Risposte WIM

Tommaso Sgarbanti

Università di Padova

tommaso.sgarbanti@studenti.unipd.it

Gennaio 2020

I. Descrivere Body Spam: caratteristiche, ambiti d'uso, pregi e difetti

È una tecnica utilizzata per aumentare il punteggio di una pagina web calcolato dai motori di ricerca. Consiste nell'inserire le keyword direttamente nel corpo del documento. E' una tecnica tanto semplice quando popolare ma bisogna accettare dei compromessi a causa del TF*IDF.
E' importante inoltre fornire una pagina sensata e ben strutturata all'utente, per questo motivo è bene nascondere dalla sua vista le keyword inserite nel body. Ciò si può fare per esempio inserendo il testo in bianco, su sfondo bianco, in una regione di 10x10 pixel, oppure reindirizzando l'utente sulla pagina che vogliamo fargli visualizzare, ma in questo caso occorre farlo con codice javascript non troppo banale, altrimenti i motori di ricerca individuerebbero il reindirizzamento.
Un'altra tecnica, non permessa e considerata eticamente scorretta, consiste nel fornire sotto lo stesso indirizzo al crawler, che si identifica come tale, la pagina con le keyword, mentre agli utenti la pagina che si vuole mostrare.
Quest'ultima tecnica, chiamata cloaking, anche se è la più efficace e la più difficile da individuare (dovrebbe venire di persona un dipendente a controllare la pagina) è anche quella che porta a penalità maggiori nel caso si venga scoperti, come per esempio la sospensione, a lungo termine, dall'indicizzazione e quindi dai risultati delle ricerche.

II. Descrivere Title Spam: caratteristiche, ambiti d'uso, pregi e difetti

È una tecnica utilizzata per aumentare il punteggio di una pagina web calcolato dai motori di ricerca. Consiste nell'inserire le keyword nel titolo della pagina e ha il vantaggio di non intaccare il contenuto della stessa.
Inoltre i motori di ricerca tendono a dare più peso ai termini inseriti nel titolo mentre gli utenti, solitamente, non gli prestano molta attenzione.

III. Descrivere Meta tag Spam: caratteristiche, ambiti d'uso, pregi e difetti

È una tecnica utilizzata per aumentare il punteggio di una pagina web calcolato dai motori di ricerca. Consiste nell'inserire le keyword negli appositi meta-dati.
Il vantaggio è che non viene intaccato il contenuto della pagina, come contro vi è che essendo una tecnica altamente utilizzata i motori di ricerca tendono a non dare molto peso alle parole inserite li, sopratutto se sono ripetute più volte.

IV. Descrivere Anchor text Spam: caratteristiche, ambiti d'uso, pregi e difetti

È una tecnica utilizzata per aumentare il punteggio di una pagina web calcolato dai motori di ricerca. Consiste nell'inserire le keyword nei nomi dei collegamenti ad altre pagine, le ancore fanno parte del body di una pagina ma vengono trattate separatamente.
I motori di ricerca, solitamente, tendono a dare un alto peso ai termini inseriti nelle ancore. Inoltre le keyword inserite nelle ancore sono, di solito, aggiunte anche alle pagine target senza che siano soggette a filtri di penalizzazione, ciò accade perchè si suppone che un'ancora dia una descrizione della pagina a cui punta.

V. Descrivere URL Spam: caratteristiche, ambiti d'uso, pregi e difetti

È una tecnica utilizzata per aumentare il punteggio di una pagina web calcolato dai motori di ricerca. Consiste nell'inserire le keyword direttamente nell'indirizzo web della pagina, i motori di ricerca infatti indicizzano e utilizzano anche gli stessi URL per calcolare i punteggi delle pagine, dando bonus simili a quelli dell'anchor text spam.
Ha il vantaggio di non intaccare il contenuto della pagina e solitamente viene usata in combinazione con altre tecniche di term spam. Anche con questa tecnica occorre agire con raziocinio, se si inserisce molteplici volte la stessa parola si verrà penalizzati.

