UNIVERSITÀ DI TRENTO

Dipartimento di Ingegneria e Scienza dell'Informazione

Corso di Laurea in
Informatica

Elaborato finale

Titolo

Sottotitolo (alcune volte lungo - opzionale)

Supervisore
......

Laureando
Gandelli Alessio

Anno accademico .../...

Ringraziamenti

...thanks to...

Contents

Sommario **2**

1 Introduction **3**
1.1 Main Project 3
1.2 My Contribution 4
1.3 Related Work 4

2 Background **5**
2.1 History Exploration 5
2.2 Dataset 6
2.3 Definitions 7
2.4 Metrics 7
2.4.1 M 7
2.4.2 G 8

3 Methods **9**
3.1 Dataset 9
3.2 Approach 9

4 Results and Discussion **10**

5 Infrastacture **11**

6 Conclusions **12**

Bibliografia **12**

A Titolo primo allegato **14**
A.1 Titolo 14
A.1.1 Sottotitolo 14

B Titolo secondo allegato **15**
B.1 Titolo 15
B.1.1 Sottotitolo 15

Sommario

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

Sommario è un breve riassunto del lavoro svolto dove si descrive l'obiettivo, l'oggetto della tesi, le metodologie e le tecniche usate, i dati elaborati e la spiegazione delle conclusioni alle quali siete arrivati.

Il sommario dell'elaborato consiste al massimo di 3 pagine e deve contenere le seguenti informazioni:

- contesto e motivazioni
- breve riassunto del problema affrontato
- tecniche utilizzate e/o sviluppate
- risultati raggiunti, sottolineando il contributo personale del laureando/a

Chapter 1

Introduction

Wikipedia is the biggest source of information currently available on the internet, there are more than 6 million articles and they are all maintained by volunteers. The value of Wikipedia is all in the hands of the editors.

Many articles means many users and therefore many potential conflicts. Avoiding these conflicts is the best way for this encyclopedia to grow.
Each Wikipedia page has four different sections:

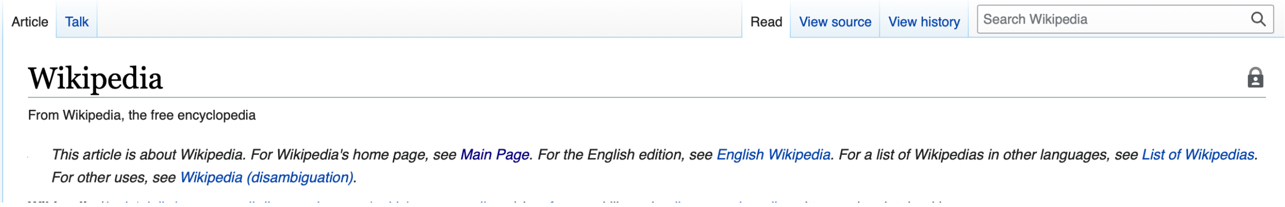
- Article: the actual content of the page.
- Talk Page: a forum where people can talk about edits.
- History: a place where everyone can see the older versions of the pages.
- Source: in this section users can edit the page.

Conflicts could happen both on the Talk page, through a discussion, and in the Article, through an edit war. It is valuable to analyze all of these aspects to get a well-rounded view of the problem.

1.1 Main Project

The project our team is working on, in collaboration with Eurecat and the Wikimedia Foundation, is named: “Community Health Metrics: Understanding Editor Drop-off“. this is an excerpt of the project idea:

“The primary value of Wikipedia is the editors. When an editor leaves the project, we lose their participation and contribution to the community, This could be related to multiple factors, also external to the project, but it could signal an issue related to internal dynamics and to the health of the community. While a big effort was dedicated to retain new editors, we lack knowledge and initiatives focused on understanding and preventing drop-off for experienced editors.”
As stated in the project description, the focus is on expert users, who are the core of Wikipedia: there are 41,741,926 Wikipedia accounts but active users are only 132,916, namely
3% of all users.



