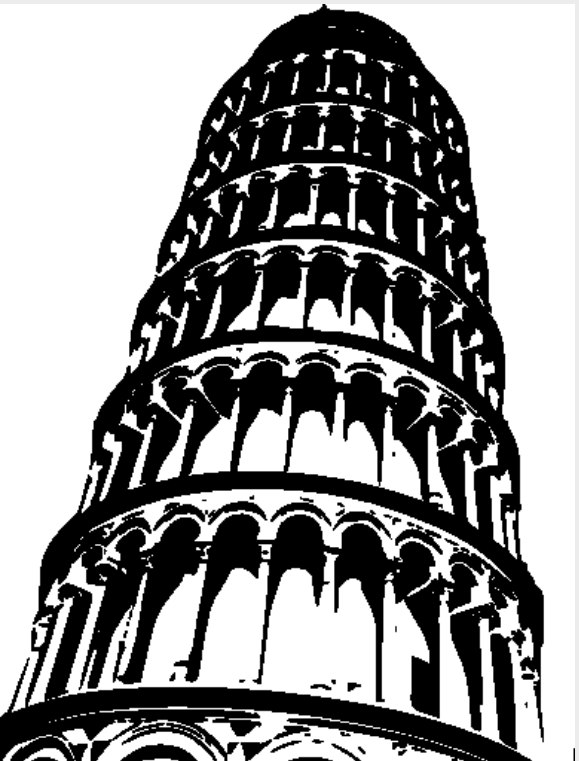
A proposito di… Wikipedia
Qualità e Attendibilità delle sue Voci

Vittoria Cozza



24 Ottobre 2015

SMS - Centro Espositivo

San Michele degli Scalzi, Pisa

Cosa è Wikipedia

- E' una enciclopedia online, libera, a contenuto aperto e multilingue, lanciata il 15 Gennaio 2001 da Jimmy Wales e Larry Sanger.
- Il concetto di enciclopedia online era stato già proposto da Rick Gates nel 1993, il concetto di libera (free as in freedom) era stato proposto da Stallman nel Dicembre 2000.
- Ward Cunningham, lo sviluppatore del primo software Wiki, WikiWikiWeb, la ha descritta come:

"the simplest online database that could possibly work"

Wikipedia: un pò di numeri

- Conta più di 35 milioni di voci in oltre 280 lingue, risultando così l'enciclopedia più grande mai scritta.
- Conta moltissimi utenti registrati e attivi:

English Wikipedia (update)	
Articles	4,991,534
Pages	37,604,072
Files	866,620
Edits	795,379,253
Users	26,499,112
Admins	1,332
Active users	125,785

