



**Motori di Ricerca e
Duplicazione dei Contenuti**
Come funzionano
i Filtri Anti-duplicazione di
Google, Yahoo e Msn e
come "impattano" sul
Posizionamento del tuo sito web...

=> Per vedere il **Video** di questo intervento [clicca qui](#)



Pagina Duplicata - definizione?

potrebbe essere tale in base a:

- % di documento copiato?
- **struttura** del documento/ template?
- **sequenza** di frasi / parole?
- **query** effettuata (v. Serp - risultati supplementari)?
- **non e' cosi' facile** da definire (dipende + mix di vari algoritmi)
-

Differenza sostanziale tra:

1. Pagina "**clone**" perfetto (1:1) di un'altra (basta un "hash")
2. Pagina con **struttura clone** (paragrafi, disposizione di links, etc.) ma contenuti diversi (puo' scattare un red-flag/sospetto?)
3. Pagina con "**pezzi di contenuti**" identici ad altre

* Andiamo con ordine...

- 1) Facciamo la **Query** (es. scarpe da ginnastica)
- 2) I **documenti trovati** per quella key vengono **confrontati fra di loro...**
- 3) Quelli **ritenuti "simili" o "identici"** (in relazione alla keyword ricercata!) finiscono nei **risultati supplementari**
- 4) I **documenti duplicati** con **piu' alto Trust/Pr** vengono *inseriti nel listing normale e posizionati nei primi posti* quindi...
non per forza vincono quelli dell'*autore "originale"* o quelli *piu' "vecchi"* !

* Ma allora, come viene rilevato un "duplicato"?

Esistono varie possibili forme di analisi dei duplicati - es:

- *Hash*
- *Shingling*
- *Check-summing*
- *Lexical comparison*
- *Struttura del documento/template*
- *etc. (v. sotto + dispense)*

Tipologie di Duplicati:

1. **Duplicati cloni:** rilevati facilmente da tutti i motori tramite una funzione di **"hash"** (cioe' "fingerprint" numerico)
2. **Near Duplicates:** piu' difficile e piu' "esoso" in termini di risorse e tempo - in questo caso *la pagina viene divisa in "blocchi" , "cluster" o "fingerprints"* che vengono poi analizzati *secondo precise sequenze ed algoritmi*

Esistono **molte brevetti rilasciati dai motori** - es. Yahoo/Altavista (patents: 5,970,497 e 6,138,113), Google (6,615,209 e 6,658,423) e Msn (20060248066 e 20050210043) proponendo metodi differenti molto interessanti (v. dispense - argomenti molto tecnici):

*** Ecco alcuni esempi molto interessanti (v. video)**

Google

il brevetto di **Google** si focalizza principalmente sulla **creazione di hash e fingerprints di "parti del documento"** e della sua capacita' di rilevare "pezzi simili" e "differenze" (togliendo le parti "duplicate") - **citazione di una parte interessante del brevetto (n° 6,658,423):**

"Improved duplicate and near-duplicate detection techniques may assign a number of fingerprints to a given document by (i) extracting parts from the document, (ii) assigning the extracted parts to one or more of a predetermined number of lists, and (iii) generating a fingerprint from each of the populated lists. Two documents may be considered to be near-duplicates if any one of their fingerprints match"....

Per **capire meglio** analizziamo questo **interessantissimo Paper** di William Pugh (v. dispense)

Situazione Attuale - filtri anti-duplicazione

Ritengo che **nessuno dei 3 motori** (G., Msn, Y) abbia adottato un unico metodo di analisi dei duplicati ma un **mix di 2 o piu' modelli analizzati**, in particolare quelli basati su **"shingles" e "fingerprints"**.

Ovviamente la rilevazione dei duplicati e l'inserimento nei risultati supplementari nelle serp, **dipende molto dalla "query" effettuata:**

*** A tal fine Google in uno dei suoi brevetti dice:**

"An improved duplicate detection technique that uses query-relevant information to limit the portion(s) of documents to be compared for similarity is described. Before comparing two documents for similarity, the content of these documents may be condensed based on the query. In one embodiment, query-relevant information or text (also referred to as "snippets") is extracted from the documents and only the extracted snippets, rather than the entire documents, are compared for purposes of determining similarity."

