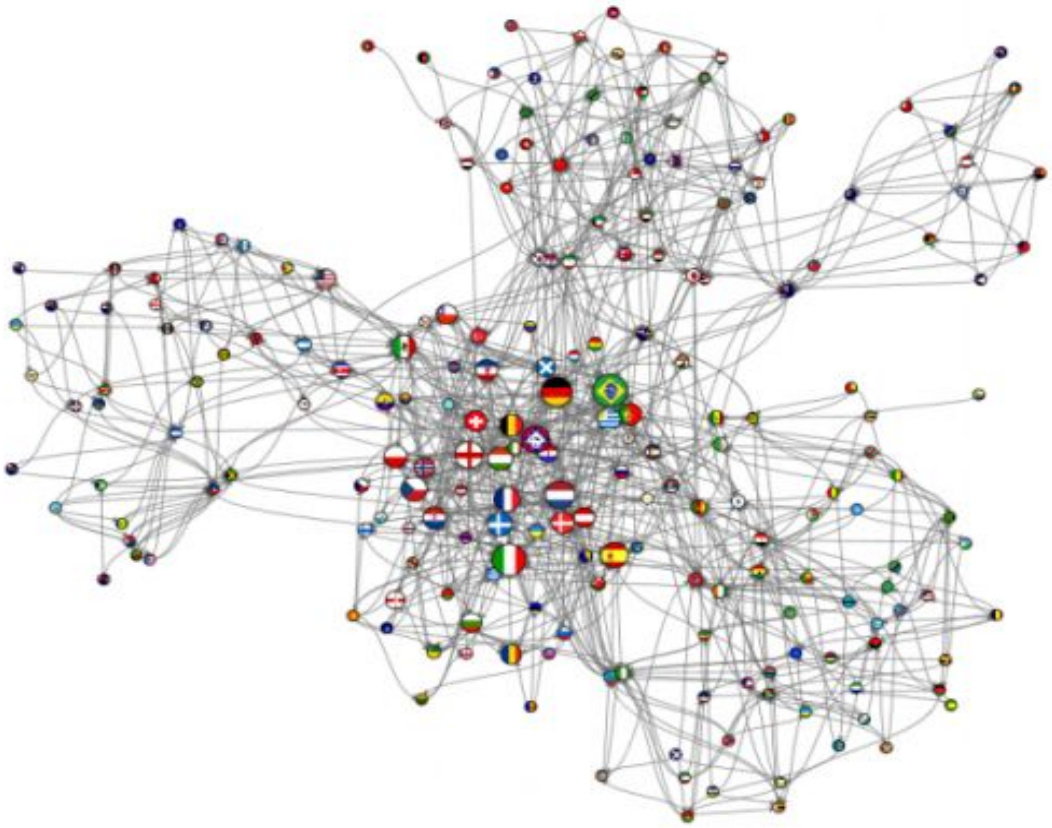
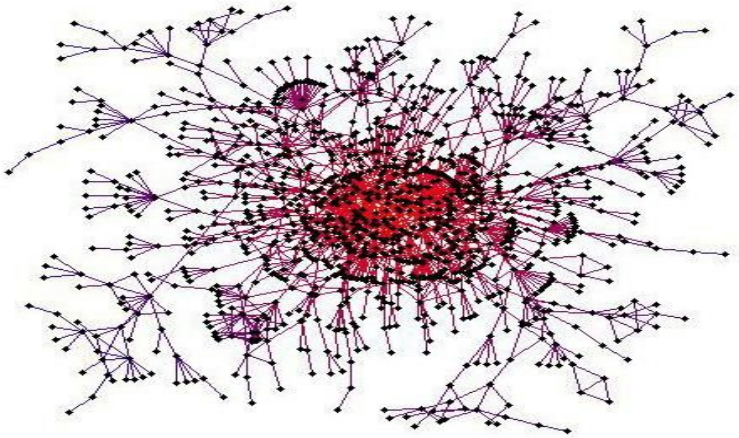
PageRank



مقدمه

دو الگوریتم مهم و مشهور HITS و PageRank در سال ۱۹۹۸ برای رتبه بندی صفحه های وب بر اساس محبوبیت و فراهم کردن نتیجه ی بهتر برای نتایج جستجوها ارائه شدند. HITS توسط جان کلاینبرگ از دانشگاه کورنل پیشنهاد شد.اما الگوریتم مورد بحث ما یعنی PageRank ، توسط لری پیج و سرجی برین که هر دو دانشجوی دانشگاه استنفورد بودند، ارائه شد.