VI. Tecnica della Repetition: caratteristiche, pregi, difetti, ambiti d'uso

È una tenica che serve ad aumentare il punteggio delle keyword in una pagina web. Consiste nel ripetere una o poche keyword più volte, stando attenti però al TF*IDF.
In questo modo si riuscirà ad incrementare la rilevanza della pagina rispetto ad una, o comunque ad un numero basso di keyword. Essendo una tecnica molto facilmente rilevabile dai crawler è necessario, se non si vuole essere penalizzati, prestare particolare attenzione alle contromisure.

VII. Tecnica del Dumping: caratteristiche, pregi, difetti, ambiti d'uso

È una tecnica utilizzata per aumentare il punteggio di una pagina web calcolato dai motori di ricerca. Consiste nell'inserire un numero alto di termini rari, anche se non sono correlati con il contenuto della pagina. In questo modo la pagina risulterà rilevante per molti termini diversi che, essendo rari, avranno soltanto poche altre pagine rilevanti.
Inserendo termini non pertinenti con il contenuto nella pagina, anche se facilità l'acesso al sito da parte degli utenti è più difficile poi farli restare, in quanto ciò che proporrà la nostra pagina sarà, già di partenza, diverso da ciò che stavano cercando (ciò può portare anche ad una perdita di trust da parte dell'utente).

VIII. Tecnica del Weaving: caratteristiche, pregi, difetti, ambiti d'uso

È una tecnica utilizzata per aumentare il punteggio di una pagina web calcolato dai motori di ricerca. Consiste nel copiare testo da altre pagine web ed inserirci all'interno le keyword in posizioni casuali.
Questa tecnica funziona meglio se l'argomento trattato dal testo copiato è raro, quindi di nicchia, ovvero per il quale ci sono solo poche pagine che risultano rilevanti. ~~Questa tecnica è inoltre utilizzata anche per "diluire" le keyword inserite, così che possano essere ripetute anche più volte riducendo la probabilità di essere penalizzati.~~

IX. Tecnica dello Stitching: caratteristiche, pregi, difetti, ambiti d'uso

È una tecnica utilizzata per aumentare il punteggio di una pagina web calcolato dai motori di ricerca. Consiste fare paste&copy da diverse sorgenti web, che siano siti o risposte di utenti nei forum, per poi assemblare il tutto con lo scopo di ottenere velocemente del contenuto rilevante.
Questa tecnica è utile per popolare velocemente un sito e per far si che la pagina abbia la possibilità di essere mostrata nei risultati di ricerca per ciascuno degli argomenti trattati nei testi copiati. Un altro vantaggio di questa tecnica è che molti motori di ricerca premiano i siti con un maggior numero di pagine.

X. Tecnica del Broadening: caratteristiche, pregi, difetti, ambiti d'uso

È una tecnica utilizzata per aumentare il punteggio di una pagina web calcolato dai motori di ricerca. Consiste in, oltre a inserire le keyword scelte, inserire anche sinonimi e frasi ad esse correlate.
Analogalmente allo stemming, eseguito da molti motori di ricerca, questo coprirà maggiormente le query degli utenti.
SVANTAGGIO: attenzione al TF*IDF perchè aggiungendo sinonimi stiamo diminuendo il TF*IDF di altre parole.

