la stringa che identifica l'autore dell'edit e.
- **Appartenenza** ad un pad pad(e) è il pad che contiene l'edit e.
- **Posizione** position(e) è la posizione relativa dell'edit e rispetto agli altri edit, ordinati per time(e) crescente.

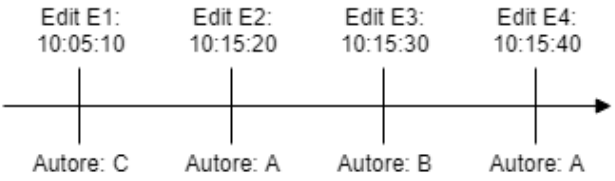


FIGURA 3.2: Questa figura mostra una successione di edit con i relativi timestamp.

3.3 Caratterizzazione della collaborazione spazio-temporale

A questo punto è possibile discutere della caratterizzazione della collaborazione spazio-temporale proposta da D'Angelo et al. nel loro articolo [6]. Questa concettualizzazione vuole classificare ogni singolo edit come collaborativo o meno in tempo, spazio o entrambi. Per far questo, vengono introdotte varie definizioni di collaborazione.

3.3.1 Collaborazione temporale

Secondo l'articolo, l'idea è quella di considerare un edit collaborativo se è "abbastanza vicino in tempo ad un edit creato da un autore diverso". Le uniche informazioni degli edit che è necessario valutare per questa proprietà sono i timestamp e gli autori di ogni edit; i cambiamenti che vengono applicati al pad sono irrilevanti, così come le loro posizioni spaziali nel documento. Inoltre, si necessita di una misura globale - chiamata finestra temporale - per esprimere la nozione di essere "abbastanza vicini" (temporalmente).

Un edit è segnato come collaborativo se esiste almeno un altro edit che (a) è stato effettuato da un'altro autore (i.e., non è possibile collaborare con sè stessi) e (b) è avvenuto entro la finestra temporale scelta in precedenza. Formalmente:

Definizione 3.1 (Collaborazione temporale). Un edit e1 è collaborativo in tempo (o time-collaborative) relativamente ad una finestra temporale wt se e solo se:

$$time(e2) \in [time(e1) - wt, time(e1)] \wedge author(e1), author(e2) \tag{3.1}$$

Si prenda in considerazione la figura 3.2 e si consideri una finestra temporale wt pari a 30 secondi. In questo caso, E4 è collaborativo in tempo con E3, perché avviene entro la finestra temporale di 30s ed è stato effettuato da un autore diverso, mentre non è collaborativo in tempo con E1 ed E2 perché nel primo caso è fuori dalla finestra di 30s e nel secondo perché effettuato dal medesimo autore.

3.3.2 Collaborazione spaziale

La collaborazione spaziale misura se degli autori "abbiano lavorato nella stessa area di un pad". Per modellarlo, l'articolo comincia definendo innanzitutto la nozione di pre-intervallo e di post-intervallo per definire l'idea di "area" di un pad su cui gli edit agiscono.

- **Pre-intervallo** il pre-intervallo è l'intervallo che denota i caratteri che saranno coinvolti da un edit prima che esso avvenga.
- **Post-intervallo** Il post-intervallo è l'intervallo che denota i caratteri coinvolti in un edit che sono rimasti dopo che questo è avvenuto.

Queste due nozioni dipendono dalla natura dell'edit:

- Gli edit INS hanno pre-intervalli vuoti (perché non coinvolgono caratteri preesistenti) e post-intervalli non vuoti (i caratteri inseriti).
- Gli edit DEL hanno pre-intervalli non vuoti (i caratteri da cancellare) e i post-intervalli vuoti.
- Gli edit UPD hanno gli stessi valori per i pre e post-intervalli poiché non aggiungono né rimuovono caratteri.

I pre- e post-intervalli sono usati per modellare la collaborazione spaziale. Un edit è collaborativo nello spazio se è "abbastanza vicino" (spazialmente) ad un edit di un altro autore.

Per determinare se un edit è collaborativo spazialmente o meno, viene mantenuta una mappatura tra i caratteri dei pad e l'autore che gli ha modificati per ultimo. Viene applicata nuovamente la cronologia degli edit e, ogni qualvolta uno di questi venga applicato, viene confrontato il suo autore con gli autori dei caratteri coinvolti e vicini: se almeno uno di questi è associato ad un autore diverso, l'edit è considerato spazialmente collaborativo.

Per catturare la nozione di essere "abbastanza vicino" viene usata una finestra spaziale ws, espressa come numero di caratteri. La finestra è divisa in due ed aggiunta prima e dopo il pre-intervallo dell'edit considerato.

Formalmente, vengono definiti:

- **Ultimo edit** in un determinato momento t, lastedit(pad,p,t) denota il più recente edit il cui post-intervallo include il carattere di pad alla posizione p.
- **Espansione** $expand(pad[l,r],w) = pad[max(0,l-w),min(r+w,length(pad))]$

Utilizzando queste definizioni si può formalizzare la nozione di collaborazione spaziale come:

Definizione 3.2 (Collaborazione spaziale). Un edit e1 è collaborativo in spazio (o space-collaborative) rispetto ad una finestra spaziale ws se e solo se:

$$\begin{gathered} \exists\, c \mid c \in expand(preinterval(e1), ws) \\ \wedge\, author(lastedit(pad, c, t)) = author(e1) \\ \wedge\, pad(e1) = pad \\ \wedge\, time(e1) = t \end{gathered} \tag{3.2}$$

i.e., un edit è spazio-collaborativo quando c'è almeno un carattere nel suo pre-intervallo, espanso utilizzando la finestra spaziale, che è stato modificato più recentemente da un autore diverso.

La collaborazione spaziale di un dato pad può quindi essere calcolata contando il numero di edit segnati come spazialmente collaborativi secondo la precedente definizione.

3.3.3 Collaborazione spaziotemporale

Il modello dell'articolo è completato con una formalizzazione della nozione di collaborazione spaziotemporale. La nozione combina le idee di collaborazione spaziale e temporale per potere identificare edit che avvengono sia (quasi) sincronicamente che su aree vicine. La definizione formale è quindi derivata dalle precedenti, ma è più di una semplice congiunzione logica AND.

Un edit e viene segnato come spazio-temporalmente collaborativo quando è sia spazio- che tempo- collaborativo rispetto allo stesso edit e' di un altro autore. Questa idea è catturata nella seguente definizione:

Definizione 3.3 (Collaborazione spaziotemporale). L'edit e1 è collaborativo in spaziotempo (o spaziotempo-collaborativo) rispetto alla finestra temporale wt e alla finestra spaziale ws se e solo se:

$$\begin{gathered} \exists\, c \mid c \in expand(preinterval(e1), ws) \\ \wedge\, author(lastedit(pad, c, t)), author(e1)) \\ \wedge\, time(lastedit(pad, c, t) \in [time(e1) - wt, time(e1)] \\ \wedge\, pad(e1) = pad \\ \wedge\, time(e1) = t \end{gathered} \tag{3.3}$$

3.4 Caratterizzazione della distanza collaborativa

Questa parte tratterà del modello che va ad espandere il precedente, introducendo il concetto di distanza collaborativa.

3.4.1 Distanza collaborativa temporale assoluta

La distanza temporale assoluta vuole misurare quanto un edit "sia vicino" (temporalmente) ad un determinato autore. E.g., una distanza temporale molto bassa indicherà che l'edit è stato effettuato lavorando (quasi) in contemporanea con un altro autore, mentre una distanza molto alta che l'edit è stato effettuato in un momento diverso.

Per essere in grado di misurare la distanza tra un edit ed un autore, viene prima introdotta una nozione simile ma applicabile tra due edit. Definiamo dunque la distanza temporale assoluta tra due edit come la differenza (misurata in secondi) tra i timestamp dei due. Formalmente:

Definizione 3.4 (Distanza temporale assoluta tra due edit). Dati due edit e1 ed e2, la loro distanza temporale è definita come:

$$absolutetimedistance(e1, e2) = |time(e1) - time(e2)| \tag{3.4}$$

A questo punto è possibile definire la distanza temporale assoluta tra un edit e ed un autore a, considerando la minima distanza temporale assoluta tra e ed un qualsiasi edit e' che (a) non sia avvenuto dopo di e, (b) appartenga all'autore a e (c) appartenga allo stesso pad di e. Formalmente:

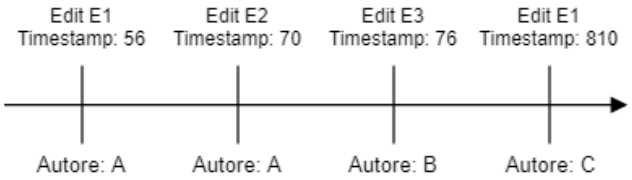


FIGURA 3.3: Un esempio di successione di edit, con i relativi timestamp e autori.

Definizione 3.5 (Distanza temporale assoluta tra un edit ed un autore). Dato un edit e1 ed un autore a, la loro distanza temporale assoluta è definita come:

$$\begin{aligned} absolutetimedistance(e1, a) = min(\infty, absolutetimedistance(e1, e2)), \\ \forall e2 \mid time(e1) \geq time(e2) \\ \wedge\, author(e2) = a \\ \wedge\, pad(e1) = pad(e2)) \end{aligned} \tag{3.5}$$

Si consideri come esempio la figura 3.3. Questa presenta una successione di edit di diversi autori. (semplificati per mostrare solamente i campi rilevanti). In questo caso:

- absolutetimedistance(1,4) = | 56 - 810 | = 754 applicando banalmente le definizioni.
- absolutetimedistance(3,A) = min(absolutetimedistance(3,2), absolutetimedistance(3,1)) = min(| 70 - 76 |, | 56 - 76 |) = 6 vengono considerati entrambi gli edit e viene presa la distanza minore.
- absolutetimedistance(3,B) = 0 poichè si considera la distanza tra 3 e 3.
- absolutetimedistance(3,C) = ∞ poichè nessun edit di C supera 3.

3.4.2 Distanza collaborativa temporale relativa

Come per la distanza temporale assoluta, la distanza temporale relativa intende catturare un'idea di "quanto è vicino" (temporalmente). Tuttavia, la distanza temporale relativa considera la distanza tra gli stessi edit (considerando un ordinamento cronologico), piuttosto che su quella tra i loro timestamp.