*** Yahoo e i Templates - v. paper (dispense)**

Molto interessante anche lo studio di Yahoo sul problema dei siti basati su "Templates" (sempre + diffusi) nel riconoscimento di duplicati "falsi-positivi" per quelle pagine con pochi contenuti e template identici e/o "duplicati non rilevati" per quelle pagine con stesso contenuto ma templates differenti

Analisi della "realtà" e Deduzioni logiche

Premessa: tutto cio' che diro' ora e' frutto di mie analisi e deduzioni personali basate su valutazioni / test / esperienza personale: *non esiste nulla di ufficialmente approvato o riconosciuto dai motori (eccetto la lettura dei brevetti e l'analisi dei risultati/test)!*

Osservazione della capacita' di **trovare/filtrare duplicati:**

Esempio1:

Frase ricercata da "article marketing" diffuso: "[cost to have them run a reverse cell phone number](#)"

- [Google](#) (2 risultati => 29 con supplementari)
- [Yahoo](#) (33 risultati => 37 con supplementari)
- [Msn](#) (2 risultati "secchi")

Esempio2:

Frase ricercata da "wikipedia - doll": "[A doll is a toy often made in the likeness of a human baby or child](#)"

- [Google](#) (13 => 68 suppl.)
- [Yahoo](#) (183 => 284 suppl)
- [Msn](#) (2 secchi)

- *** Google**
sembra rilevare i duplicati **in funzione dei "blocchi di testo/frasi" che contornano la "parola/frase" ricercata** con risultati eccellenti (il 90-95% dei duplicati finisce sempre nei "risultati supplementari"). In questo modo e' in grado di rilevare tutti quei siti e blog (che sono migliaia) che creano pagine mixando pezzi di frase da fonti differenti (es. article marketing, rss feed, press releases).
- *** Yahoo**
Il **peggiore**. Non sembra avere lo stesso grado di capacita' e precisione (o forse non ha le "risorse" hardware/software per farlo): **bastano semplici modifiche** (title, description, struttura) e mix/remix di blocchi di testo da fonti differenti per superare il filtro anti duplicazione. I risultati supplementari sono proporzionalm. sempre molto pochi rispetto ai risultati naturali (segno che i "near duplicates" sfuggono ai controlli)
- *** Msn**
Sembra essere il **piu' "severo" e "potente"** nella rilevazione ed **eliminazione** dei duplicati. In realta' questo **accade soprattutto con determinate query da fonti note** - es. wiki, article marketing, etc. (forse Msn utilizza dei filtri piu' forti e rapiti x qs. siti) mentre in altre query sembra meno efficace (meglio di Yahoo comunque).

La GRANDE DIFFERENZA tra Msn, Yahoo e Google: Google usa il filtro anti-duplicazione anche per PENALIZZARE !

*** La grande DIFFERENZA:**

Yahoo e Msn usano il **filtro antiduplicaz.** principalmente per "filtrare i risultati" mentre **Google lo utilizza anche per Penalizzare!**

Negli **ultimi 2 anni** Google ha iniziato ad associare alle pagine duplicate una **qualche forma di penalizzazione** (piu' o meno intensiva in base a determinati fattori - v.dopo) che **puo' colpire non solo la pagina stessa ma, nei casi piu' estremi, anche l'intero sito (v. dopo)!**

*** Le RIVELAZIONI**

Molto interessante l'intervento di Adam Lasnik sul Blog ufficiale "Google Webmaster Central":

*"During our crawling and when serving search results, we try hard to index and show pages with distinct information... In the rare **cases in which we***

perceive that duplicate content may be shown with intent to manipulate our rankings and deceive our users, we'll also make appropriate adjustments in the indexing and ranking of the sites involved. However, we prefer to focus on filtering rather than ranking adjustments = PENALIZZAZIONE !"

*** Perche' questa forma "estrema" di penalizzazione?**

Per combattere la sempre piu' **crescente ondata di spam** che utilizza contenuti duplicati/grabbati per creare tonnellate di pagine e siti spazzatura (v. la grande fine degli spam engine a partire da 2 anni fa):

Grazie a questa forma di penalizzazione **Google e' in grado in brevissimo tempo** (poche settimane) di **riconoscere / penalizzare** e infine **bannare** le tonnellate di siti "spam" e di **"spam engine" che creano contenuti remixando fonti diverse**, attraverso contenuto rubato e di "scraping". Nonche' di **penalizzare** chiunque basi la creazione dei propri siti e pagine **"principalmente su contenuti duplicati"** eccetto casi rari (v. dopo press relases e article marketing).

La stessa cosa **NON accade con Yahoo e Msn** che usano tale filtro solo per "pulire" le serp ma non per penalizzare (eccetto rari casi di banning => siti al 90% cloni - ma banning che spesso richiede molti mesi)

Google - Livelli di Penalizzazione dei duplicati

*** Penalizzazione "ridotta" (della sola pagina):**

- **singola pagina duplicata** interamente o "parzialmente" da un'altra
- se la pagina **non ha sufficiente trust/pagerank** subisce una sorta di "svalutazione" che puo' incidere nel ranking di quella pagina, per piu' keyword/query.
- **rischi:** *chi ci copia/ruba contenuto*, nonostante possa non avere sufficiente Pr per posizionarsi prima di noi, *puo' in qualche modo influire sul nostro ranking*

* Penalizzazione "diffusa" (di piu' pagine e/o dell'intero dominio):

