الگوریتم جستجوی گوگل

محمد علی کرمی

گوگل[1] موفق‌ترین موتور جستجو[2]ی دنیاست و شاید یکی از رمزهای موفقیت آن، الگوریتم جستجوی آن باشد. در این نوشتار سعی می‌کنیم ایده‌ی استفاده شده توسط گوگل برای جستجو در شبکه‌ی اینترنت را توضیح دهیم.

مدلی که برای شبکه‌ی اینترنت می‌توان در نظر گرفت، یک گراف جهت‌دار است که در آن راس‌ها نماینده‌ی سایت‌های اینترنتی (یا سرور آن سایت‌ها) و یال‌های جهت‌دار نماینده‌ی وجود یک پیوند[3] از سایت مبدا به سایت مقصد است. اگر بتوانیم به هر سایت، یک عدد به عنوان ارزش یا معیار اهمیت نسبت دهیم، آن‌گاه می‌توانیم برای جستجو در اینترنت، در میان سایت‌هایی که واژه‌ی جستجو شده را دربردارند، آن‌هایی که ارزش بیشتری گرفته‌اند را زودتر نشان داده و در اولویت بالاتر قرار دهیم. بنابراین گوگل در کنار یک الگوریتم سریع و کارآمد جستجوی رشته[4] از یک الگوریتم تخصیص ارزش به صفحات وب نیز استفاده می‌کند که Page Rank نام دارد.

پس هدف یافتن تابعی مانند $PR : V \to \mathbb{R}^+$ است که در آن V مجموعه‌ی راس‌های گراف یا همان مجموعه‌ی کل سایت‌های اینترنت است و فرض می‌کنیم هر سایت یک Page Rank مثبت دارد (ارزش هیچ سایتی را صفر در نظر نمی‌گیریم)، البته این موضوع زیاد اهمیتی ندارد ولی چیزی که مهم است این است که ارزش هیچ سایتی منفی نیست، یعنی به نحوی دنبال یک اندازه احتمال روی شبکه‌ی اینترنت هستیم به گونه‌ای که سایت‌های با اهمیت‌تر، احتمال بالاتری بیابند.

مبتکر ایده‌ی Page Rank لری پیج[5]، که خود یکی از دو موسس شرکت گوگل نیز هست، ایده‌ی این ارزش‌گذاری را از روش ارزش‌گذاری مقالات علمی اخذ کرد. مقالات علمی براساس تعداد ارجاع[6] و تعداد نقل قول[7] ارزش‌گذاری می‌شوند و هر چه به مقاله‌ای بیشتر ارجاع داده شده باشد، نشان از با اهمیت‌تر بودن و پرکاربردتر بودن آن مقاله دارد. از طرفی اگر در مقاله‌ای با اهمیت، به یک مقاله‌ی دیگر ارجاع داده شده باشد، ارزش آن مقاله‌ی دیگر هم بالا می‌رود، زیرا به نوعی تبلیغ شده است. بنابراین ما انتظار داریم دو عامل در افزایش Page Rank یک سایت اینترنتی تاثیر مثبت داشته باشند:

۱. تعداد بالای سایت‌هایی که به سایت موردنظر پیوند دارند.

۲. وجود سایت‌های با اهمیت، یا با Page Rank بالا که به سایت موردنظر پیوند دارند.

بنابراین تعریف Page Rank به نوعی بازگشتی است و دور دارد؛ زیرا ارزش سایتی بالاتر است که تعداد بیش‌تری سایت با ارزش بالا به آن پیوند داشته باشند. به زبان احتمالات PR (همان Page Rank) یک توزیع احتمال روی شبکه‌ی اینترنت است که نشان می‌دهد اگر شخصی با شروع از یک سایت به تصادف و با کلیک کردن روی پیوندها به تصادف و احتمال برابر، از سایتی به سایت دیگر برود، بعد از مدت نسبتا طولانی با توزیع PR در سایت‌ها خواهد گشت، یعنی با احتمال بیش‌تری در سایت‌های دارای PR بیش‌تر قرار خواهند داشت. (به نسبت $\frac{PR(s)}{\sum_t PR(t)}$ از زمان را در سایت s خواهد گذراند.) بنابراین می‌توان این کار را به منزله‌ی یک فرآیند تصادفی مارکوف نگریست که احتمال گذر از سایتی به سایت دیگر برای تمام پیوندهای درون یک سایت مساوی فرض می‌شود و ما به دنبال توزیع پایا و نهایی این فرآیند هستیم که همان PR است.