Questa distanza collaborativa fornisce un'idea migliore del grado di collaborazione su di un documento in relazione al suo utilizzo. Un documento che viene rielaborato nell'arco di mesi avrà una distanza temporale elevata, ma la sua distanza temporale relativa non sarà necessariamente alta.

Più due autori si alternano, più la distanza temporale relativa tra uno di essi e gli edit dell'altro tenderà ad 1. Ciò può avvenire sia nel caso in cui essi lavorino contemporaneamente, che nel caso di piccolissime modifiche alternate tra loro in un arco temporale più ampio (Fig. 3.4).

Se al contrario due autori lavorano in momenti temporali separati, la distanza temporale relativa tenderà ad aumentare con il numero di modifiche effettuate dallo stesso autore (Fig. 3.5).

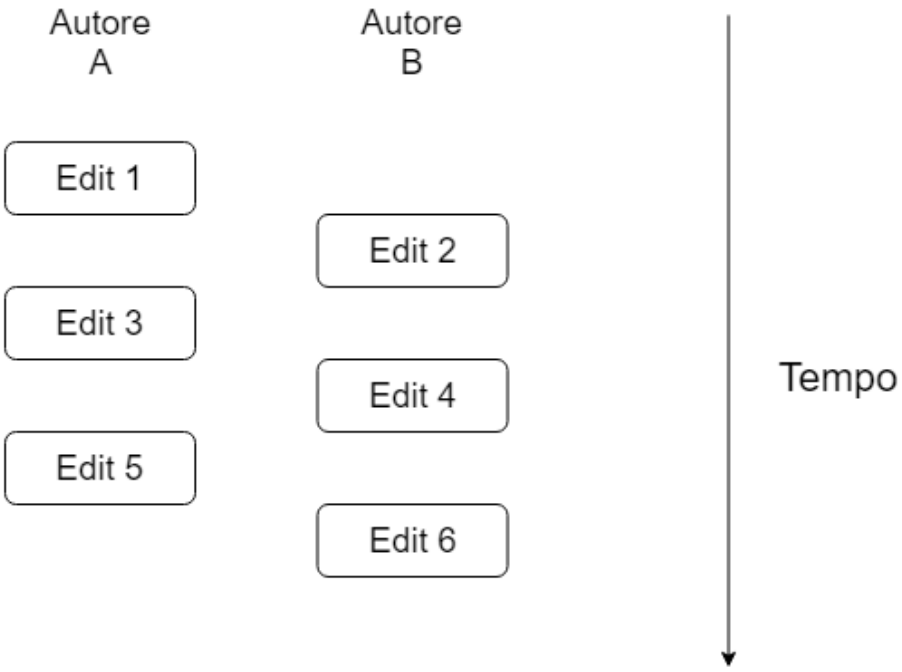


FIGURA 3.4: Esempio di una serie di edit che portano ad una collaborazione temporale relativa bassa.

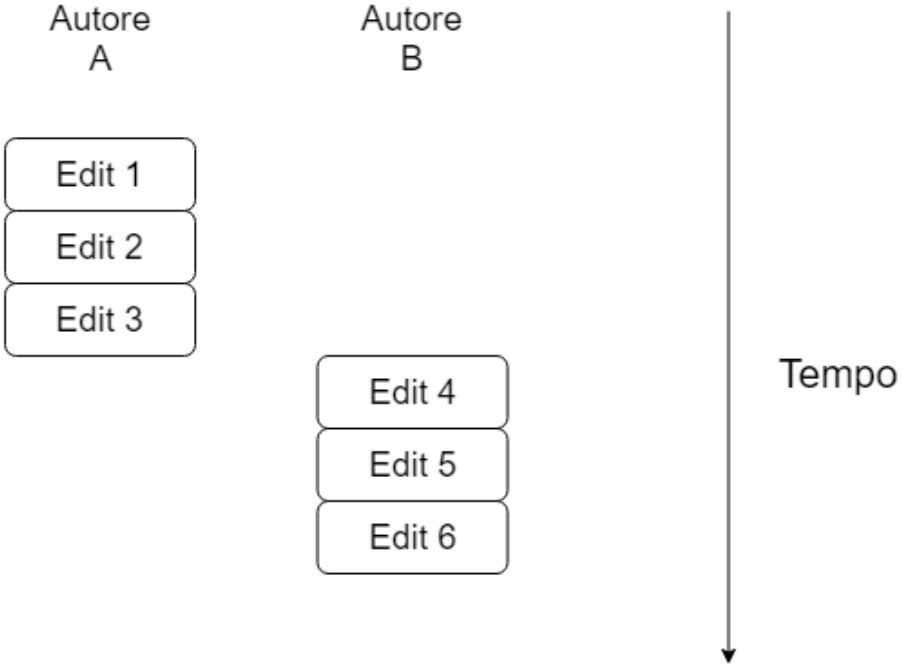


FIGURA 3.5: Esempio di una serie di edit che portano ad una collaborazione temporale relativa alta.

TABELLA 3.1: Confronto tra le distanze temporali relative e assolute degli edit in figura 3.3.

Argomenti	Distanza temporale assoluta	Distanza temporale relativa
$(1, 4)$	754	3
$(3, A)$	6	1
$(3, B)$	0	0
$(3, C)$	∞	∞

Come nel caso precedente verrà introdotto il concetto di distanza tra due edit per poi estenderlo ad un edit e un autore. La distanza relativa tra due edit appartenenti allo stesso pad è banalmente la differenza tra le loro posizioni. Se però i due edit non appartengono allo stesso pad, la loro distanza relativa verrà considerata come infinita, per catturare l'idea di non collaborazione. Formalmente:

Definizione 3.6 (Distanza temporale relativa tra due edit). Dati due edit e1 ed e2, la loro distanza temporale è definita come:

$$\begin{gathered} relativetimedistance(e1, e2) = |position(e1) - position(e2)| \; se \; pad(e1) = pad(e2) \\ relativetimedistance(e1, e2) = \infty \; altrimenti \end{gathered} \tag{3.6}$$

È dunque possibile ragionare in maniera analoga alla definizione di 3.5 per definire la distanza temporale tra un edit e ed un autore a, considerando la minima distanza temporale tra e ed un qualsiasi edit e' che (a) non sia avvenuto dopo di e, (b) appartenga all'autore a e (c) appartenga allo stesso pad di e. Formalmente:

Definizione 3.7 (Distanza temporale tra un edit ed un autore). Dato un edit e1 ed un autore a, la loro distanza temporale è definita come:

$$\begin{gathered} relativetimedistance(e1, a) = min(\infty, relativetimedistance(e1, e2)), \\ \forall e2 \mid time(e1) \geq time(e2) \\ \wedge \; author(e2) = a \\ \wedge \; pad(e1) = pad(e2)) \end{gathered} \tag{3.7}$$

Si consideri di nuovo l'immagine 3.3 come esempio. In questo caso:

- timedistance(1,4) = 754 come visto in precedenza.
- relativetimedistance(1,4) = | 1 - 4 | = 3 seguendo banalmente le definizioni.
- relativetimedistance(3,A) = min(relativetimedistance(3,2), relativetimedistance(3,1)) = min(| 3 - 2 |, | 3 - 1 |) = 1 poichè in questo caso vengono considerati entrambi gli edit e viene presa la distanza minore.
- relativetimedistance(3,B) = 0 banalmente perché l'edit 3 è di B.
- relativetimedistance(3,C) = ∞ perché, come nel caso assoluto, C non ha edit che non superino l'edit 3.

La tabella 3.1 riassume i valori delle due distanze temporali.

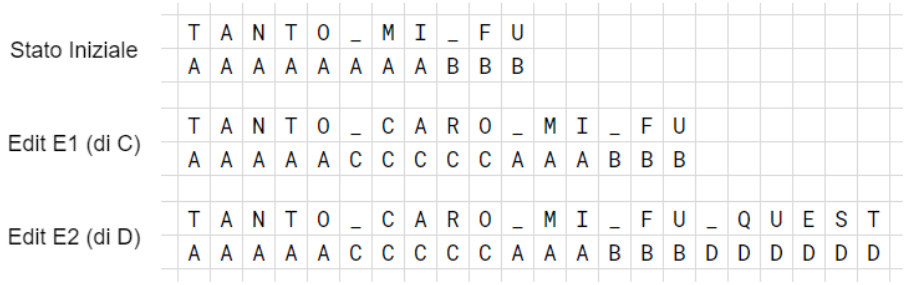


FIGURA 3.6: Questa figura mostra una successione di edit applicata a dei caratteri, e la mappatura presente tra i caratteri e gli autori.

3.4.3 Distanza collaborativa spaziale

La distanza spaziale vuole trasformare l'idea della collaborazione spaziale da "è abbastanza vicino?" (spazialmente) in "quanto è vicino?" (spazialmente). Un edit avrà una distanza spaziale bassa con gli autori che hanno lavorato su aree del pad vicine a quelle su cui ha effetto, mentre avrà una distanza molto più alta con autori che hanno lavorato su parti lontane del documento.

Analogamente alle altre distanze, si partirà dalla distanza tra due edit per poi passare alla definizione della distanza tra un edit ed un autore.

Per potere misurare la distanza spaziale tra i due edit, è necessario avere una mappatura che colleghi ogni carattere all'edit che ne è direttamente responsabile. Viene quindi ripreso il concetto di lastedit utilizzato per 2.

Siano e1, e2 due edit appartenenti allo stesso pad P. Ora è possibile definire la loro distanza spaziale come la minima distanza testuale tra un carattere appartenente ad e1 ed un carattere appartenente ad e2 nel momento in cui il più vecchio tra i due edit viene applicato. Formalmente:

Definizione 3.8 (Distanza temporale tra due edit). Dati due edit e1 ed e2, la loro distanza spaziale è definita nel momento max(time(e1), time(e2)) come:

$$\begin{gathered} spacedistance(e1, e2) = min(|p1 - p2|, lastedit(pad(e1), p1) = e1 \\ \wedge\, lastedit(pad(e2), p2) = e2)\ se\ pad(e1) = pad(e2) \\ spacedistance(e1, e2) = \infty\ altrimenti \end{gathered} \tag{3.8}$$

A questo punto in maniera identica a quella utilizzata per definire 3.5 è possibile definire la distanza spaziale tra un edit e ed un autore a, considerando la minima distanza spaziale tra e ed un qualsiasi edit e' che (a) non sia avvenuto dopo di e, (b) appartenga all'autore a e (c) appartenga allo stesso pad di e. Formalmente:

Definizione 3.9 (Distanza spaziale tra un edit ed un autore). Dato un edit e1 ed un autore a, la loro distanza spaziale è definita come:

$$\begin{gathered} spacedistance(e1, a) = min(\infty, spacedistance(e1, e2)), \\ \forall e2 \mid time(e1) \geq time(e2) \\ \wedge\, author(e2) = a \\ \wedge\, pad(e1) = pad(e2)) \end{gathered} \tag{3.9}$$

Per esempio, si consideri l'immagine 3.6. L'immagine rappresenta tre stati diversi del pad: lo stato a cui era ad un certo punto, lo stato subito successivo dopo

l'applicazione dell'edit E1 e lo stato subito dopo a quest'ultimo, dopo l'applicazione dell'edit E2.