Wikipedia

From Wikipedia, the free encyclopedia

This article is about Wikipedia. For Wikipedia's home page, see Main Page. For the English edition, see English Wikipedia. For a list of Wikipedias in other languages, see List of Wikipedias. For other uses, see Wikipedia (disambiguation).

Figure 1.1: page structure

Focusing on this category of users and understanding the reasons that lead to a drop-off can give a big help to Wikipedia. Several people are working on this project, this work is just a part of the whole. In the team, everyone is working on a specific topic with the idea of then merging the different results to obtain an analysis of the phenomenon from different points of view in order to have a greater understanding of the life cycle of users.

The prevention of the drop-off is not the only goal of the project, improving the community health is also important to let users be in a good environment without being held back from editing.

1.2 My Contribution

The topic explored in this study is the revert analysis - i.e., when the version of a page is restored to that of a specific date - for all the articles of Wikipedia.

This project consisted of the analysis of the edit history of different language editions of Wikipedia to study patterns of reverts and edit wars to understand their potential effect on individual user activity.

We implemented state-of-the-art metrics of controversy based on reverts and mutual reverts and developed a new metric based on revert chains. Metrics have been computed per page and per user monthly.

The results can be viewed in an interactive dashboard available online.

1.3 Related Work

There are several works involving reverts: An interesting tool that allows visualizing conflicts is the one developed by Suh *et al.* [2]. The problem is that it is from 2007 but Wikipedia started to grow around 2010; now we have new technologies and much more data to analyze so more interesting conclusions can be reached. There have been analyses of antisocial behavior caused by vandalism [1], but since the focus of the project is on experienced users, this is not relevant to this study.

Chapter 2

Background

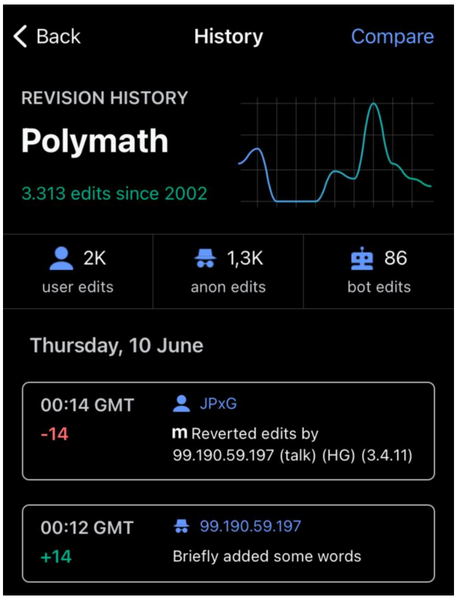
Everyone knows what is Wikipedia and how to read an article, but there are many features that most people are not aware of, *e.g.* see all the versions of a page and being able to edit it. Anyone with a browser and without much effort can see and compare all the edits in a Wikipedia page. For developers, there are many powerful resources such as big datasets containing a lot more information.

2.1 History Exploration

In the history section of a wikipedia article is possibile to see every version of the page. There are several tools anyone can use to explore the revision history:

- Mobile application: this resource is only available on mobile device and provides us some statistics about the edits of the page like the total number of revisions (Fig 2.1).

- Website: it is possible to compare two versions with an interactive tool that shows the progress of the modified page: each change corresponds to a bar indicating the number of bytes added or removed from the revision(Fig. 2.2).