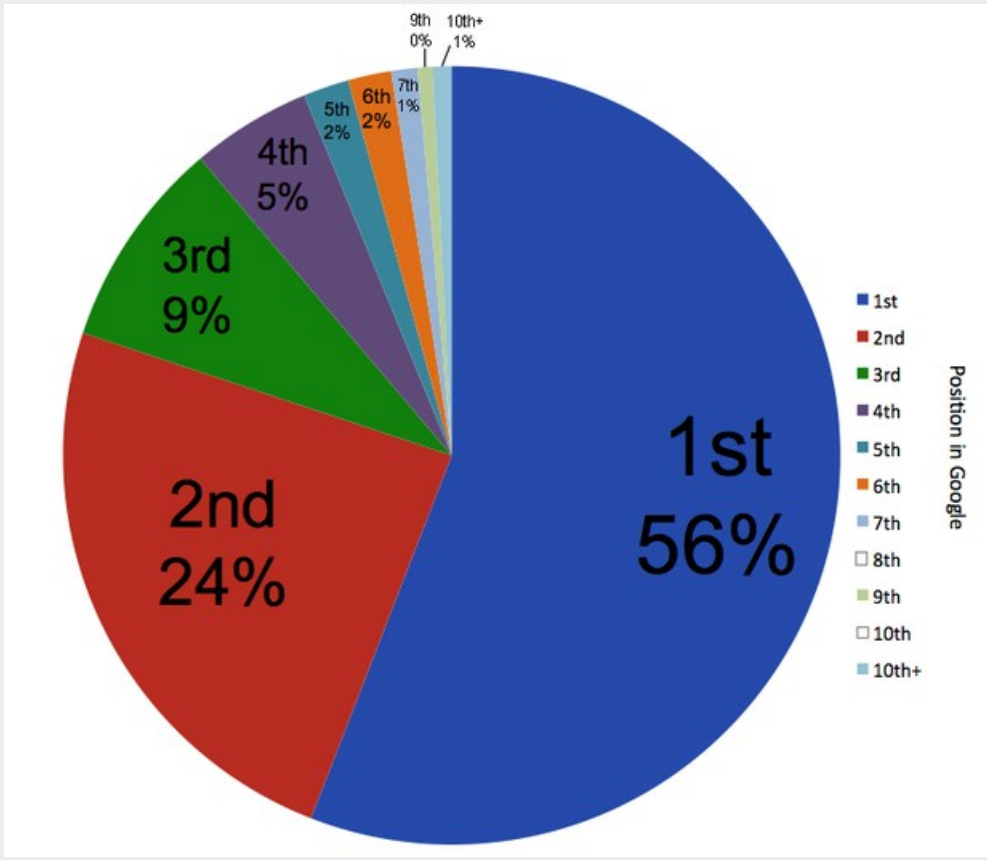
Source: https://en.wikipedia.org/wiki/Wikipedia:Statistics

- Conta moltissimi visitatori: *wikipedia.org* dal 2007 è sempre stata tra i 10 siti web più visitati al mondo

(source: http://www.alexa.com/siteinfo/wikipedia.org)

Perché tanti visitatori?

- Google mostra le Voci di Wikipedia fra i primi risultati di una ricerca; Wikipedia è [la principale rivale dei SEO](#) ☺
- Se c'è una Voce di Wikipedia corrispondente al termine della ricerca, questa apparirà nel risultato!



Wikipedia's positions on Google for 1,000 randomly generated search terms; 2012

Chiunque può **creare o ritoccare una voce** di **Wikipedia:** pro

- Ne deriva dunque una enciclopedia ricca di contenuti eterogenei, e che tengon conto di varie opinioni.
- La natura collaborativa del wiki fa si che i contributors si correggano vicendevolmente gli errori.

Legge di Linus: «Dato un numero sufficiente di occhi, tutti i bug vengono a galla» [Raymond]

Chiunque può **creare o ritoccare una voce** di **Wikipedia:** contro

- **Ritoccare una voce per estorsione:** Chiunque, scrivendo come se stesse esprimendo una opinione imparziale, può screditare la reputazione di una persona o una organizzazione per fini pubblicitari, propagandistici o addirittura a fine di ricatto.
- **Vandalism:** Chiunque con l'obiettivo di danneggiare l'immagine di Wikipedia può fare delle editing con contenuti inappropriati nelle voci: delle edit prodotte il 7% sono atti vandalici (Potthast 2010)
- **Lost in quality:** chiunque può scrivere testo non pertinente e/o in modo incoerente e sgrammaticato

Illustration by Sarah Rogers The Daily Beast

Ritoccare una Voce per estorsione

Londra, Settembre 2015

- Black hat editing: Utenti Wikipedia fingevano di essere amministratori e hanno ricattato imprenditori e celebrità minacciando di alterare i profili delle vittime su Wikipedia e quindi la loro reputazione online.
- Il crimine è noto come a COI scam: a conflict-of-interest editing on Wikipedia (COI) scam.
- Operazione Orangemoody: Sono stati bloccati gli account di 381 ricattatori, pare che abbiano estorto migliaia di sterline a diverse centinaia di persone.