- **numerose pagine duplicate** interamente e/o parzialmente ma prevalenza sull'intero sito di pagine originali
- il livello di penalizzazione **puo' colpire l'intero dominio** se la duplicazione e' molto diffusa (es. >50% delle pagine del sito) e peggiorare ulteriormente se e' frutto di copia/clone di "**fonti note**" di cui esistono gia' molte copie in rete (es. articoli redistribuibili, wikipedia, serp, etc.)
- **soluzione:** bloccare lo spider su tutte o buona parte delle pagine clonate - esempio di www.TheFreeDictionary.com (area con clone di Wiki bloccata agli spiders con robots.txt)

* Banning

- accade **solo su domini con un numero elevato di pagine duplicate da fonti note** (es. sito con 1.000 pagine di cui 90% duplicate da wikipedia, article marketing, etc.) e con **basso Trust/PR** (es. 90% di links da siti con basso trust e/o spam)
- **n.b.** *I siti "clone" con medio/alto Trust non vengono bannati! (altrimenti tutti i siti di article marketing, press releases, etc. farebbero questa fine).*
- **unica contromisura:** alzare il Trust (tanto piu' la % di pagine e' duplicata) e/o modificare i contenuti per renderli unici

Note importanti:

- *** Risultato supplementare = non per forza penalizzazione!**
un documento che finisce nei "risultati supplementari" **non significa per forza che e' stato penalizzato** (dipende dalla query e dal grado di duplicazione della pagina e dell'intero sito!)

- *** Article Marketing e Press Releases**
chi diffonde comunicati o articoli già presenti nel proprio sito farebbe bene a **creare (almeno nel proprio sito) una versione unica, differente da quella diffusa**. Chi invece ripubblica deve considerare alcuni rischi (v. dopo)
- *** Struttura replicata = fingerprint e red-flag**
la struttura di un documento replicata su migliaia di pagine secondo schemi di contenuto-ridotto e struttura-identica (es. directory, database di e-commerce con pochi contenuti, oppure doorways malfatte, etc.) **non può essere causa automatica e immediata di penalizzazione** (altrimenti Dmoz, elenchi telefonici, etc. sarebbero tutti bannati) ma, al massimo, può diventare **indizio/traccia** (fingerprint) di un possibile "sito spam" => ovvero **aumenta le probabilità che scatti un "red-flag"** (sospetto = controllo quality rater?) soprattutto su siti con **pochi contenuti**
- *** Tempi di Penalizzazione**
la penalizzazione per duplicazione **non "scatta immediatamente"** in concomitanza allo spidering/indicizzazione (forse in futuro sì - v. phraserank) ma solitamente dopo almeno **10-30 gg. dall'indicizzazione** cioè dimostra che le pagine devono "passare" sotto l'analisi di algoritmi/filtri successivi (causa probabilmente elevata richiesta di risorse hardware/software per questo tipo di analisi - v. forse BigTable 2a tabella)

Motivi di Duplicazione e Soluzioni

* Domini differenti - stesso contenuto (es. Virgilio.com e Virgilio.it) - soluzione:

- **redirect 301** sul dominio con più forte trust
- differenziazione dei contenuti con **welcome su quello più debole** (usarlo per linkare quello più forte)

* Versione Stampabile

blocco con **robots.txt** - [esempio](#)

* Violazione di Copyright e/o Scraping da siti spam

- Ricerca di siti e pagine duplicate tramite **CopyScape** - www.copyscape.com (n.b. non spaventarti se copiano solo qualche pezzetto di frase => non basta per penalizzazione!)
- **se sito spam 100%** = segnalazione "**spam report**" tramite **Google SiteMap** (Google Webmaster Tools - www.google.com/webmasters/sitemaps/ (come specificato anche "ufficialmente" = probabilita' che una segnalazione di spam sia presa in consideraz. tramite sitemap sono molto + alte)
- **se sito non spam** (quindi non bannabile da quality raters) e clonazione di numerose pagine (o intero sito):
 - *soluzione1*: raccomandata dell'avvocato
 - *soluzione2*: segnalazione a Google => Digital Millennium Copyright Act - <http://www.google.com/dmca.html>

Vuoi saperne di più? Quale sara' il **futuro** dei motori di ricerca e dei suoi **algoritmi**? Come funzionano tutte le **penalizzazioni di Google** e come e' possibile **evitarle**?
=> Per saperne di piu' [clicca qui](#).



COPYRIGHT: Questo documento e' realizzato da [Madri Internet Marketing](#) ed e' tratto dall'ultimo corso [Seo Extreme](#).

DISTRIBUZIONE del contenuto: Questo documento puo' essere **ripubblicato e distribuito gratuitamente** nel tuo sito o nella tua newsletter a patto che venga sempre **citata la fonte** (con relativo link al sito <http://www.madri.com/>) e che **non** venga mai alterato o modificato il **contenuto** o i links in esso presenti.