XI. Spiegare il problema dei posizionamento dei banner pubblicitari

E' importante che i banner pubblicitari seguano determinate regole per attirare l'attenzione dell'utente, di base, diversamente a quanto siamo abituati, nel web le cose nuove, strane e che creano un effetto dissonante attirano maggiormente l'attenzione.
Le immagini web e dunque anche i banner pubblicitari, godono di meno visibilità e attenzione da parte dell'utente rispetto al testo e questo è dovuto da un processo automatico e subconscio chiamato effetto zapping, che porta l'utente, durante lo scanning, a saltare tutto ciò che è immagine e, ancor di più, tutto ciò che sembra essere pubblicità. Per questo motivo, una tecnica utile per attirare l'attenzione degli utenti è il blending, ovvero eliminare le zone di contorno delle pubblicità confondendole insieme al contenuto della pagina, il blending che funziona meglio è quello testuale, per questo motivo è consigliato inserire del testo negli annunci.
Di pari passo con il blending è importante che i banner pubblicitari non presentino colori troppo vivaci e contrastanti rispetto al testo, in quanto trasmetterebbero

chiaramente di essere annunci pubblicitari con il risultato che non verrebbero presi in considerazione dall'utente.
Un altro aspetto importante è la dimensione infatti, nonostante non risulti essere una proprietà dominante, banner grandi tendono ad essere visti più facilmente rispetto che banner più piccoli, bisogna però assolutamente evitare che occupino la maggior parte della pagina o che ne coprano il contenuto, in questo caso infatti si otterrebbe l'effetto contrario. Inoltre un banner pubblicitario non deve caricarsi lentamente e non deve cercare di far cliccare l'utente in maniera fraudolenta.
Banner che si espandono al passaggio del mouse vengono visti di più, bisogna però cercare di posizionarli in punti strategici, in cui è probabile che l'utente passi con il mouse.
In ogni caso è importante che, un banner pubblicitario, sia capace di trasmettere il messaggio con una sola occhiata.
Assolutamente da evitare sono gli annunci pop-up, funzionano male perchè fanno inervosire l'utente e, se proprio vengono usati, deve essere chiaro da capire e facile da atturare il modo in cui vengono chiusi (questa è una regola che vale per qualsiasi banner pubblicitario, ma ancora di più per gli annunci pop-up).
Riguardo alle aree della pagina in cui è possibile posizionare banner pubblicitari vi è un ordine crescente per utilità ben definito: colonna di sinistra, top della pagina, colonna di destra, bottom della pagina. Un annuncio posto in alto nella pagina viene dunque visto un numero discreto di volte, ma comunque meno che un banner posto nella colonna sinistra della pagina.
Vi sono poi alcune cose che sono completamente da evitare, come suoni, spostamenti di posizione o lampeggiamenti. Uno stratagemma che invece sembra funzionare bene è di inserire pubblicità all'interno di mini-giochi, in quanto un giochino attira l'attenzione dell'utente e ne azzera i timer finchè è impegnato a giocare. E' importante però che il gioco segua il contenuto in cui è inserito, altrimenti si incorre nel distraction effect e i timer dell'utente diminuiscono di circa il 40% e la sua voglia di ritornare nel sito cala adirittura dell'80%.

XII. Descrivere che cos'è il fenomeno Lost in navigation, quando si verifica e come si può porrere rimedio

Il fenomeno lost in navigation si verifica quando l'utente è in una pagina e prende coscienza del fatto di essersi perso. È considerato un problema in quanto l'utente tende a chiudere la pagina o a premere sul pulsante indietro (richiede uno sforzo computazionale basso).
Questo fenomeno può accadere in seguito all'occultamento dell'asse where o in seguito all'apertura di una pagina dal layout completamente diverso. Una

soluzione è l'utilizzo dei breadcrumb (location, atttribute o path) in modo che l'utente sappia sempre dove si trovi; risulta inoltre utile colorare di un colore diverso i link alle pagine già visitate dall'utente, così che sia sempre a conoscenza di quali pagine ha già visitato.

XIII. Dublin Core, caratteristiche, pro/contro e ambiti d'uso

Il Dublin Core è uno dei primi tentativi di strutturare il web in maniera semantica. Permette di definire le proprietà di base fondamentali ed essenziali che modellano l'intera struttura dei dati semantici e, a tale scopo, sono stati definiti 15 elementi di metadati di base, opzionali, ripetibili e indipendenti dal dominio in cui vengono utilizzati (Titolo, Autore, Argomento, ID, Fonte, Data, Lingua, ecc...).
Come contro è troppo generico per descrivere in maniera adeguata risorse specifiche.