- space(e2,e1) = | 17 - 10 | perchè il carattere toccato dall'edit più vicino di e2 è quello in posizione 17 e quello più vicino per e1 è in posizione 10
- space(e2,d) = 0 perché a è responsabile di quell'edit
- space(e2,b) = | 18 - 17 | = 1 perché è la distanza tra il carattere di b più vicino a e2 (nella posizione 17) e il carattere di e2 più vicino a quel carattere (nella posizione 18)
- space(e1,d) = ∞ perchè d in quel momento non è responsabile di nessun carattere nella stringa.

3.4.4 Distanza spaziotemporale

È opportuno precisare il perché mentre la collaborazione spaziale e temporale vengano affiancate da concetti simili (distanza spaziale e temporale), non venga introdotto un concetto di distanza spaziotemporale.

La collaborazione spaziotemporale introduce un nuovo concetto che considera sia la temporalità che la spazialità allo stesso tempo, i.e., un edit collaborativo spazialmente e temporalmente, può non esserlo spazio-temporalmente. Invece, un'eventuale distanza spazio-temporale dovrebbe necessariamente consistere nel semplice affiancamento della distanza spaziale e di quella temporale assoluta, poiché non avrebbe senso combinare i due in quanto aventi dimensioni diverse (uno è misurato in numero di caratteri e l'altro in secondi).

A questo punto, siccome il nuovo concetto non offrirebbe nessuna nuova prospettiva, è stato preferito non introdurlo.

3.4.5 Definizione della funzione di mappatura

A questo punto, per concludere il discorso del metodo, non rimane altro che discutere della funzione di mappatura utilizzata per la distanza spaziale. Questa funzione prende in input un numero che identifica la posizione di un carattere nel documento di testo in un determinato momento, e restituisce un edit associato a quel carattere.

Per definire questa assegnazione per ogni stato del documento si utilizza la ricorsione:
Passo 0: Ogni eventuale carattere già presente nel documento di testo è assegnato ad E0 (Fig. 3.7)
Passo N+1: Si parte dall'assegnazione al passo N, e si eseguono i seguenti passi a seconda dell'opCode dell'edit:

- INS: i caratteri inseriti appartengono a E N+1 (Fig. 3.8)
- DEL: i caratteri vengono eliminati insieme alle assegnazioni (Fig. 3.9)
- UPD: i caratteri modificati appartengono a E N+1 (Fig. 3.10)

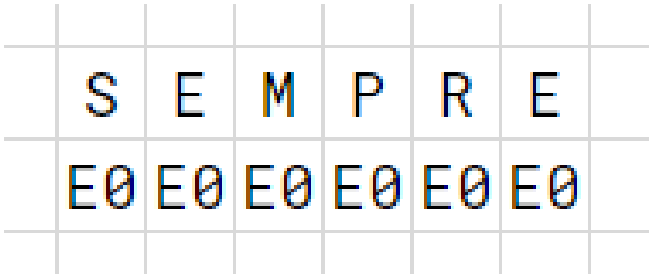


FIGURA 3.7: Esempio di stato iniziale della funzione di mappatura.

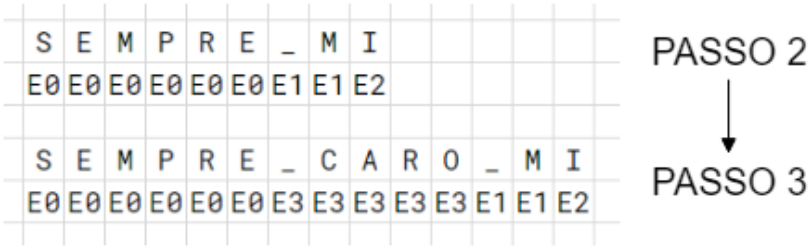


FIGURA 3.8: Cambiamento della funzione di mappatura con un operazione INS.

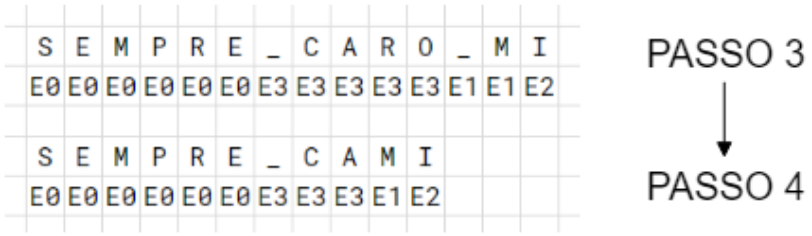


FIGURA 3.9: Cambiamento della funzione di mappatura con un operazione DEL.

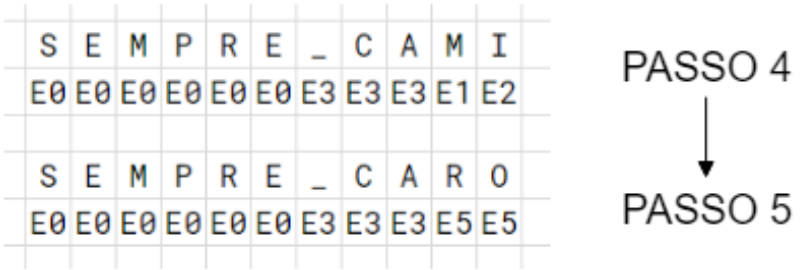


FIGURA 3.10: Cambiamento della funzione di mappatura con un operazione UPD.

Capitolo 4

Applicazione del modello

Una volta definito il modello, questo è stato applicato definendo un metodo per l'analisi dei documenti di testo. Questo capitolo presenta la struttura dei dati prodotti in output e varie particolarità osservabili su di essi.

4.1 Le matrici delle distanze

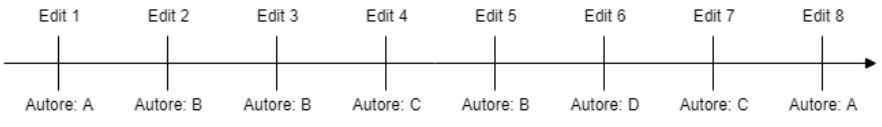
A partire da ogni file, l'analisi produce una matrice autore per edit (fig. 4.1) per ognuna delle tre distanze collaborative. Per ogni matrice:

- Ogni cella contiene la distanza di un determinato edit rispetto ad un preciso autore.
- Ogni riga contiene la distanza di ogni edit rispetto ad un preciso autore.
- Ogni colonna contiene la distanza di un determinato edit rispetto ad ogni autore.

Le sezioni successive presentano i tre diversi tipi di matrice, osservando alcune delle loro peculiarità.

4.2 Analisi della distanza temporale relativa

La figura 4.1 mostra un esempio di matrice finale per la distanza temporale relativa. Ogni cella misura il numero di edit tra quello più vicino di ogni autore e quello



Edit	Edit 1 (A)	Edit 2 (B)	Edit 3 (B)	Edit 4 (C)	Edit 5 (B)	Edit 6 (D)	Edit 7 (C)	Edit 8 (A)
A	0	1	2	3	4	5	6	0
B	∞	0	0	1	0	1	2	3
C	∞	∞	∞	0	1	2	0	1
D	∞	∞	∞	∞	∞	0	1	2

FIGURA 4.1: Esempio di matrice finale per la distanza temporale relativa

	18:12:40	18:12:50	18:12:55	18:12:55
	Edit 1 (A)	Edit 2 (B)	Edit 3 (B)	Edit 4 (C)
A	0	10	15	15
B	∞	0	0	0
C	∞	∞	∞	0

FIGURA 4.2: Esempio di matrice finale per la distanza temporale assoluta

attuale. Per esempio, la cella evidenziata indica che l'edit 7 ha distanza temporale relativa pari a 6 edit dall'autore A, i.e., il settimo edit è stato effettuato sei edit dopo l'edit più vicino effettuato da A.

Si noti anche come:

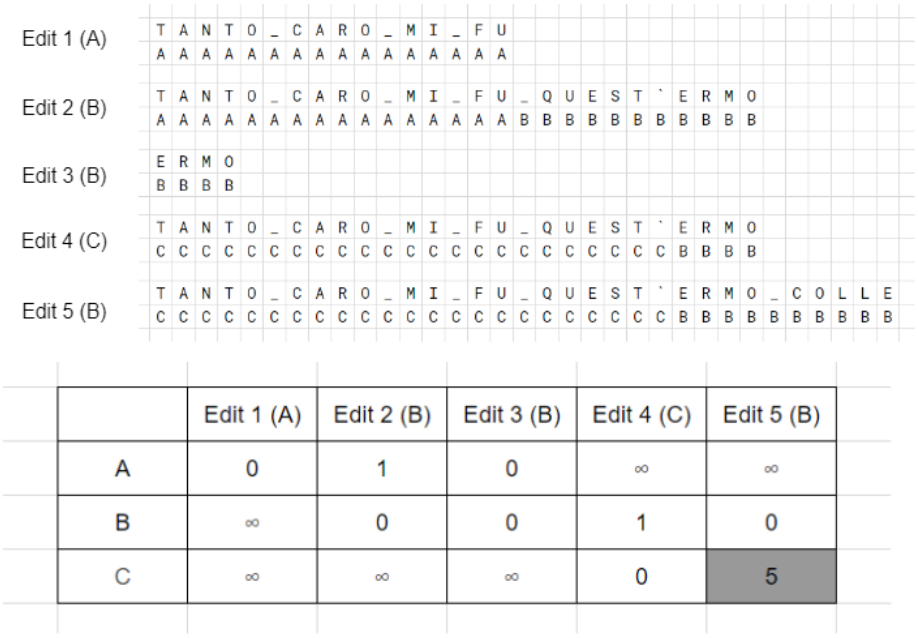
- L'autore che effettua l'edit n ha sempre distanza 0 da esso. Questa è una banale conseguenza della definizione della distanza temporale relativa. Inoltre, siccome due edit non possono essere applicati contemporaneamente, se un edit ha distanza 0 rispetto ad un autore, significa che è stato effettuato da quell'autore.
- La distanza di un autore da un edit è infinita se questo autore non ha ancora effettuato edit. Nel caso della distanza temporale relativa, questo è l'unico modo con il quale può apparire un valore infinito.
- Nel caso della distanza temporale relativa in ogni riga il valore di ogni cella dopo la prima può solo essere 0 o il valore della cella precedente aumentato di uno. Questo perché, o l'edit appartiene all'autore di quella riga e quindi il valore è 0, oppure è di un altro autore e quindi il numero di edit di distanza è lo stesso con l'aggiunta di uno (quello precedente che prima non era stato contato).