طبق این الگوریتم ،ساختار وب تشکیل یک گراف عظیم جهت دار میدهد . نقاط در این گراف صفحه های وب را نشان می دهند و یالهای جهت دار لینک ها بین این صفحه ها را نشان می دهند. لینک هایی که به یک نقطه وارد می شوند را لینک درونی و لینک هایی که از یک نقطه در گراف خارج می شوند را لینک های بیرونی می نامند.



اگر بخواهیم این الگوریتم که قلب موتور جستجوگر گوگل است را در یک جمله معرفی کنیم باید گفت طبق این روش ، یک صفحه ی اینترنتی مهم و با اهمیت است اگر آن توسط صفحه های مهم دیگر ارجاع داده شده باشد .

مفهوم رتبه و روش اعمال الگوریتم

این الگوریتم به هر صفحه ی وب یک رتبه نسبت می دهد . رتبه ی صفحه ، یک مقدار عددی است که اهمیت در حال حاضر یک صفحه بر روی وب را نمایش می دهد.

هنگام ارجاع یک لینک از یک صفحه به یک صفحه ی دیگر ، میزانی از رای و رتبه صفحه اول به دیگری انتساب داده می شود . به همین شکل رای و رتبه های بیشتر اعتبار بیشتری را برای یک صفحه می آورد و همچنین اهمیت یک رای به اهمیت صفحه ای است که آن را انتساب می دهد.

موتور جستجوی گوگل اهمیت یک صفحه را برحسب میزان رای و رتبه ی آن می سنجد و براین اساس رتبه بندی می کند .

از این به بعد برای راحتی کار ، PageRank را با PR نمایش می دهیم.

بررسی و شرح الگوریتم PR

فرمول اصلی که در این روش استفاده می شود به شکل زیر می باشد که به بررسی و توضیح آن می پردازیم :

$$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

PR(A) : رتبه ی یک صفحه

PR(Ti) : رتبه ی صفحات Ti که به یک صفحه ی خاص لینک دارند

C(Ti) : تعداد لینک های بیرونی در صفحه ی Ti

d : فاکتور تعدیل با همان دمپ است که می تواند بین صفر تا یک باشد

حال برای شرح و توضیح این فرمول با یک مثال کار را شروع می کنیم :

فرض کنید d = 0.85 . پس

(PR) = 0.15 + 0.85 *

* مجموعی از رتبه ی تمامی صفحه هایی که به این صفحه لینک دارند .

حال میزانی که یک صفحه می تواند با آن به صفحه های دیگر رای دهد برابر با ارزش واقعی خودش است که اینجا برابر با ۰/۸۵ است . این مقدار به طور مساوی بین تمامی صفحاتی که به آن لینک دارد پخش می شود.

برای مثال یک صفحه با رتبه صفحه ی ۴ و ۵ لینک بیرونی از اهمیت بیشتری نسبت به یک صفحه با رتبه صفحه ی ۸ و ۱۰۰ لینک بیرونی دارد .

مقدار دقیق رتبه ها در گروی تکرار های زیاد است . برای مثال دوصفحه را درنظر بگیرید که تنها به یکدیگر ارجاع لینک بدهند . طبیعتا مقدار دقیقی نمیتوان به این دو صفحه انتساب داد و البته از اهمیت کمتری برخوردار هستند. در واقع یک دوره گردش تمام نشدنی در این شکل داریم . چون برای رسیدن به صفحه اول نیاز داریم صفحه دوم را دیده باشیم و بالعکس. اما ایجاد لینک ها و تکرار های بیشتر به کسب رتبه ی بهتر برای صفحه ها کمک می کند .

بررسی انواع مختلف لینک ها

در این قسمت به انواع لینک هایی که در گراف صفحه های وب وجود دارد می پردازیم .

