گوگل چگونه کار می‌کند؟

« مقدمه‌ای بر پیوندکاوی »

نویسنده

سید جمال میرکمالی

گروه دانش آماری– مهر ۱۳۸۸

ایده‌ی گوگل

وب‌گردها در هنگام گشت و گذار در اینترنت معمولا از موتورهای جستجو برای یافتن صفحات وب مرتبط با موضوع مورد علاقه‌ی خود استفاده می‌کنند. متاسفانه (!) اغلب هزاران صفحه‌ی وب در ارتباط با این پرس و جوها[1] وجود دارد. بنابراین وجود یک فهرست از صفحات وب که بر اساس اهمیت مرتب شده‌اند ضروری به نظر می‌رسد. این فهرست بایستی به طور منظم و مکررا روزآمد[2] شود. بنابراین یافتن الگوریتم‌های سریع برای محاسبه‌ی رتبه‌ی صفحه بگونه‌ای که تاخیر در روزآمدسازی را کاهش دهد از اهمیت بسزایی برخوردار است. به نظر می‌رسد که این مساله‌ی دشواری است. نه فقط به دلیل اینکه تعداد صفحات وب در اینترنت بسیار زیاد است بلکه تعداد آن‌ها به سرعت در حال افزایش است.

رتبه‌ی صفحه[3] توسط پیج و همکاران (۱۹۹۸) برای بیان اهمیت هر صفحه‌ی وب پیشنهاد شد (اینجا را ببینید). لری پیج و سرگی برین پایه‌گذاران گوگل هستند. در واقع، می‌توان عبارت‌های زیر را در سایت گوگل (در اینجا) مشاهده کرد: "قلب نرم‌افزار ما رتبه‌ی صفحه است؛ سامانه‌ای که برای رتبه‌بندی صفحات وب توسط موسسین ما یعنی لری پیج و سرگی برین در دانشگاه استنفورد تولید شده است. در حالی که مهندسان بسیاری به طور روزمره جنبه‌های مختلف گوگل را بهبود می‌بخشند با این وجود رتبه‌ی صفحه به عنوان مبنای همه‌ی ابزارهای جستجوی ما همچنان به کار خود ادامه می‌دهد."

[1] Query
[2] Update
[3] PageRank

ایده‌ی مشابهی در خصوص رتبه‌بندی مجله‌ها توسط گارفیلد (۱۹۵۵ و ۱۹۷۲) به عنوان درجه‌ی اعتبار مجله‌ها تحت عنوان *ضریب تاثیرگذاری*[1] معرفی شد. ضریب تاثیرگذاری یک مجله از تقسیم تعداد ارجاع‌ها به آن مجله بر تعداد مقالات آن مجله به دست می‌آید. در واقع ضریب تاثیرگذاری متوسط تعداد ارجاع‌ها به یک مقاله از یک مجله است. با در نظر گرفتن هر صفحه‌ی وب به عنوان یک مجله، این ایده برای اندازه‌گیری اهمیت صفحات وب در الگوریتم رتبه‌ی صفحه گسترش داده شده است.

رتبه‌ی صفحه به صورت زیر تعریف می‌شود. فرض کنید N تعداد کل صفحات وب موجود در اینترنت باشد و ماتریس Q را به نام ماتریس *فراپیوند*[2] به صورت زیر تعریف کنید:

$$Q_{ij} = \begin{cases} ۱/k & \text{اگر صفحه‌ی وب } i \text{ یک پیوند بیرونی صفحه‌ی وب } j \text{ باشد} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

و k تعداد کل پیوندهای بیرونی صفحه‌ی وب j است. برای سادگی در اینجا فرض می‌کنیم که برای هر i داشته باشیم $Q_{ii} > ۰$. این یعنی هر صفحه‌ی وب یک پیوند به خودش دارد. بنابراین Q را می‌توان به عنوان یک ماتریس احتمال انتقال یک زنجیر مارکوف در یک قدم زدن تصادفی در نظر گرفت. مساله شبیه به این است که وب‌گرد را به عنوان شخصی که به طور تصادفی قدم برمی‌دارد و صفحات وب را به عنوان وضعیت‌های یک زنجیر مارکوف در نظر بگیریم. با فرض اینکه زنجیر مارکوف ناشکننده[3] باشد توزیع احتمال ایستای

$$(p_۱, p_۲, \ldots, p_N)^T$$

[1] Impact Factor
[2] Hyperlink
[3] Irreducible

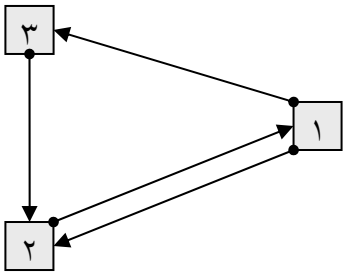
برای وضعیت‌ها (صفحات وب) وجود دارد. در اینجا p_i نسبتی از زمان است که شخصی که به طور تصادفی قدم برمی‌دارد (وب‌گرد) وضعیت (صفحه‌ی وب) i را ملاقات می‌کند. هر چه مقدار p_i بیشتر باشد به معنی اهمیت صفحه‌ی وب i است. بنابراین رتبه‌ی صفحه‌ی i همان p_i تعریف می‌شود.