4.3 Analisi della distanza temporale assoluta

La figura 4.2 mostra un esempio di matrice finale per la distanza temporale assoluta. Ogni cella misura il numero di secondi tra l'edit più vicino di ogni autore e quello attuale. Per esempio, la cella evidenziata indica che l'edit 3 ha distanza temporale assoluta pari a 15 secondi dall'autore A, i.e., il terzo edit è stato effettuato 15 secondi dopo l'edit più vicino effettuato da A.

Si noti anche come:

- L'autore che effettua l'edit n ha sempre distanza 0 da esso. Come per la distanza temporale relativa, questa è una banale conseguenza della definizione. Tuttavia, siccome due edit possono accadere nello stesso secondo, un edit può avere distanza 0 da più autori. Pertanto, non è detto che un edit con distanza 0 da un autore sia stato effettuato da quell'autore.



	Edit 1 (A)	Edit 2 (B)	Edit 3 (B)	Edit 4 (C)	Edit 5 (B)
A	0	1	0	∞	∞
B	∞	0	0	1	0
C	∞	∞	∞	0	5

FIGURA 4.3: Esempio di matrice finale per la distanza spaziale

- La distanza di un autore da un edit è infinita se questo autore non ha ancora effettuato edit. Come per la distanza temporale relativa, anche in questo caso è l'unico modo con il quale può apparire un valore infinito.
- Nel caso della distanza temporale assoluta in ogni riga il valore di ogni cella dopo la prima può solo essere 0 o maggiore o uguale del valore della cella precedente. Questo perché, o l'edit appartiene all'autore di quella riga e quindi il valore è 0, oppure è di un altro autore e quindi il tempo passato è per forza almeno quello indicato nell'edit precedente (perchè è necessariamente avvenuto prima).

4.4 Analisi della distanza spaziale

La figura 4.3 mostra un esempio di matrice finale per la distanza spaziale. Ogni cella misura il numero di secondi tra l'edit più vicino di ogni autore e quello attuale. Per esempio, la cella evidenziata indica che l'edit 5 ha distanza spaziale pari a 5 secondi dall'autore C, i.e., il quinto edit è a 5 caratteri di distanza dall'edit più vicino effettuato da C.

Si noti anche come:

- L'autore che effettua l'edit n ha sempre distanza 0 da esso. Come per la distanza temporale relativa, questa è una banale conseguenza della definizione. Tuttavia, siccome un edit può coinvolgere caratteri associati ad altri autore (per esempio, cancellando dei caratteri), un edit può avere distanza 0 da più autori. Pertanto, non è detto che un edit con distanza 0 da un autore sia stato effettuato da quell'autore.
- La distanza di un autore da un edit è infinita se questo autore non ha ancora effettuato edit o se non rimangono più caratteri associati a quell'autore nel documento.

Capitolo 5

Implementazione del modello

In questo capitolo si discuterá di alcuni dettagli riguardanti le implementazioni del modello discusso precedentemente. A partire da esso infatti sono stati realizzati due script Python [1], uno per l'analisi delle distanze collaborative spaziali e delle distanze temporali relative, e uno per l'analisi delle distanze collaborative temporali assolute.

5.1 Analizzatore delle distanze collaborative spaziali e temporali relative

Il primo programma utilizza due classi principali per analizzare le distanze collaborative spaziali e temporali relative.

5.1.1 La classe matrice autore-edit

La prima classe rappresenta una matrice autore per edit in cui ogni cella contiene le distanze collaborative, e supporta come azioni principali l'aggiunta o la ricerca di un nuovo autore e l'inserimento di una distanza per un autore già esistente. È realizzata banalmente tramite delle liste: una lista contiene ogni struttura dati rappresentante un autore, che a sua volta contiene una lista contenente ogni distanza autore-edit associata a quell'autore (Fig. 5.1). Le operazioni di ricerca e inserimento sono quindi delle semplici operazioni su liste.

5.1.2 La classe mappatura carattere-autore

La seconda classe invece rappresenta la funzione di mappatura che associa ad ogni carattere del documento di testo l'autore che ne è responsabile.

Si prenda la fig. 5.2 come esempio: la classe funzione di mappatura associa al documento di testo una stringa, che mappa ad ogni carattere del documento in un determinato momento un autore, per essere in grado di effettuare il calcolo della distanza spaziale. Si noti che a differenza del modello presentato nella parte del metodo la stringa non mantiene l'informazione di quale edit abbia effettuato la modifica che ha coinvolto il carattere, ma solo di quale autore abbia effettuato quell'edit; questo per velocizzare l'operazione di calcolo eliminando informazioni non necessarie.

Per far questo, la classe ha tre funzioni che aggiornano la stringa a seconda dell'edit effettuato e una funzione che misura la minima distanza di un carattere associato ad un determinato autore da una determinata area della stringa.

Le tre funzioni aggiornano la stringa in questo modo:

[1] Il codice è liberamente disponibile alla seguente pagina: `https://github.com/Lucalta/EtherPad-Collaboration-Analizer`

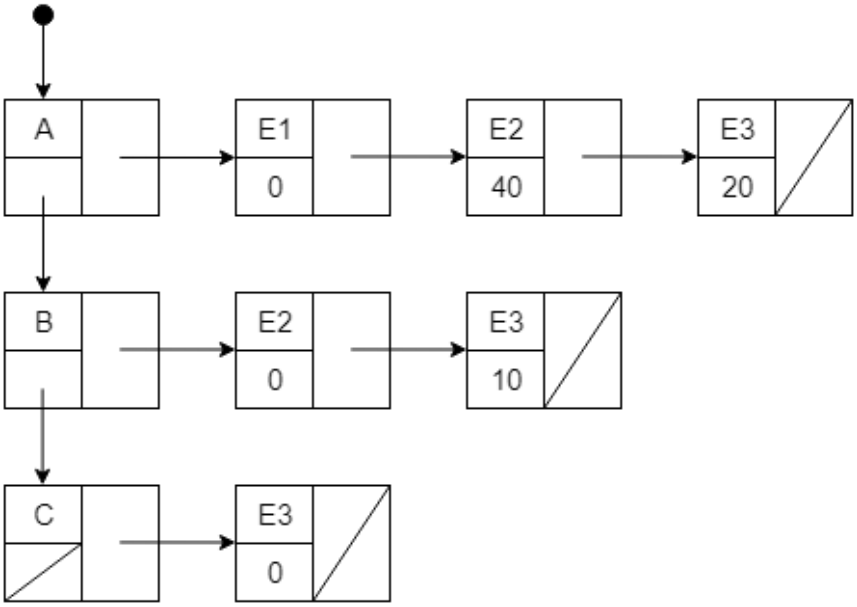


FIGURA 5.1: Schema della struttura dati che rappresenta la matrice autore-edit.

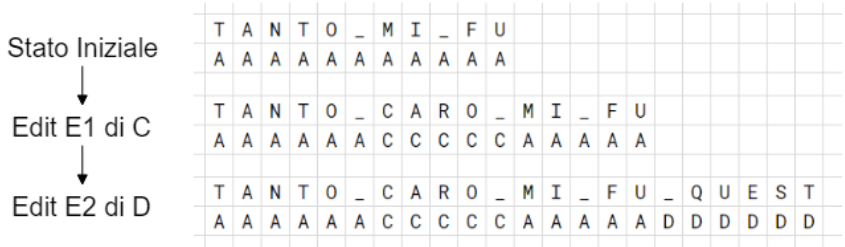


FIGURA 5.2: Esempio di un'associazione carattere/autore della funzione di mappatura.

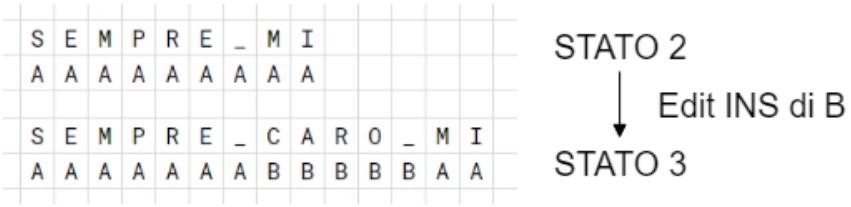


FIGURA 5.3: Cambiamento della classe funzione di mappatura con un operazione INS.

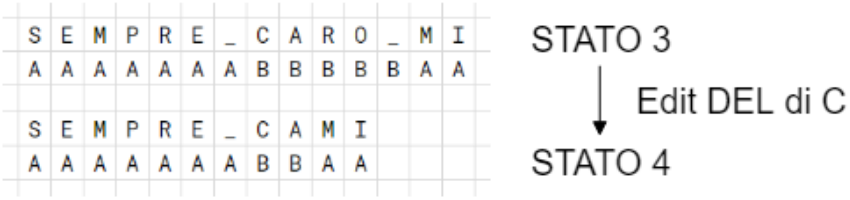


FIGURA 5.4: Cambiamento della classe funzione di mappatura con un operazione DEL.

- INS: i caratteri inseriti appartengono all'autore dell'edit (Fig. 5.3)
- DEL: i caratteri vengono eliminati insieme alle assegnazioni (Fig. 5.4)
- UPD: i caratteri modificati appartengono all'autore dell'edit (Fig. 5.5)

La funzione che misura la distanza collaborativa si limita invece a partire dai due estremi del preintervallo dell'edit, e a contare il minimo numero di caratteri incontrati prima di trovare quello dell'autore scelto.