۱. لینک های ورودی :

لینک هایی هستند که از خارج وارد یک وب سایت می شوند . این لینک ها یکی از راه های افزایش رتبه PL یک سایت هستند . البته درصورت کم بودن رتبه یا محتوای نامربوط سایت ها مقصر نیستند و از اهمیت آنها به این دلیل کاسته نمی شود .

۲. لینک های خروجی :

لینک هایی هستند که از یک سایت مورد نظر به سایت ها و صفحات مختلق دیگر ارجاع داده می شوند.

۳. لینک های آویزان:

لینک های آویزانی هستند که به هرصفحه ی دلخواهی بدون لینک خروجی ارجاع داده می شوند.

بررسی ریاضی الگوریتم PR :

در این قسمت به بخش های مختلفی از جمله فرمول جمع الگوریتم ، نمایش ماتریسی معادلات ، مساله ها با پردازش های تکراری ، نشانه گذاری های PR ، ارائه مدل ساده و محاسبه ی آن می پردازیم .

فرمول جمع الگوریتم :

r مجموع رتبه ی تمام صفحاتی است که به صفحه ی مورد نظر ما لینک داده شده اند .

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}$$

B_{P_i} ---- مجموعه صفحه هایی که به اشاره P_i می کنند.

$|P_j|$ --- تعداد لینک های بیرونی از صفحه ی P_j

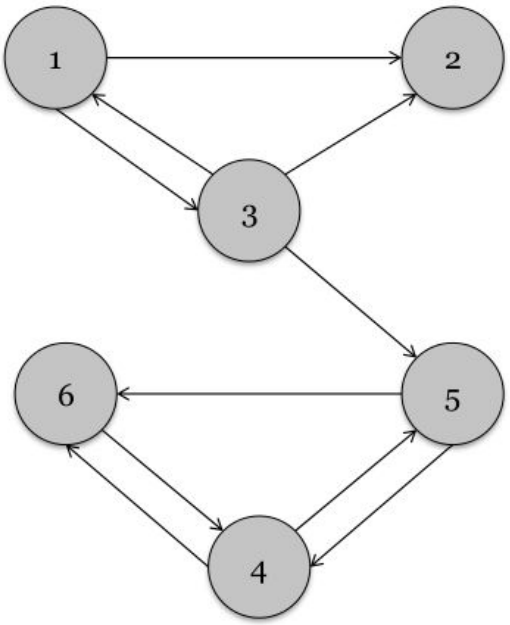
$r(P_j)$ --- مقدار در ابتدای کار نا معلوم است

تمامی صفحه ها دارای رتبه ی یکسان $\frac{1}{n}$ هستند. n تعداد صفحات است .

در معادله به شکل بازگشتی از مقادیر گذشته استفاده می شود . چون الگوریتم PL ، یک روش بازگشتی است .

با یک مثال محاسبه ی مقادیر بازگشتی را توضیح می دهیم:

گراف زیر با ۶راس را در نظر بگیرید.



در این جا به بررسی میزان رتبه ی یه گراف جهت دار از ۶ صفحه ی وب می پردازیم . بدیهی است که این یک مدل ساده برای نشان دادن نحوه ی کارکرد روش بازگشتی است .

در صفحه ی بعد با یک جدول روال را توضیح می دهیم .

رتبه در دومین تکرار	دومین تکرار	اولین تکرار	تکرار صفرم
5	R2(P1)=1/36	R1(P1)=1/18	R0(P1)=1/6
4	R2(P2)=1/18	R1(P2)=5/36	R0(P2)=1/6
5	R2(P3)=1/36	R1(P3)=1/12	R0(P3)=1/6
1	R2(P4)=17/72	R1(P4)=1/4	R0(P4)=1/6
3	R2(P5)=11/72	R1(P5)=5/36	R0(P5)=1/6
2	R2(P6)=14/72	R1(P6)=1/6	R0(P6)=1/6

نمایش ماتریسی معادلات :

در اینجا علامت سیگما را با ماتریس ها جایگزین می کنیم و در هر مرحله ی تکرار ، بردار رتبه صفحه (که یک بردار $n \times 1$ است) تمام مقادیر رتبه صفحه ی محاسبه شده را در خود نگه می دارد .