REVISION HISTORY

Polymath

3.313 edits since 2002

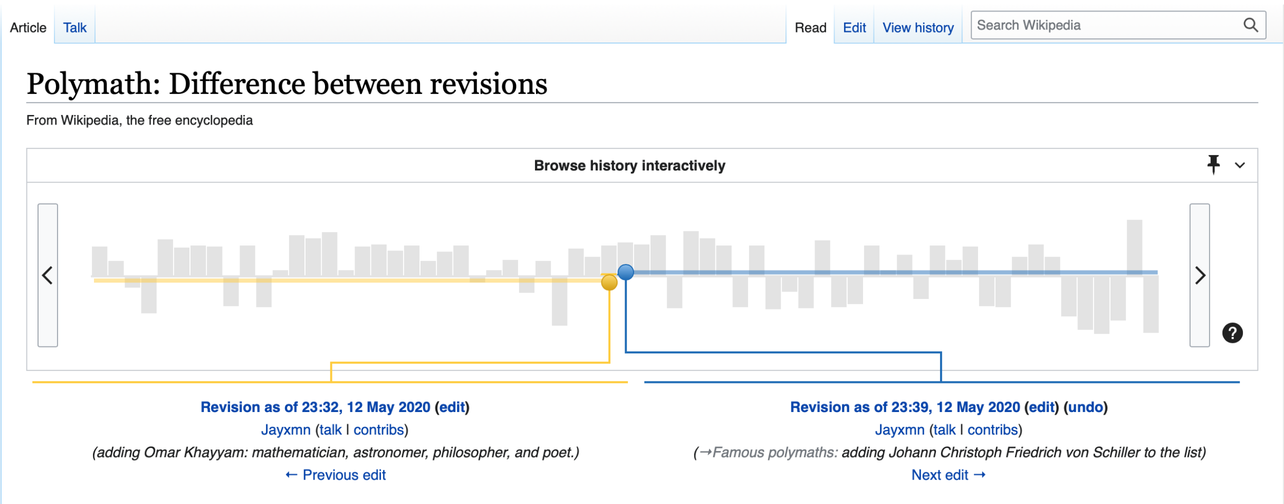
2K	1,3K	86
user edits	anon edits	bot edits

Thursday, 10 June

00:14 GMT -14 JPxG
m Reverted edits by 99.190.59.197 (talk) (HG) (3.4.11)

00:12 GMT +14 99.190.59.197
Briefly added some words

Figure 2.1: Mobile interactive visualization of the history



Polymath: Difference between revisions

From Wikipedia, the free encyclopedia

Revision as of 23:32, 12 May 2020 (edit)
Jayxmn (talk | contribs)
(adding Omar Khayyam: mathematician, astronomer, philosopher, and poet.)
← Previous edit

Revision as of 23:39, 12 May 2020 (edit) (undo)
Jayxmn (talk | contribs)
(→Famous polymaths: adding Johann Christoph Friedrich von Schiller to the list)
Next edit →

Figure 2.2: Interactive visualization of the history

2.2 Dataset

There are two datasets that store info about Wikipedia edits made available from the WikiMedia Foundation: a) the MediaWiki History and b) the MediaWiki History Dumps

The only difference other than the format (XML the former, TSV the latter) is that the former has the page content. The dataset used in this study is the MediaWiki History Dumps.

Each line of the TSV represent an event and, since it is denormalized, the events for user, page and revision are stored in the same schema. All event entity have different event types:

- Page: create, delete, move, reatore, merge
- User: create, rename, altergroup (change user rights), alterblocks (block user)
- Revision: create (edit a page)

In this analysis only revision events are of interest, there are 68 fields but only a few were needed. The entry could be divided in different sections: one section with general information of the revision like timestamp and comment, a section with information about the user who did the revision, one for the page where the revision was made, and the last one with more specific information about the revision. The most interesting fields of each section are represented in the Tables 2.1, 2.2, 2.3. In the caption are present, if needed, the descriptions of the fields.

id	username	groups	is_anonymous	registration	revision_count
42081	Checco	autopatrolled	False	2006-02-10 14:52:44.0	10479

Table 2.1: Data about the user who did the revision, the *groups* field helps to identify if the user is an admin, the *revision_count* is needed to calculate complex metrics like M and G.

id	title	namespace	revision_count
116530	Pino_Rauti	0	195

Table 2.2: Data about about the page where the revision took place, the *namespace* field is used to filter only the revisions because we are only interested in articles, i.e., the actual encyclopedia.

id	parent_id	is_reverted	reverter_id	is_reverter
73507165	73506955	True	73511400	False

Table 2.3: Data about the revision itself, we are able to identify if the revision is reverting another one, if it is been reverted and who is the reverter.

language	size
English	540 GB
Spanish	72 GB
Italian	54 GB
Catalan	12 GB

Table 2.4: Size of the dataset in different languages.