Soluzione attuale

- Le attività di creazione e di editing delle Voci vengono controllati da programmi automatici detti bots
- I contenuti vengono supervisionati dagli Amministratori di Wikipedia… circa 4000 in totale (full statistics here!)
- Il caso di Wikipedia inglese: 1903 tipi di bots e circa 1000 amministratori incaricati di supervisionare un numero di modifiche al secondo che va dai 25.000 ai 60.000

Bot

- Si tratta di un programma automatico o semiautomatico che opera sulle pagine di Wikipedia come se fosse un utente
- Esempi sono:
  - User:ClueBot NG – reverts vandalism
  - User:CorenSearchBot – checks for copyright violations on new pages
  - User:Lowercase sigmabot III – archives talk pages
- Per un elenco completo: https://en.wikipedia.org/wiki/Wikipedia:Bots

Is a small number of people running the show?

- Polemiche imperversano sui blogs: su alcuni argomenti caldi (sport, politica, etc.) alcuni revisori sono accusati di non essere completamente imparziali nella valutazione dei contenuti.

Source: http://www.scientificethics.org/

Wikipedia:
a "shining example of Web democracy“
or
“a small number of people are running the show”?

Verso soluzioni automatiche intelligenti

- E’ possibile demandare il controllo di qualità e di affidabilità dei contenuti web, quindi anche di Wikipedia, a sofisticati strumenti basati su algoritmi di intelligenza artificiale e l’analisi del linguaggio naturale noti come strumenti di:
  - Automatic quality assessment
  - Vandalism detection
  - Opinion spamming e opinion spammer detection

Quality Assessment di Wikipedia

- Wikipedia fornisce delle linee guida per valutare la qualità degli articoli rispetto a 7 livelli di qualità che sono: Stub, Start, C, B, A, GA: Good article, FA: Featured article
- e si basano su diversi aspetti: linguistico, strutturale, storico e reputazionale.

Wikipedia Medicine articles by quality	
FA	61
GA	190
B	1,978
C	3,761
Start	11,853
Stub	8,777

Esempio: distribuzione articoli di Wikipedia Medica

Automatic Quality Assessment di Wikipedia

- Analisi automatica della qualità delle Voci di Wikipedia
- Esempi di apprendimento: dagli articoli già etichettati, vengono estratte delle features avanzate.
- Machine learning: un classificatore automatico, basato su algoritmi di IA, è capace di apprendere nuova conoscenza dall'osservazione di questi esempi

Quali features?

- Legate ai metadati della pagina: numero di editing, numero di revisori.
- Legate al contenuto testuale: alla struttura html della pagina (numero di link, immagini, riferimenti) e alla quantità dell'informazione contenuta nel testo (ripulito dai tag mediawiki)
- Legate al contenuto semantico: sofisticati tools per analisi del linguaggio naturale e ontologie di riferimento sono alla base dell'estrazione delle features semantiche

.**Automatic** Quality Assessment del portale di Medicina di Wikipedia

- Analisi della qualità delle Voci del progetto di Medicina raggiungibile da: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine)

- Pericoloso considerare attendibili questo tipo di informazioni se non lo sono!

> *Le informazioni riportate non sono consigli medici e potrebbero non essere accurate. I contenuti hanno solo fine illustrativo e non sostituiscono il parere medico: leggi le avvertenze.*

- Analisi automatica sfruttando anche features del dominio biomedicale

Joint work with Marinella Petrocchi, Angelo Spognardi – IIT CNR, Pisa, Italy

Istituto di Informatica e Telematica del CNR

Conclusioni

- Gli studi fatti dimostrano che :
  - la qualità di una Voce non è collegata al numero dei revisori e quindi delle revisioni;
  - la qualità di una Voce è legata alla reputazione degli autori;
  - analizzare caratteristiche del dominio di riferimento è importante per valutare la qualità di una Voce.
- Ogni idea, suggerimento o contributo è apprezzato!