PageRank

È importante dare un grado di interesse alle pagine. Il più delle volte questo dipende dal lettore, ma esistono dei meccanismi che giudicano il grado di una pagina in base all'importanza nel web.

Introduzione e motivazione

Per produrre un grado di importanza globale a tutte le pagine si analizza la struttura dei link. Questo approccio è chiamato *PageRank* e aiuta gli utenti dei motori di ricerca.
Definiamo come *backlink* una pagina che linka la pagina in questione.

Diversità delle pagine Web
Se si pensa semplicemente alle differenze tra una pubblicazione accademica ed una pagina web si può notare subito che le pubblicazioni sono riviste e corrette, le pagine web no, e può contenere delle citazioni (anche compromesse) che stanno dentro le pubblicazioni.

PageRank
Per misurare la qualità media di una pagina web è stato proposto PageRank, un metodo per calcolare un grado per tutte le pagine web basato sul grafo del web.
PageRank ha applicazioni in ricerca, navigazione e stima del traffico.

Un grado per tutte le pagine del web

Struttura dei link del Web
La struttura del web ha all'incirca 150 milioni di nodi e 1,7 miliardi di archi. Ogni pagina ha degli archi uscenti e degli archi entranti.
Generalmente "una pagina molto linkata è più importante di una pagina poco linkata".
PageRank fornisce un modo più sofisticato per fare questa ricerca. PageRank è interessante perché ci sono diversi casi in cui la citazione non corrisponde al senso comune di importanza.
Per esempio se una pagina ha un link che indica la home di Yahoo è un link molto importante. Questa pagina dovrebbe essere graduata più alta rispetto ad altre che hanno più link ma da posti oscuri. PageRank cerca di dare un'approssimazione sulla bontà dell'importanza dei link che può essere ottenuta dalla loro semplice struttura.

Definizione intuitiva di PageRank
Una pagina ha un grado più alto se la somma dei gradi dei suoi backlinks è alta.
Questo copre sia il caso in cui la pagina ha molti backlink, sia una pagina che ha pochi backlink ma di alto grado.

Definizione di PageRank
Sia u una pagina Web, e sia F_u l'insieme delle pagine puntate da u, e sia B_u l'insieme delle pagine che puntano ad u.
Sia $N_u = |F_u|$ il numero di link che partono da u e sia c un fattore usato per normalizzare così che il grado totale di tutte le pagine Web è costante. Noi possiamo iniziare a dare una definizione semplice di ranking: R che è una versione "leggermente modificata" di PageRank è:

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

Questo formalizza l'intuizione della sezione precedente. Notiamo che il grado di una pagina è diviso attraverso i suoi link uscenti regolarmente per contribuire ai gradi delle pagine a cui punta. Si noti inoltre che $c<1$ perché ci sono un numero di pagine senza archi uscenti, e il loro peso è perso dal sistema.
L'equazione è ricorsiva ma può essere calcolata partendo da un insieme di gradi e iterando il calcolo fino al raggiungimento di una convergenza.

Un altro modo è: sia A la matrice quadrata con righe e colonne corrispondenti alle pagine web. Sia $A_{u,v} = 1/N_u$ se c'è un arco da *u* a *v*, altrimenti $A_{u,v} = 0$.
Se trattiamo R come un vettore sulle pagine web allora R=cAR, così R è un autovetture di A con un autovalore *c*. Infatti noi l'autovettore dominante di A. Esso può essere calcolato tramite applicazioni ripetute di A ad ogni vettore iniziale non corrotto.
C'è un piccolo problema con questa funzione di ranking semplificata: consideriamo due pagine che si puntano a vicenda ma che non puntano a nessun'altra pagina e supponiamo che ci sia una pagina che punta ad una delle due. Allora durante l'iterazione, questo loop accumula rank, ma non distribuisce alcun rank! Così il loop forma una specie di trappola che viene chiamata calo di grado (*rank sink*). Per risolvere il problema introduciamo *rank source*.