XIV. RDF e RDFS, caratteristiche, pro/contro e ambiti d'uso

RDF, che sta per Resource Description Framework, è uno strumento che permette di strutturare l'informazione e di rimuovere le ambiguità, fornisce una semantica ai dati permettendo dunque alle macchine di "capire" le informazioni presenti nel web.
RDF ha "sconfitto" XML in quanto i molti dialetti esistenti di XML rendevano impossibile l'aggregazione dell'informazione. Questo framework permette di descrivere metadati, relazioni e concetti garantendo l'interoperabilità tra essi, ossia lo scambio di informazioni, senza necessità di traduzioni di linguaggio. Permette di descrivere l'informazione basandosi sulla grammatica di base composta dalla tripla soggetto, predicato, complemento oggetto e ciascuno di questi tre elementi può contenere stringe o URI. Questa struttura può essere visualizzata come un grafo detto grafo della conoscenza e l'RDF permette l'aggregazione di più grafi della conoscenza collegando risorse identificate da uno stesso URI.
Gli oggetti devono però essere anche classificati garantendo un costo computazionale minimo e ciò viene permesso dall'RDF Schema. L'RDF offre infatti uno schema con una struttura informativa che per gli oggetti è fatta di classi, sotto-classi e indivui, mentre per i verbi è fatta di proprietà, sotto-proprietà, domini e intervalli (range) dove la proprietà è applicata. L'RDF ha fissato delle regole semantiche di livello base, l'RDF Schema permette di definire ontologie per categorizzare l'informazione, ma si parla di un livello di classificazione tassonomico.
I principali vantaggi dell'RDF è che è ben specificato, con diverse implementazioni e permette ai dati di essere decentralizzati e distribuiti in quanto chiunque può

creare un vocabolario o può pubblicare dati riguardo altre risorse, inoltre un grafo della conoscenza è concettualmente semplice da comprendere e analizzare. Come contro RDF è molto astratto e anche abbastanza verboso, risulta quindi snervante da scrivere o leggere manualmente, inoltre programmare interfacce richiede una conoscenza dei dettagli di base quali per esempio che cos'è un URI e che cos'è una tripla.
Due esempi di vocabolari espressi con RDF sono Dublin Core e Friend Of A Friend.

XV. FOAF, descrizione, ambiti d'uso, pregi e difetti

FOAF, che sta per Friend Of A Friend, è un vocabolario RDF nato per la socialità e non per le pagine web. Permette infatti di strutturare informazioni sulle persone, dunque sui loro interessi, sulle loro relazioni e sulle loro attività.
Essendo un'applicazione RDF può essere facilmente aggregato e combinato con altri vocabolari, permettendo così di catturare un insieme molto ricco di metadati.

XVI. SPARQL, caratteristiche, pro/contro e ambiti d'uso

SPARQL (SPARQL Protocol And RDF Query Language), è un linguaggio che permette le interrogazioni, formulate da delle query, all'interno del web semantico. Segue le stesse scelte di design di SQL, dunque è vincolato da alcuni limiti con il vantaggio che è decidibile, ovvero è garantita la terminazione, con una complessità computazionale PSPACE.
SPARQL lavora sulle strutture a grafo basandosi sul pattern matching tra triplette, una query SPARQL ha la seguente struttura:

- PREFIX ...
- SELECT ...
- FROM ...
- WHERE ...
- ORDERED BY ...

Un'aspetto interessante di questo linguaggio è la possibilità di ricerca con dati opzionali, ossia che potrebbero essere nulli o non presenti. Nel web semantico, infatti, è facile avere informazioni parziali e per questo in SPARQL è stato previsto l'operatore OPTIONAL. In questo modo, se nella query vi sono parametri in OPTIONAL e, durante la ricerca, alcuni dati associati a questi parametri risultano

assenti, la query non darà errore.
Uno svantaggio di SPARQL lo si ha in query particolarmente complesse, in quanto il costo computazionale si fa notare.