ماتریس $n \times n$ H برای نشان دادن لینک ها می باشد که عنصر های آن به شکل $Hij = \frac{1}{|Pi|}$ برای نقاطی که باهم لینک دارند (i,j هایی که به هم لینک دارند) یا اینکه 0 برای بقیه.

ماتریس H برای گراف صفحه ی قبل به شکل زیر است :

H =

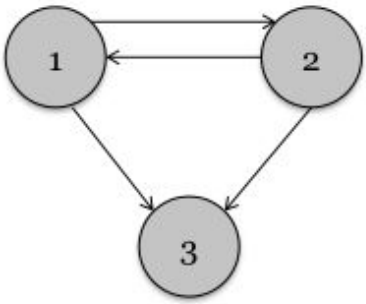
	P1	P2	P3	P4	P5	P6
P1	0	1/2	1/2	0	0	0
P2	0	0	0	0	0	0
P3	1/3	1/3	0	0	1/3	0
P4	0	0	0	0	1/2	1/2
P5	0	0	0	1/2	0	1/2
P6	0	0	0	1	0	0

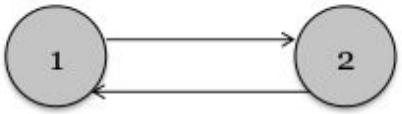
هر تکرار شامل یک ضرب ماتریسی است که هزینه ی محاسباتی آن از مرتبه ی $O(n^2)$ است. لازم به ذکر است که H یک ماتریس اسپارسی با 0 های زیاد است که یک فضای حافظه ی حداقلی می خواهد که مرتبه محاسباتی آن $O(nnz(H))$ است که $nnz(H)$ نشان دهنده ی تعداد غیر صفر های ماتریس H است .

عملیات محاسبه در هر تکرار یک عملیات خطی تابت است . H همانند یک ماتریس احتمال انتقال تصادفی برای زنجیره ماروف است که نقطه های آویزان در گراف ردیف های صفر در ماتریس را ایجاد می کنند.

حال به بررسی مشکلاتی که در تکرار ها ایجاد می شود می پردازیم:

پیج و برین وقتی این الگوریتم رو ارائه کردند به دو مشکل دور داشتن و رتبه دهی نزولی برخوردند.





رتبه دهی نزولی:

صفحه ای که در یک تکرار رتبه صفحه ی بیشتری دارد و امتیاز بیشتری تجمع کرده است ، در ادامه ی تکرار ها رتبه ها را انحصاری به سمت خود می کشد. در گراف ۳ نقطه ای بالا نقطه ۳ یک نقطه ی نزولی است .

دور:

در شکل گراف ۲ نقطه ای بالا یک دور بی نهایت وجود دارد و پایان این پردازش هیچ وقت معلوم نیست .

یک راه حل ساده :

در اینجا قصد داریم یک راه حل برای این دو مشکل ارائه دهیم . برای این کار از مدل پیمایشگر تصادفی استفاده می کنیم. پیمایشگر تصادفی با توجه به اسمش به شکل تصادفی بین لینک های ساختار وب حرکت می کند.

به این شکل که یکی از چندین لینک بیرونی حاضر در صفحه را انتخاب می کند. اهمیت یک صفحه با میزان زمانی که پیمایشگر برای یک صفحه اختصاص داده است سنجیدی می شود. البته این مدل ساده مشکل هایی نیز به همراه خود دارد . از جمله اینکه اگر پیمایشگر با یک نقطه ی آویزان همانند عکس ،فایل و .. روبه رو شود، گرفتار می شود .

برای این منظور یک اصلاحیه ای برای این روش ایجاد شد . اول اینکه پیمایشگر باید مشکل لینک های آویزان را حل کند . دوم اینکه مشکل انتقال از راه دور لینک ها نیز باید برطرف شود .