2.3 Definitions

It is worth defining some terms that will be used several times in the discussion.

Definition 1 *(Revert) On Wikipedia, reverting means undoing or otherwise negating the effects of one or more edits, which results in the page (or a part of it) being restored to a previous version.*

Definition 2 *(Revert chain) On a Wikipedia page, a revert chain occurs when an edit that reverts an edit is itself reverted.*

Definition 3 *(Mutual revert) A "mutual revert" is recognized if a pair of editors (x, y) is observed once with x and once with y as the reverter [3].*

Definition 4 *(Editor weight) The weight of an editor x is defined as the number of edits N performed by him or her [3].*

Definition 5 *(Mutual revert weight) The weight of a mutually reverting pair MW is defined as the minimum of the weights of the two editors [3].*

Definition 6 *(Chain weight) The weight of a revert chain CW is defined as the minimum of the weights of the editors involved in the chain.*

2.4 Metrics

Two complex controversiality metrics have been computed in this study: the first one, M, is the state of the art metric introduced by Yasseri *et al.* [3] which give us a score of the controversiality of the page based on the presence mutual reverts. The second one that we designed, called G, is very similar to M, but instead of using mutual reverts, it uses revert chains to evaluate the controversiality of the page.

2.4.1 M

The controversiality M of an article is defined by summing the weights of all mutually reverting editor pairs, excluding the topmost pair, and multiplying this number by the total number of editors E involved in the article.

$$M = E \sum_{all\ mutual\ reverts} MW \tag{2.1}$$

2.4.2 G

The controversiality G of an article is defined by summing the weights off all the chains there are on a page and multiplying by the total number of editors N involved in at least one chain.

$$G = N \sum_{all\ revert\ chains} CW \tag{2.2}$$

Chapter 3

Methods

3.1 Dataset

The dataset used is the wikimedia history dumps which is a large dataset derived from - TSV - each line is an event - event types: revision page user

i'm interested in revision event

here you the schema overview

more detailed info there

3.2 Approach

Considering the huge dimension of the dataset and the fact that a large portion of its content was useless, smaller datasets have been computed with the aim of expediting the analysis even for future usages.

the first skimming was to filter the rows, after this only the revisions which made o received a revert is keeped. than i ordered it by page, this sataset which is 10 % of the original one is the one i used to computeal the now dataset. . .

while this is just the biggest one without some rows, the structure of computed one is different. there are 2 modules

- Chains
- Group

image slide 5

Chapter 4

Results and Discussion

Chapter 5

Infrastacture

Chapter 6

Conclusions

Bibliography

[1] Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. Spatio-temporal analysis of reverted wikipedia edits. In *ICWSM*, 2017.

[2] Bongwon Suh, Ed H. Chi, Bryan A. Pendleton, and Aniket Kittur. Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 163–170, 2007.

[3] Yasseri T., Spoerri A., Graham M., and Kertész J. The most controversial topics in wikipedia: A multilingual and geographical analysis. *In: Fichman P., Hara N., editors, Global Wikipedia: International and cross-cultural issues in online collaboration. Scarecrow Press*, 2014.

Allegato A Titolo primo allegato

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

A.1 Titolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

A.1.1 Sottotitolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

Allegato B Titolo secondo allegato

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

B.1 Titolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.

B.1.1 Sottotitolo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec sed nunc orci. Aliquam nec nisl vitae sapien pulvinar dictum quis non urna. Suspendisse at dui a erat aliquam vestibulum. Quisque ultrices pellentesque pellentesque. Pellentesque egestas quam sed blandit tempus. Sed congue nec risus posuere euismod. Maecenas ut lacus id mauris sagittis egestas a eu dui. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque at ultrices tellus. Ut eu purus eget sem iaculis ultricies sed non lorem. Curabitur gravida dui eget ex vestibulum venenatis. Phasellus gravida tellus velit, non eleifend justo lobortis eget.