5.1.3 Calcolo della distanza spaziale

Una volta implementate queste due classi, per riempire la prima matrice autore per edit con le distanze spaziali è solo necessario caricare ogni edit su una lista, ordinarla per revisione/timestamp, e poi elaborare ogni edit nel seguente modo:

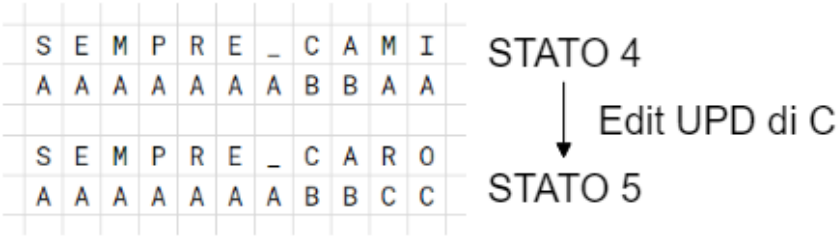


FIGURA 5.5: Cambiamento della classe funzione di mappatura con un operazione UPD.

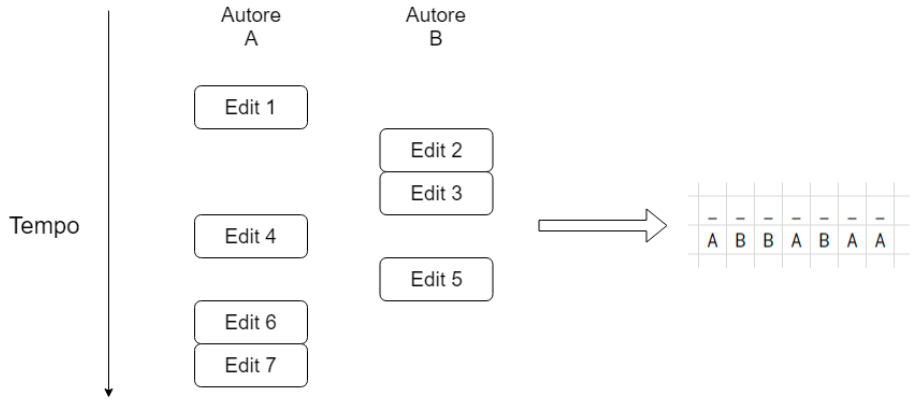


FIGURA 5.6: Esempio di rappresentazione della distanza temporale relativa come distanza spaziale.

1. Per ogni autore, si calcola la distanza autore-edit con l'apposita classe.
2. Si utilizza la funzione più consona della classe per aggiornare la stringa.

5.1.4 Calcolo della distanza temporale relativa

Per il calcolo delle distanze temporali relative, si pùo notare come si possa rappresentare la successione di edit come una stringa a cui si aggiunge solo in coda.

È quindi possibile calcolare le distanze temporale relative allo stesso modo di quelle spaziali, fingendo che ogni edit inserisca un carattere qualsiasi in fondo al documento di testo. (fig. 5.6)

5.2 Analizzatore delle distanze collaborative temporali assolute

Il secondo programma analizza le distanze collaborative temporali assolute. Il funzionamento è molto simile all'altro programma: è sempre presente la classe rappresentante la matrice autore per edit, e gli edit vengono analizzati in maniera quasi identica. L'unica differenza è che per memorizzare le distanze invece che utilizzare una classe apposita il programma salva in una semplice struttura dati l'ultimo timestamp incontrato per ogni autore. A quel punto per avere le distanze basta solamente calcolare la differenza tra il timestamp attuale e quello di ogni altro autore.

Capitolo 6

Esperimento

In questo capitolo si parlerà delle osservazioni risultanti dall'analisi dei dati ottenuti eseguendo il programma descritto nel capitolo precedente, con l'intento di:

- Notare eventuali regolarità nei dati.
- Esaminare la potenzialità delle finestre proposte da D'Angelo et al. in [6].
- Proporre delle nuove finestre che potrebbero risultare interessanti se prese in esame.

6.1 Introduzione

Una volta realizzati i programmi per la misura delle distanze, sono stati presi in input gli stessi file utilizzati nel lavoro di D'Angelo et al., ottenuti da `https://etherpad.wikimedia.org`. La tabella 6.1 ricapitola il numero di pad ed edit che sono stati esaminati. È inoltre importante notare come questi documenti di testo siano documenti pubblici non realizzati appositamente per l'esperimento descritto nell'articolo. I vari script hanno quindi generato per ogni file le tre matrici delle distanze (discusse nel capitolo 4).

A questo punto, per ogni file sono state prese tutte le matrici, e da ognuna sono state estratte le distanze collaborative (escludendo quelle di un autore da sé stesso e quelle degli autori che non hanno ancora lavorato sul documento, in modo tale da seguire completamente il modello del lavoro precedente), a partire dalle quali sono state calcolate:

- La media e la deviazione standard, per avere un'idea della distribuzione degli edit.
- Il primo, secondo e terzo quartile, utilizzati poi per giustificare i valori delle finestre scelte.

La figura 6.1 mostra l'andamento della media e della deviazione standard nei vari file. Ogni punto dell'asse orizzontale rappresenta un diverso file, mentre l'asse verticale rappresenta il valore della media o della deviazione standard, misurate in numero di edit, secondi o caratteri a seconda del tipo di distanza.

TABELLA 6.1: Numero di Pad ed Edit analizzati

Pad analizzati	13871
Edit analizzati	6065234

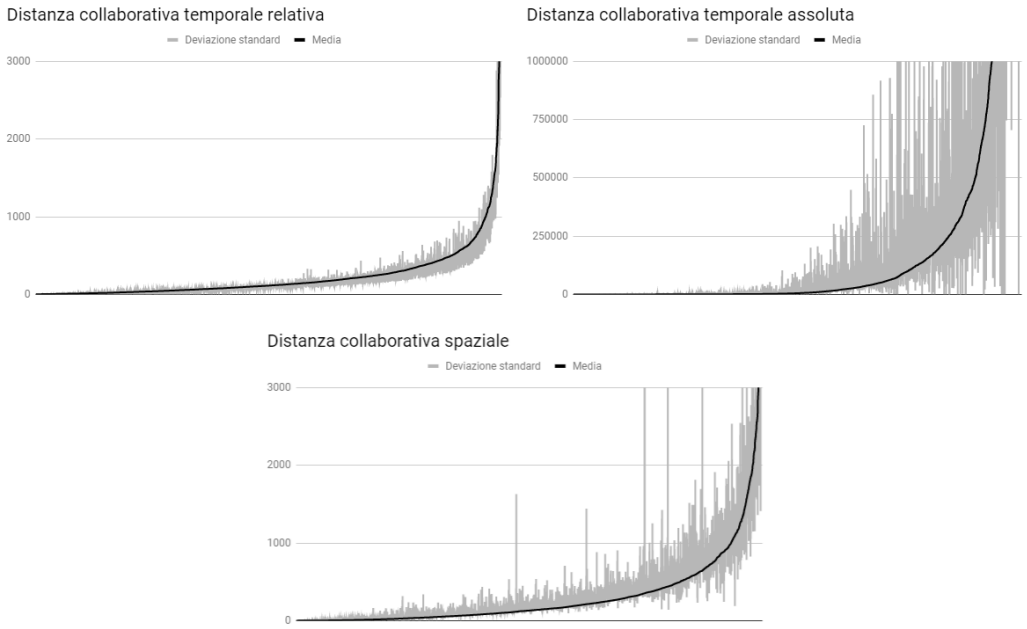


FIGURA 6.1: Le medie e le deviazioni standard di tutti i file esaminati. Ogni punto dell'asse orizzontale rappresenta un diverso file, mentre l'asse verticale rappresenta il valore della media o della deviazione standard, misurate in numero di edit, secondi o caratteri a seconda del tipo di distanza.

6.2 Analisi delle finestre spaziali e temporali

6.2.1 Parametri per la ricerca delle finestre

L'idea di base dietro alla ricerca delle finestre spaziali e temporali è quella di cercare dei valori tali da permettere di osservare una percentuale di collaborazione significativa. Si vuole quindi evitare di utilizzare finestre sia troppo grandi che troppo piccole, siccome in entrambi i casi la collaborazione risulterebbe appiattita verso uno dei due estremi (o completamente assente, o presente in ogni modifica).

Prima di passare a parlare in dettaglio dell'analisi effettuata sulle finestre spaziali e temporali, è necessario descrivere tre nuovi insiemi considerati per ogni tipo di distanza, chiamati Q1, Q2 e Q3. Questi tre insiemi contengono rispettivamente il primo, secondo e terzo quartile di ogni file, e.g., un valore contenuto nel Q2 della

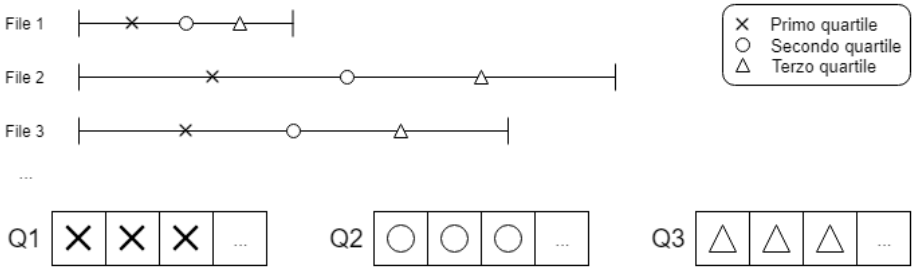


FIGURA 6.2: Esempio dei valori contenuti in Q1, Q2 e Q3.

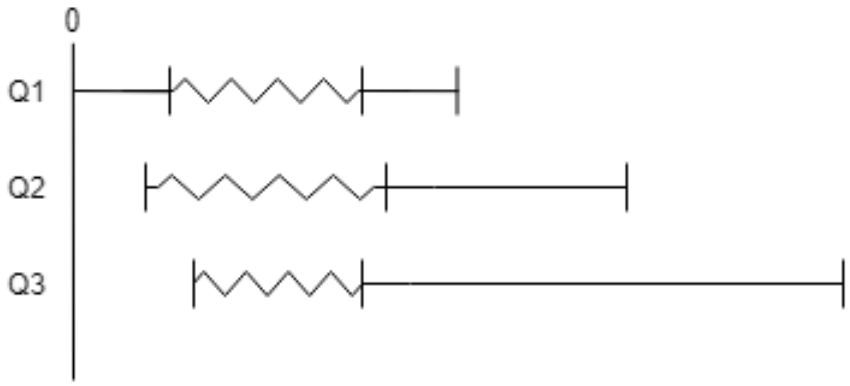


FIGURA 6.3: Gli intervalli scelti per definire le finestre. L'immagine mostra la situazione che intendono catturare.

distanza collaborativa spaziale sarà pari al valore del secondo quartile delle distanze collaborative spaziali in uno dei file esaminati (fig. 6.2).