برای مشکل اول تغییر ماتریس H به یک ماتریس تصادفی پیشنهاد شد . برای مشکل دوم یک ماتریس جدید تعریف شد که به ماتریس گوگل معروف بود . این ماتریس با G نشان داده می شدکه برای محاسبه فرمولی به شکل زیر داشت :

$$G = \alpha S + (1-\alpha)1/nee\ T$$

دراینجا α یک مقدار بین صفر و یک است .

نکات جمع بندی و چند مثال:

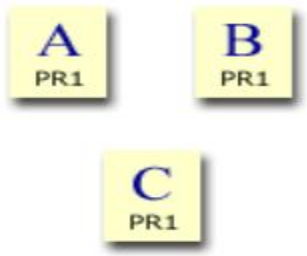
یک وب سایت دارای رتبه صفحه ی بیشترین است که این رتبه بین تمام صفحاتش با لینک درونی توزیع شده باشد .

میزان رتبه صفحه یک سایت با افزایش صفحات یک سایت افزایش می یابد و رابطه ی مستقیم دارد.

ایجاد لینک در بین صفحه های سایت برای کسب رتبه صفحه ی بالاتر لزومی است .

حال به بررسی چند مثال می پردازیم :

مثال ۱: شکل زیر را در نظر بگیرید.



اگر هر کدام از صفحات رتبه یک داشته باشند پس رتبه صفحه ی سایت سه می باشد.

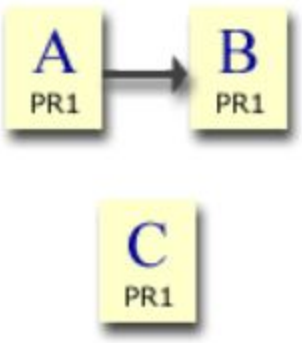
اما اگر d=0.85 باشد می بینیم که هر کدام از صفحه ها فقط رتبه ی 0.15 دارند .

این به دلیل عدم وجود لینک های درونی در گراف است.

پس به همین دلیل امکان دارد میزان جمع رتبه ی سایت 0.45 باشد به جای اینکه 3 باشد که نشان دهنده ی اتلاف رتبه ی سایت است.

مثال ۲ :

گراف زیر را در نظر بگیرید.



اگر هر کدام از صفحات رتبه یک داشته باشند پس رتبه صفحه ی سایت سه می باشد.

اگر d=0.85 باشد می بینیم که :

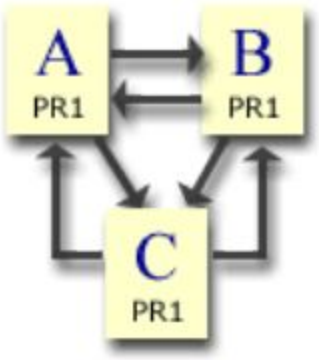
- ➔ PR(A)=0.15
- ➔ PR(B)=1
- ➔ PR(C)=0.15

بعد از ۱۰۰ تکرار رتبه A و رتبه C یکسان اند اما رتبه ی B به 0.2775 کاهش می یابد.

بنابراین جمع رتبه ی سایت به 0.5775 می رسد که در این حالت بهتر از مدل قبلی است اما باز هم با مدل ایده آل مقدار ۳ فاصله زیادی دارد .

مثال ۳:

گراف زیر را در نظر بگیرید:



اگر هر کدام از صفحات رتبه یک داشته باشند پس رتبه صفحه ی سایت سه می باشد.

اگر d=0.85 باشد می بینیم که :

- ➔ PR(A)=1
- ➔ PR(B)=1
- ➔ PR(C)= 1

بعد از هر تعداد تکرار رتبه ی تمام صفحات همچنان یک خواهد بود که این بیشترین میزان مورد نظر است و نشان دهنده ی کارآمد بودن لینک های درونی در گراف برای افزایش رتبه می باشد.

مرجع ها :

Google's PageRank and Beyond by Amy N.Langville and Carl D.Meyer

http://www.webworkshop.net/pagerank.html#how_is_pagerank_calculated

http://pr.efactory.de/e-further-factors.shtml

http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html

به کوشش :

گروه پنج درس جبرخطی عددی (رضاقنبری - نگین لوا - هنگامه نثاری زاده - روزبه ایزدیان)