توزیع پایا توزیعی است که تحت انجام فرآیند مارکوف بدون تغییر باقی بماند. بنابراین برای هر سایت u، اگر B_u مجموعه‌ی سایت‌هایی باشد که به u پیوند دارند و اگر برای هر سایت w، $L(w)$ را تعداد پیوندهای سایت w به بیرون بگیریم (تکرار مجاز نیست و از هر سایت به سایت دیگری حداکثر یک پیوند داریم و از هیچ سایتی به خودش پیوند نداریم) در این صورت:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} \qquad (۱)$$

زیرا احتمال این که در گام بعد در u باشیم از طرفی $PR(u)$ است و از طرف دیگر، در حال حاضر باید در یکی از سایت‌های B_u باشیم تا بتوانیم در گام بعد به u برویم و برای هر سایت که به u پیوند دارد مانند v، احتمال رفتن از v به u در یک گام $\frac{1}{L(v)}$ است. (قدم زدن تصادفی با گام‌های مستقل و در هر راس به احتمال برابر را در نظر بگیرید.)

برای پاره‌ای از سهولت‌های ریاضی در جهت حل یا محاسبه‌ی تقریبی جواب دستگاه معادلات (۱) یک فرض را اضافه می‌کنیم؛

[1] Google
[2] Search Engine
[3] Link
[4] String Matching
[5] Larry Page
[6] Reference
[7] Citation

فرضی که اضافه می‌کنیم این است که قدم زدن تصادفی روی وب، در هر گام به احتمال d به یکی از پیوندهای موجود می‌رود و به احتمال $1-d$، یک سایت کاملا تصادفی درون شبکه‌ی اینترنت را انتخاب می‌کند و به آن می‌رود. یک دلیل برای چنین فرضی این است که ممکن است بعد از چند کلیک درون سایتی برویم که هیچ پیوندی به بیرون ندارد، در این صورت داخل آن گیر می‌افتیم. برای رفع این مشکل کار را با یک سایت کاملا تصادفی ادامه می‌دهیم. دلیل دیگر افزودن این عامل (به نام Damping Factor) این است که یک کاربر ممکن است بعد از دنبال کردن لینک‌های متوالی خسته شده و به کلی با یک سایت جدید کار را ادامه دهد. دلیل سوم هم این است که معادلاتی که با در نظر گرفتن Damping Factor نوشته می‌شوند از لحاظ تحلیلی بهتر و راحت‌تر حل می‌شوند و از لحاظ محاسباتی نیز بهتر می‌توان PR را با الگوریتم‌هایی تخمین زد. (مثلا در حالت اول که $d=1$ بود ممکن بود فرآیند مارکوف نقاط جاذب داشته باشد یا تناوبی شود و در این دو حالت توزیع پایای خوش‌تعریفی نمی‌شد داد.) پس معادلات به این شکل در می‌آید:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \qquad (2)$$

که p_j سایت iام، N تعداد کل سایت‌ها، d عددی بین صفر و یک (معمولا d را ۰٫۸۵ می‌گیرند، که به صورت تجربی به دست آمده است) و $M(p_i)$ همان B_{p_i} است.

دستگاه N معادله، N مجهول (۲) را به شکل فشرده می‌توان چنین نوشت:

$$R = \frac{1-d}{N}\begin{bmatrix}1\\1\\1\\\vdots\\1\end{bmatrix} + d\begin{bmatrix} l(p_1,p_1) & l(p_1,p_2) & \cdots & l(p_1,p_N) \\ l(p_2,p_1) & l(p_2,p_2) & \cdots & l(p_2,p_N) \\ \vdots & \vdots & \ddots & \vdots \\ l(p_N,p_1) & l(p_N,p_2) & \cdots & l(p_N,p_N) \end{bmatrix} R \qquad (3)$$

که

$$R = \begin{bmatrix} PR(p_1) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

و

$$l(p_i,p_j) = \begin{cases} 0 & \text{اگر } p_j \text{ به } p_i \text{ پیوند نداشته باشد} \\ \frac{1}{L(p_j)} & \text{اگر } p_j \text{ به } p_i \text{ پیوند داشته باشد} \end{cases}$$

بنابراین برای هر j: $\sum_{i=1}^{N} l(p_i,p_j) = 1$.

ماتریس $N \times N$ در (۳) یک ماتریس تصادفی است و اگر جمله‌ی

$$\frac{1-d}{N}\begin{bmatrix}1\\1\\1\\\vdots\\1\end{bmatrix}$$

را نیز با آن ترکیب کنیم به معادله‌ای به شکل $R = AR$ می‌رسیم که A یک ماتریس تصادفی متناظر با یک فرآیند مارکوف غیرتناوبی تحویل‌ناپذیر است. کافی است بنویسیم:

$$R = MR + LR$$

که M ماتریسی است که تمام درایه‌های آن $\frac{1-d}{N}$ است و L همان ماتریس حاوی درایه‌های $l(p_i,p_j)$ است. بنابراین بایستی بردار ویژه‌ی متناظر با مقدار ویژه‌ی ۱ را پیدا کنیم که طبق قضیه‌ای در فرآیندهای تصادفی وجود دارد و یکتاست و درایه‌های آن بردار، مثبت است و می‌تواند در صورت نرمالیزه شدن به عنوان R یا همان Page Rank استفاده شود.