I tre insiemi appena descritti potranno quindi essere utilizzati per la ricerca delle finestre. Q1 conterrà un valore per ogni file, corrispondente al suo primo quartile, e si potrà usare per giustificare la scelta di finestre molto piccole e vicine al valore minimo, mostrando come molti file abbiano comunque un ingente numero di distanze sotto ad una certa soglia. Insieme a Q2 e Q3 invece, potrà essere usato in maniera opposta per cercare una soglia massima per la scelta delle finestre. Un eventuale insieme Q4 non viene considerato così come non viene calcolato il quarto quartile di ogni file poichè banalmente corrispondenti al più alto valore che compare.

Pertanto, partendo da queste considerazioni, sono stati scelti i seguenti intervalli indicativi entro i quali dovranno essere compresi i valori delle finestre. (fig. 6.3):

- Tra il 25% e il 75% dei valori in Q1
- Tra lo 0% e il 50% dei valori in Q2
- Tra lo 0% e il 25% dei valori in Q3

A questo punto, si può cominciare con l'analisi delle finestre.

6.2.2 Finestre spaziali

Si parta considerando le finestre spaziali. L'articolo prende in considerazione quattro finestre, da 10, da 80, da 400 e da 800 caratteri. Utilizzando Q3 si possono velocemente scartare le finestre da 400 e 800 perché estremamente grandi. Infatti, per circa il 60% dei file, più di tre quarti di tutte le distanze è inferiore al 400 (fig. 6.4). Questo significa che, prendendo come finestra 400, in più della metà dei file avrei più del 75% degli edit marcati come collaborativi.

La finestra da 80 caratteri risulta sempre piuttosto grande poiché include più della metà delle distanze per più del 50% dei file (fig. 6.5), tuttavia può avere senso come valore più alto.

Quindi, una finestra massima da 80 caratteri sembra risultare adeguata, rimane solamente da cercare un valore minimo. A tal proposito, si può notare come 10 sia un valore troppo elevato per consistere nella minima finestra analizzata. Infatti,

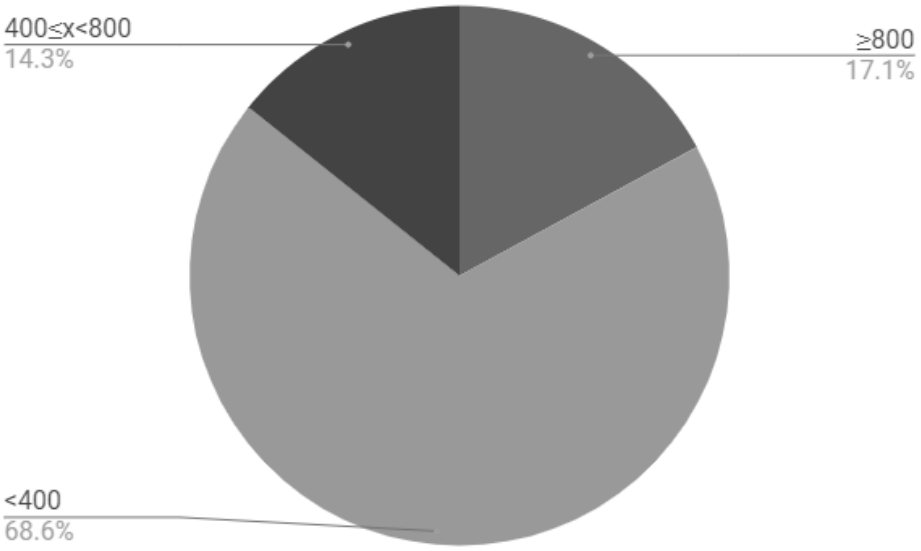


FIGURA 6.4: Distribuzione dei valori del Q3 delle distanze spaziali. Il 68.6% dei suoi valori è inferiore a 400.

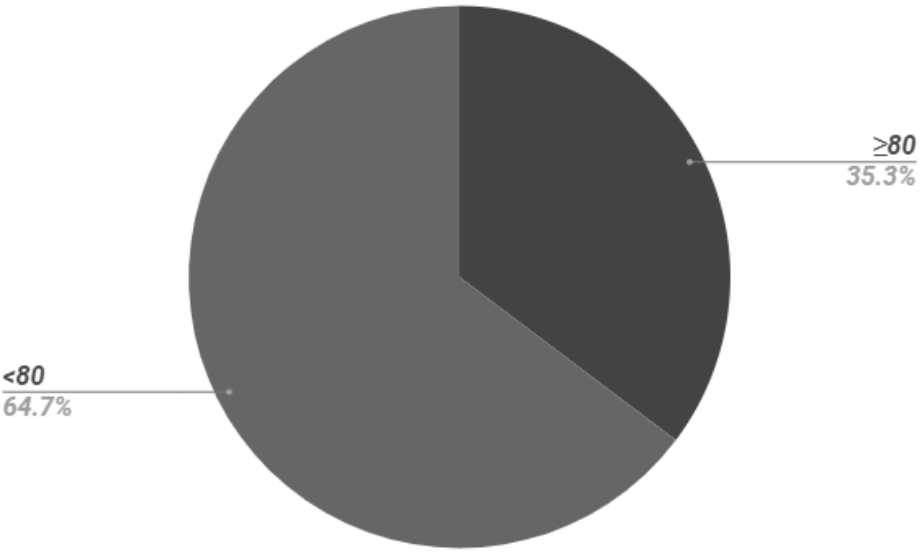


FIGURA 6.5: Distribuzione dei valori del Q2 delle distanze spaziali. Il 64.7% dei suoi valori è inferiore a 80.

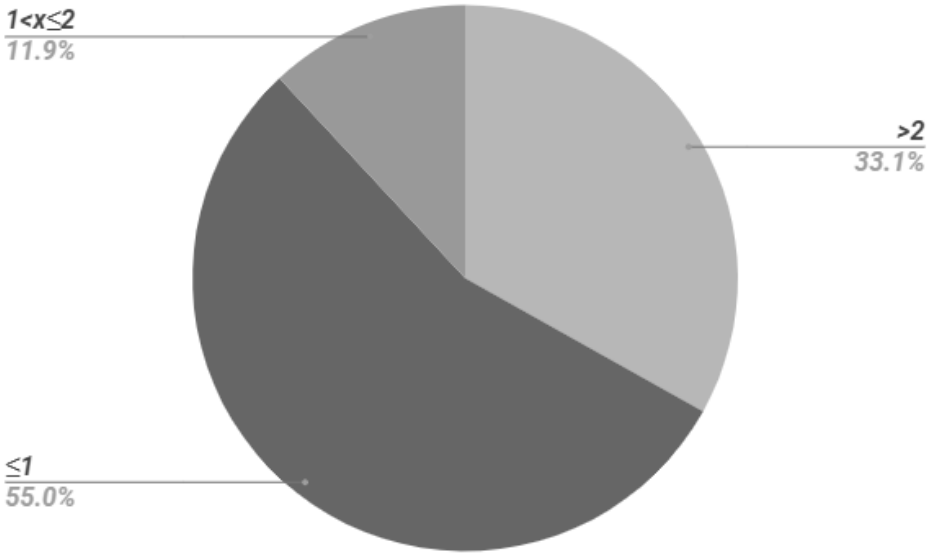


FIGURA 6.6: Distribuzione dei valori del Q1 delle distanze spaziali. Il 66.9% dei suoi valori è inferiore a 2.

osservando il Q1 (fig. 6.6) delle distanze spaziali nei vari file, si può notare come per più della metà di questi più di un quarto delle distanze risulta inferiore o uguale a 1. Può quindi aver senso partire da una finestra molto piccola, come 1 o 2 caratteri.

Pertanto:

- Tra le finestre spaziali prese dall'articolo, quelle da 10 e 80 caratteri sembrano le più interessanti, che quindi può avere senso riprendere.
- Difficilmente le finestre spaziali da 400 e 800 caratteri avranno risultati degni di nota.
- Una finestra spaziale da 1 o 2 caratteri può essere molto interessante.
- Può avere senso inserire un'ulteriore finestra spaziale da 5 caratteri per osservare l'andamento tra la prima finestra e quella da 10, così come una finestra da 40 caratteri per quello tra la finestra da 10 e quella da 80.

6.2.3 Finestre temporali

Si prendano in considerazione le finestre temporali. L'articolo prende in considerazione tre finestre, da 5, 10 e 60 secondi. A differenza delle finestre spaziali, quelle temporali sembrano molto più adeguate. Infatti, il Q1 delle distanze temporali assolute contiene valori molto più elevati rispetto a quello delle distanze spaziali. Mentre nel caso spaziale si aveva un valore inferiore a 2 caratteri nel 66% dei casi, in questo il valore è **superiore** a 60 secondi in più del 60% dei casi (Fig. 6.7). Si deve quindi ragionare in maniera diversa, cercando finestre temporali più alte.

Osservando il Q3 delle distanze temporali assolute (fig. 6.8), si nota come la finestra di 900 secondi scelta dai ricercatori di Google per i loro studi [15] può avere risultati interessanti. Può quindi avere senso considerare la finestra come massima e prenderne una intermedia tra quella e 60 secondi per avere un'idea dell'andamento.