یک مثال

یک شبکه‌ی اینترنت شامل ۳ صفحه‌ی وب در نظر بگیرید: ۱، ۲، ۳ بگونه‌ای که

$$\begin{aligned} &۱ \to ۱,۱ \to ۲,۱ \to ۳ \\ &۲ \to ۱,۲ \to ۲, \\ &۳ \to ۲,۳ \to ۳. \end{aligned}$$

این روابط را می‌توان به صورت زنجیر مارکوف زیر نمایش داد.



شکل ۱: یک مثال با سه صفحه‌ی وب

ماتریس احتمال انتقال زنجیر مارکوف به صورت زیر به دست می‌آید:

$$Q = \begin{array}{c} ۱ \\ ۲ \\ ۳ \end{array} \begin{pmatrix} ۱/۳ & ۱/۲ & ۰ \\ ۱/۳ & ۱/۲ & ۱/۲ \\ ۱/۳ & ۰ & ۱/۲ \end{pmatrix}.$$

توزیع احتمال ایستای زنجیر مارکوف $\mathbf{p} = (p_۱, p_۲, p_۳)$ در شرط $\mathbf{p} = Q\mathbf{p}$ و $p_۱ + p_۲ + p_۳ = ۱$ صدق می‌کند. با حل این معادلات داریم:

$$\mathbf{p} = (\frac{۳}{۹}, \frac{۴}{۹}, \frac{۲}{۹}).$$

بنابراین ترتیب اهمیت صفحات وب به صورت زیر است:

صفحه‌ی وب ۳ > صفحه‌ی وب ۱ > صفحه‌ی وب ۲

از آنجا که زنجیر مارکوف بسیار بزرگ است و مدت زمان محاسبه‌ی رتبه‌ی صفحه توسط گوگل فقط چند روز است، روش مستقیم محاسبه‌ی توزیع احتمال ایستا مطلوب نیست. به همین دلیل روش‌های متنوعی بر اساس تجزیه و تکرار ارائه شده‌اند. مطالعات گسترده‌ای در این خصوص تحت عنوان *بهینه‌سازی موتورهای جستجو*[1] (SEO) شکل گرفته است که نتیجه‌ی آن معرفی شیوه‌های تطبیقی، الگوریتم‌های موازی و روش هیبرید بوده است.

از آنجا که ماتریس Q ممکن است شکننده باشد محاسبه‌ی توزیع احتمال ایستا با مشکل مواجه است. در این صورت می‌توان از ماتریس P

$$P = \alpha Q_{N\times N} + \frac{(۱-\alpha)}{N}\mathbf{1}_{N\times N}$$

استفاده کرد که در آن $۰ < \alpha < ۱$ و $\mathbf{1}_{N\times N}$ یک ماتریس $N \times N$ با مقادیر ۱ است. در این صورت ماتریس P ناشکننده و نادوره‌ای[2] است و توزیع احتمال ایستا وجود دارد و یکتاست (راس ۲۰۰۰). مقادیری که برای α در نظر گرفته می‌شود معمولا ۰/۸۵ یا $۱ - ۱/N$ است.

تعبیری که برای ماتریس P می‌توان ارائه کرد این است که در یک شبکه شامل N صفحه‌ی وب، هر صفحه دارای اهمیت ذاتی به اندازه‌ی $(۱-\alpha)/N$ است. اگر صفحه‌ی وب i

[1] Search Engine Optimization
[2] Aperiodic

دارای اهمیت p_i باشد آنگاه به اندازه‌ی αp_i در اهمیت صفحات وبی که به آن‌ها پیوند داده است شریک خواهد بود. اهمیت صفحه‌ی وب i از حل معادله‌ی زیر به دست می‌آید:

$$\mathbf{p} = \alpha Q\mathbf{p} + \frac{(1-\alpha)}{N}\mathbf{1}_{N\times 1}$$

برای توضیحات بیشتر در خصوص الگوریتم‌های محاسباتی به وایکی و میخاییل (۲۰۰۶) یا لنگویل و می‌یر (۲۰۰۶) مراجعه کنید.

در آخر به این نکته توجه کنید که در سامانه‌ی اندازه‌گیری اهمیت صفحات وب گوگل هیچ‌گونه هوشمندی از نوع تفسیر محتوا صورت نگرفته است و آنچه که موجب محبوبیت موتور جستجوی گوگل شده است در واقع استفاده از هوش نهفته‌ای است که کاربران اینترنت در پیوندها جاسازی می‌کنند. قاعدتا این پایان راه گوگل نخواهد بود و ممکن است روزی برسد که بتوان در گوگل یک قطعه‌ی آواز، یک فیلم یا یک عکس شبیه

را جستجو کرد و گوگل شما را مثلا به اینجا راهنمایی کند.

مرجع‌ها:

Garfield E (1955) *Citation Indexes for Science: A New Dimension in Documentation Through Association of Ideas*, Science, 122:108–111.
Garfield E (1972) *Citation Analysis as a Tool in Journal Evaluation*, Science, 178:471–479.
Langville A N, Meyer C D (2006) *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University.
Page L, Brin S, Motwani R and Winograd T (1998) *The PageRank Citation Ranking: Bring Order to the Web*, Technical Report, Stanford University.
Ross S (2000) *Introduction to Probability Models, 7th Edition*, Academic Press.
Wai-Ki C, Michael K (2006) *Markov Chains Models, Algorithms and Applications*, Springer.