XVII. OWL, caratteristiche, pro/contro e ambiti d'uso

L'RDF permette l'aggregazione automatica della conoscenza attraverso gli URI, ma vi possono essere problemi di ambiguità. Infatti una stessa risorsa potrebbe essere presente sul web sotto URI diversi, questo problema è detto problema delle varianti degli URI ed è una delle motivazioni principali che ha portato all'aggiunta di un ulteriore strato ontologico, descritto attraverso il linguaggio OWL (Web Ontology Language).
OWL è un linguaggio W3C per le ontolgie, viene utilizzato per connettere vocabolari definendo relazioni tra di essi e permettendo di indicare quando due classi o proprietà sono la stessa e quando sono diverse.
Alcune funzionalità principali di OWL sono:

equivalentClass: per capire se due classi sono equivalenti;

equivalentProperty: per capire se due proprietà hanno lo stesso significato;

sameIndividualAs: per definire se due individui sono lo stesso;

differentFrom & allDifferent: per definire invece le ineguaglianze.

Per garantire sia una buona espressività del linguaggio che una logica decidibile facile da gestire vi sono tre diversi sottoinsiemi di OWL:

OWL Lite: con espressività limitata ma decidibile con una complessità computazionale SHIFT;

OWL DL: meno limitato nell'espressività, sempre decidibile ma con una complessità computazionale SHOIN;

OWL Full: sfrutta logiche più avanzate, quindi gode di un'alta espressività, ma non è decidibile.

XVIII. Descrivere la classificazione LOD: caratteristiche, ambiti d'uso ed eventuali pregi e difetti

Il web semantico è costituito dai cosidetti linked data, dove il web semantico è il tutto mentre i linked data sono le parti che lo compongono. I LOD, linked

open data, sono linked data che sono fruibili gratuitamente ed è molto importante facilitarne la creazione e la gestione in modo tale da creare nuova conoscenza e agevolare l'innovazione, rendendo possibile la realizzazione di servizi ed applicazioni sempre migliori.
I LOD sono classificati in una scala da 1 a 5 stelle.

1: Le informazioni sono disponibili sul web, in qualsiasi formato esse siano, sotto una licenza open (fruibili gratuitamente).

2: Le informazioni sono liberamente disponibili sul web e si presentano con un formato strutturato, ovvero machine readable.

3: Le informazioni sono liberamente disponibili sul web e si presentano con un formato strutturato non proprietario.

4: Le informazioni sono liberamente disponibili sul web e si presentano con formato semantic web, sono dunque impiegati degli identificatori URI affinchè sia possibile puntare a un singolo dato.

5: Le informazioni sono liberamente disponibili sul web, con formato semantico, e sono collegate ad altri dati così da fornire un contesto.

E' possibile, partendo da dati a 3 stelle (ma, volendo, anche da meno), raggiungere le 4 o 5 stelle dandogli un formato semantico. Questa operazione si chiama operazione di lifting ed esistono anche strumenti appositi per fare ciò. A livello pratico questo avviene unendo i grafi della conoscenza e, per vedere quali grafi è possibile collegare, si utilizzano misure di similarità come la distanza di Levensthein. L'operazione contraria, di diminuzione delle stelle, si chiama operazione di lowering e, entrambe le operazioni, sono dette operazioni di mashup.
Il principale problema dei LOD è che, per l'elaborazione di questi dati, è necessario ricorrere ad algoritmi di analisi di big data che sono computazionamente onerosi e, inoltre, non sempre l'utente si sente all'altezza di riuscire ad interfacciarsi con un database per estrarre i dati a lui necessari.
Esempi concreti di LOD sono DBpedia e schema.org.