Pertanto:

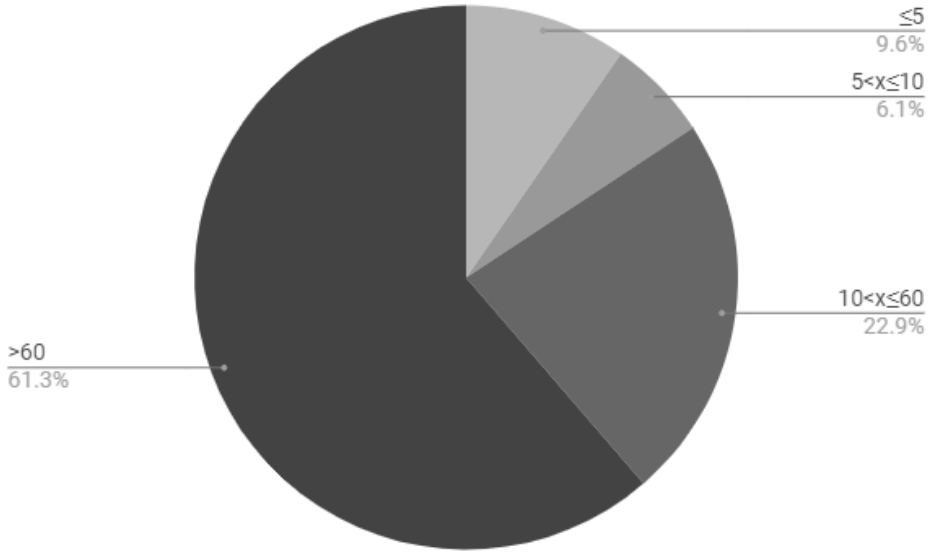


FIGURA 6.7: Distribuzione dei valori del Q1 delle distanze temporali assolute. Il 61,3% dei suoi valori è superiore a 60.

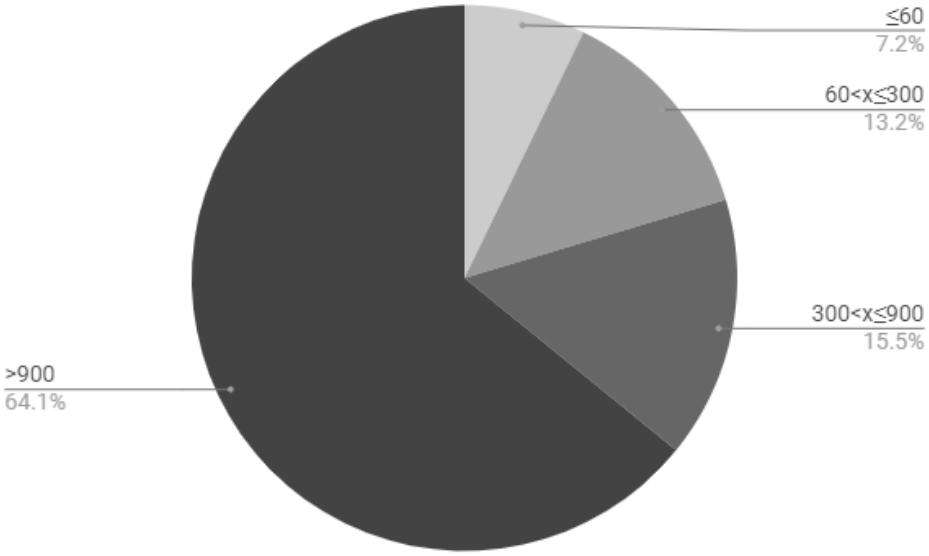


FIGURA 6.8: Distribuzione dei valori del Q3 delle distanze temporali assolute. Il 64,1% dei valori è superiore a 900.

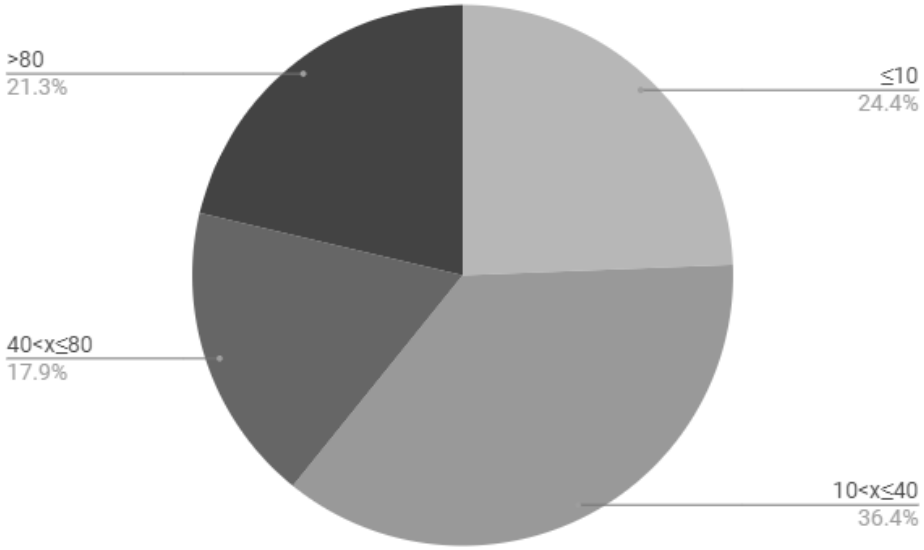


FIGURA 6.9: Distribuzione dei valori del Q1 delle delle distanze temporali relative. Circa il 25% dei valori è inferiore a 10.

- Tra le finestre temporali prese dall'articolo, quelle da 5, 10 e 60 secondi sembrano le più interessanti, che quindi può avere senso riprendere.
- Non ci dovrebbero essere differenze molto rilevanti tra la finestra da 5 secondi e quella da 10, può avere senso considerarne solo una.
- Una finestra spaziale sopra ai 60 secondi può essere molto interessante.
- Può avere senso considerare la stessa finestra da 900 secondi scelta dai ricercatori di Google in [15], e un eventuale finestra intermedia da 300 o 600 secondi.

6.2.4 Finestre temporali relative

Potrebbe essere interessante proporre delle finestre temporali relative attraverso l'analisi delle distanze temporali relative, in maniera analoga a quanto fatto negli altri casi. A tal proposito, osservando il Q1 delle distanze temporali relative (fig. 6.9) si può notare come una finestra minima adeguata potrebbe essere di 10 edit, poiché risulta superiore solo al 24,4% dei valori. Una finestra di 1 risulta estremamente piccola, così come una finestra minima di 5, superiore al 13,9% dei valori.

Un valore massimo potrebbe essere di 80 edit, che risulta inferiore circa al 30% dei terzi quartili di tutti i file (Fig. 6.10), mentre un eventuale valore intermedio potrebbe essere 40, posto a metà tra i due in entrambi i grafici.

Pertanto per la misura di un concetto di collaborazione temporale relativa, si potrebbero considerare delle finestre da 10, 40 e 80 edit.

6.3 Dipendenza tra i quartili

6.3.1 Distanze temporali assolute e spaziali

Osservando i dati, è interessante notare come per le distanze temporali assolute e quelle spaziali il valore del secondo e terzo quartile sia estremamente variabile con

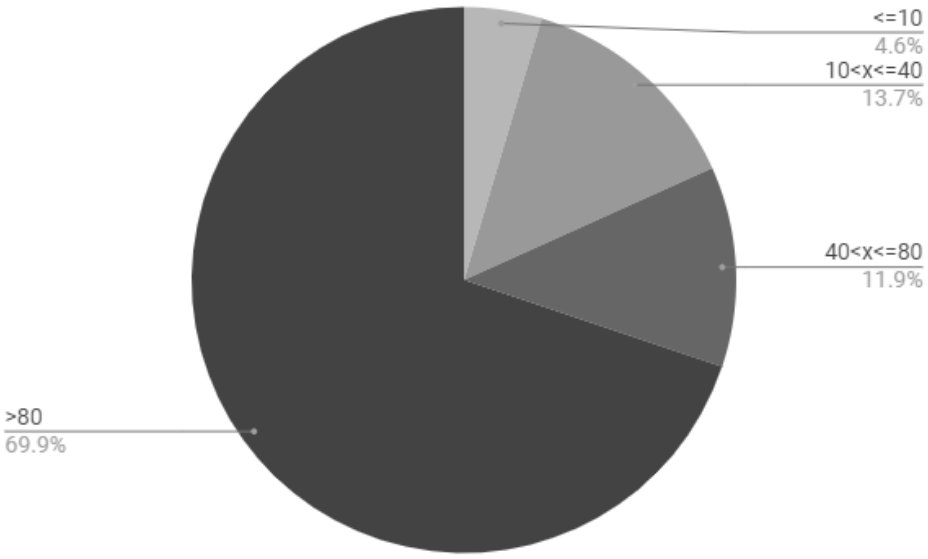


FIGURA 6.10: Distribuzione dei valori del Q3 delle distanze temporali relative. Il 30% dei valori è inferiore o uguale a 80.

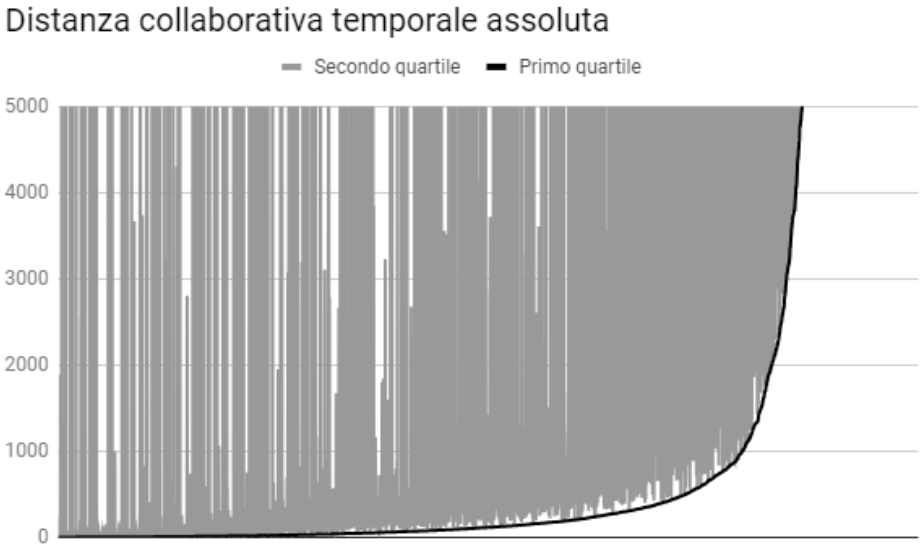


FIGURA 6.11: Valori del primo e secondo quartile delle distanze spaziali in ogni file, ordinati per il valore del primo quartile. Ogni punto dell'asse orizzontale rappresenta un diverso file analizzato.

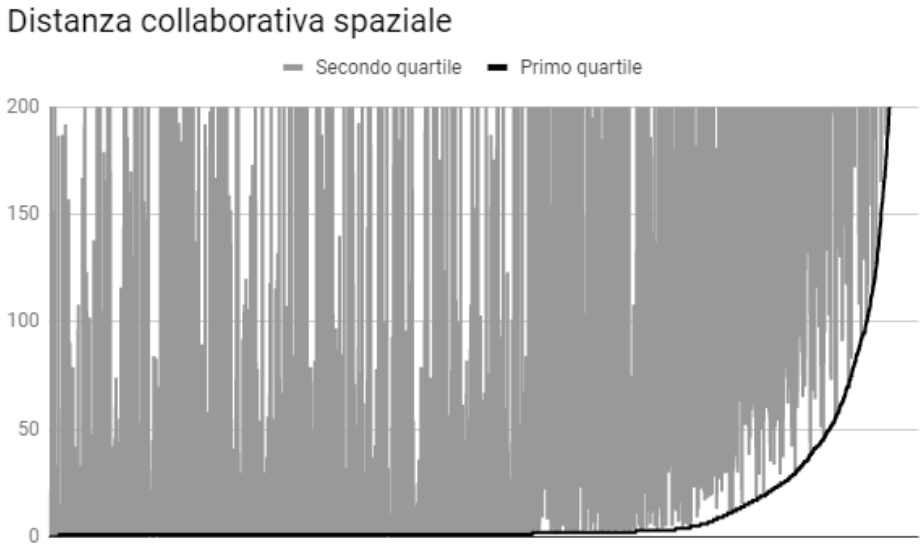


FIGURA 6.12: Valori del primo e secondo quartile delle distanze temporali assolute in ogni file, ordinati per il valore del primo quartile. Ogni punto dell'asse orizzontale rappresenta un diverso file analizzato.

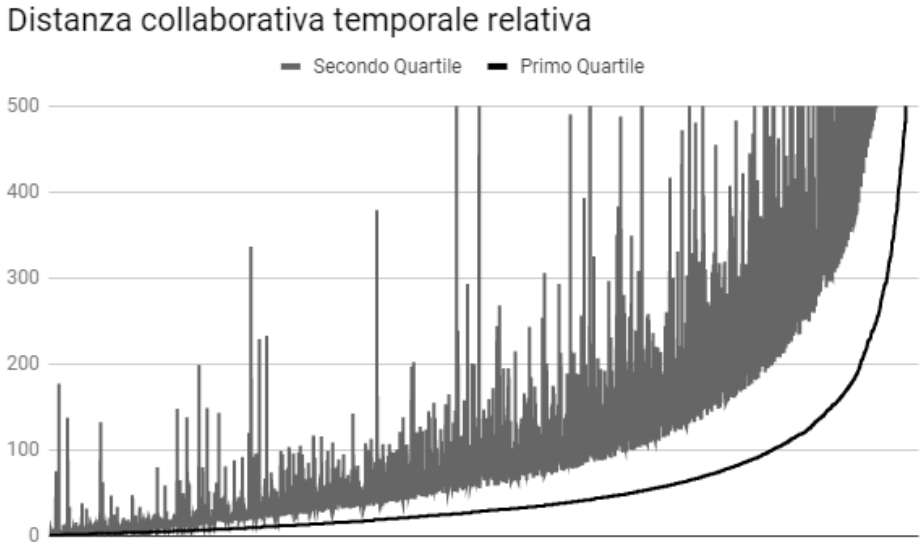


FIGURA 6.13: Valori del primo e secondo quartile delle distanze temporali relative in ogni file, ordinati per il valore del primo quartile. Ogni punto dell'asse orizzontale rappresenta un diverso file analizzato.

primi quartili simili (Fig. 6.11 e 6.12). Ciò significa che sarà molto difficile avere un'idea del valore del secondo e terzo quartile di un file osservandone solo il primo (oltre al banale fatto che questi saranno superiori ad esso), i.e., se le collaborazioni più strette in un documento di testo sono più vicine rispetto ad un altro, non è garantito che valga altrettanto per quelle più lontane.

6.3.2 Distanze temporali relative

Osservando invece i quartili delle distanze temporali relative, si può notare come ci sia una regolarità tra i loro valori (Fig. 6.13). Inversamente al caso precedente quindi, sembra che se le collaborazioni più strette in un documento di testo risultano più vicine rispetto ad un altro, quelle più lontane tendano ad essere più vicine a loro volta.

Capitolo 7

Conclusioni

In questo lavoro è stato descritto il modello definito in uno studio precedente [6], consistente nella caratterizzazione della collaborazione temporale, spaziale e spaziotemporale. Una volta descritto, questo modello è stato espanso con la definizione dei concetti di distanza spaziale e temporale assoluta. Subito dopo, è stato introdotto un nuovo concetto di distanza temporale relativa, utile per misurare la collaborazione in un documento di testo indipendentemente dall'estensione del periodo in cui vi si è lavorato. A questo punto, è stato fornito un esempio di applicazione del modello definendo il concetto di matrice delle distanze. È stato quindi implementato un programma capace di analizzare il registro delle modifiche di vari file di testo per ottenere le relative matrici delle distanze. Infine, le matrici delle distanze ottenute sono state esaminate, estraendo diversi dati, come la media delle distanze in ogni file, o altri dati utilizzati poi per osservare eventuali regolarità o cercare dei valori interessanti da utilizzare nell'applicazione del modello originale.

Per concludere, ecco alcune delle altre idee considerate, a partire dalle quali sarebbe possibile continuare questo studio:

- Potrebbe risultare interessante effettuare un'analisi identica a quella descritta negli ultimi capitoli a partire da dei documenti realizzati su Google Docs, per confrontare le differenze di utilizzo tra quella piattaforma ed Etherpad.
- Si potrebbe realizzare un'analisi a partire da documenti provenienti da autori di cui si conosce l'identità e il contesto in cui hanno lavorato, con un approccio che pone più enfasi sull'ambito relazionale, in maniera simile a quella del lavoro di Birnholtz et al. [10], consistente nell'analisi dell'uso di Google Docs da parte di 150 studenti, per cercare di capire quanto questi comunicassero durante la collaborazione su un documento.
- Si potrebbe riproporre lo studio effettuato da D'Angelo et al. in [6], utilizzando tuttavia le nuove finestre individuate con questo lavoro.

Bibliografia

[1] Tun Lu Bin Shao Du Li e Ning Gu. «An Operational Transformation Based Synchronization Protocol for Web 2.0 Applications». In: *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)* (2011), 563–572. URL: https://doi.org/10.1145/1958824.1958910.

[2] Cornelia Brodahl e Nils Kristian Hansen. «Education Students' Use of Collaborative Writing Tools in Collectively Reflective Essay Papers». In: *Journal of Information Technology Education: Research* 13 (2014), pp. 91–120. URL: http://www.jite.org/documents/Vol13/JITEv13ResearchP091-120Brodahl0463.pdf.

[3] Sam Kai Wah Chu e David M. Kennedy. «Using Online Collaborative Tools for Groups to Co-Construct Knowledge». In: *Online Information Review* 35.4 (apr. 2011), pp. 581–597. URL: https://doi.org/10.1108/14684521111161945.

[4] Olivia Fox Valerie L. Shalin Claudia-Lavinia Ignat Gérald Oster e François Charoy. «How Do User Groups Cope with Delay in Real-Time Collaborative Note Taking». In: *ECSCW 2015: Proceedings of the 14th European Conference on Computer Supported Cooperative Work, 19-23 September 2015, Oslo, Norway, Nina Boulus-Rødje, Gunnar Ellingsen, Tone Bratteteig, Margunn Aanestad, and Pernille Bjørn (Eds.)* (2015), pp. 223–242. URL: https://doi.org/10.1007/978-3-319-20499-4_12.

[5] Haodan Tan Dakuo Wang e Tun Lu. «Why Users Do Not Want to Write Together When They Are Writing Together: Users' Rationales for Today's Collaborative Writing Practices». In: *Proc. ACM Hum.-Comput. Interact* 3 (dic. 2017). URL: https://doi.org/10.1145/3134742.

[6] Gabriele D'Angelo, Angelo Di Iorio e Stefano Zacchiroli. «Spacetime Characterization of Real-Time Collaborative Editing». In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (nov. 2018), 41:1–41:19. ISSN: 2573-0142. DOI: 10.1145/3274310. URL: http://doi.acm.org/10.1145/3274310.

[7] Etherpad Foundation. «Etherpad». In: (2016). URL: http://etherpad.org/.

[8] Pascal Molli Gérald Oster Pascal Urso e Abdessamad Imine. «Data Consistency for P2P Collaborative Editing». In: *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW '06)* (2006), 259–268. URL: https://doi.org/10.1145/1180875.1180916.

[9] Google Inc. «Google Drive». In: (2016). URL: https://www.google.com/drive/.

[10] Stephanie B. Steinhardt Jeremy P. Birnholtz e Antonella Pavese. «Write here, write now!: an experimental study of group maintenance in collaborative writing». In: *ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13* (2013), pp. 961–970. URL: https://doi.org/10.1145/2470654.2466123.

[11] Gary M. Olson Judith S. Olson Dakuo Wang e Jingwen Zhang. «How People Write Together Now: Beginning the Investigation with Advanced Undergraduates in a Project Course». In: *ACM Trans. Comput.-Hum. Interact* 24.1 (mar. 2017), pp. 36–43. URL: https://doi.org/10.1145/3038919.

[12] Du Li e Rui Li. «A Performance Study of Group Editing Algorithms». In: *Proceedings of the 12th International Conference on Parallel and Distributed Systems - Volume 1 (ICPADS '06)* (2006), pp. 300–307. URL: https://doi.org/10.1007/978-3-319-20499-4_12.

[13] Gérald Oster Hyun-Gul Roh Mehdi Ahmed-Nacer Claudia-Lavinia Ignat e Pascal Urso. «Evaluating CRDTs for real-time document editing». In: *Proceedings of the 2011 ACM Symposium on Document Engineering* (2011), 103–112. URL: https://doi.org/10.1145/2034691.2034717.

[14] Judith S. Olson Ricardo Olenewa Gary M. Olson e Daniel M. Russell. «Now That We Can Write Simultaneously, How Do We Use That to Our Advantage?» In: *Commun. ACM* 60.8 (lug. 2017), pp. 36–43. URL: https://doi.org/10.1145/2983527.

[15] Makoto Uchida Nicolas Remy Yunting Sun Diane Lambert. «Collaboration in the Cloud at Google». In: *WebSci '14* (giu. 2014), pp. 239–240. URL: https://doi.org/10.1145/2615569.2615637.