با توجه به این که طیف ماتریس A به نحوی است که فقط چند مقدار ویژه‌ی اول آن از اندازه‌ی قابل توجهی برخوردارند، لذا بردار R (توزیع پایا) با دقت نسبتا بالا و فقط با تعداد کمی مرحله از الگوریتم تکرر، یعنی محاسبه‌ی $v_0, Av_0, A^2v_0, \ldots$ قابل محاسبه است [۱].

یک نقطه‌ی ضعف الگوریتم Page Rank این است که به صفحات قدیمی اهمیت بیش‌تری می‌دهد و یک صفحه‌ی وب جدید و تازه تاسیس حتی در صورت خوب بودن، PR بالایی نمی‌گیرد مگر این که تا حد زیادی محبوب شده باشد و تعدادی از صفحات با PR بالای قدیمی به آن ارجاع داده باشند.

روش‌های تقریبی و محاسباتی زیادی برای محاسبه‌ی سریع و کاربردی Page Rank ارائه شده است [۲] و [۳].

یکی از این الگوریتم‌ها یک الگوریتم توزیعی[۸] و تقریبی ساده و سریع است که از ایده‌ی قدم‌زدن تصادفی استفاده می‌کند [۲]. الگوریتم توزیعی، الگوریتمی است که بتوان پردازش موردنیاز برای آن را بر روی تعدادی پردازنده تقسیم کرد و به طور موازی اجرا کرده و سپس نتایج هر بخش را گرفته و با هم ترکیب کرد و جواب نهایی را به دست آورد. (مثلا می‌توان مرتب‌سازی تعداد زیادی عدد را به روش توزیعی انجام داد که خود مساله‌ی جالبی است.) این الگوریتم در $O(\frac{\log n}{\epsilon})$ مرحله، با احتمال زیاد Page Rank را محاسبه می‌کند که n تعداد کل سایت‌ها و ϵ همان d در (۲) است. الگوریتم‌هایی با هزینه‌ی $O(\frac{\sqrt{\log n}}{\epsilon})$ و $O(\sqrt{\frac{\log n}{\epsilon}})$ نیز در شرایط خاص دیگری وجود

[۸] Distributed

دارد. روش دیگر برای محاسبه‌ی تقریبی PR، پیدا کردن بردار ویژه‌ی ماتریس با روش تکرری توانی است. روش دیگری هم می‌شد برای محاسبه‌ی PR به کار برد و آن پیدا کردن بردار ویژه‌ی نظیر مقدار ویژه‌ی ۱ از طریق حل دستگاه معادلات مذکور است. ولی این روش دارای کاربرد عملی نیست، زیرا حل دستگاه N معادله و N مجهول، یا وارون کردن یک ماتریس $N \times N$ وقتی N از مرتبه‌ی ۱.۵ میلیارد صفحه‌ی وب است، چندان کار به صرفه‌ای نیست و علاوه بر هزینه‌ی بالای محاسباتی و زمانی، ممکن است به دلیل خطای محاسباتی در پیدا کردن وارون ماتریس و پایدار نبودن الگوریتم‌های وارون‌سازی ماتریس، PR به کلی غلط به دست بیاید. روش تکرری توانی برای این مقصود بهتر است که در آن از توزیع اولیه‌ی v_0، مثلا توزیع یکنواخت شروع می‌کنیم و دنباله‌ی $v_0, Av_0, A^2v_0, \ldots$ را محاسبه می‌کنیم. بعد از تعداد کمی مرحله، مثلا k مرحله طبق قضیه‌ای در فرآیندهای تصادفی با تقریب خوبی $A^k v_0$ با R برابر می‌شود و همین روش تکرری نیز الگوریتم‌های توزیعی دارد.

تلاش‌هایی در راستای دستکاری Page Rank

برخی مدیران سایت‌ها تلاش می‌کنند PR خود را به شکل مصنوعی بالا ببرند، مثلا تعداد زیادی پیوند از سایت‌های بی‌اهمیت یا ضعیف به سایت خود ایجاد کنند. هم‌چنین نوعی کسب و کار پدید آمده که در آن برخی سایت‌های مهم و پربازدید و با PR بالا، با گرفتن مبالغی، پیوندهایی به سایت‌هایی که به دنبال کسب PR بالاتر در موتور جستجوی گوگل هستند، ایجاد می‌کنند. خود گوگل با این کار مخالف است و معتقد است کیفیت موتور جستجو را کاهش می‌دهد و تهدید کرده است که این نوع پیوندها را در صورت کشف در نتیجه‌ی جستجو بی‌اثر خواهد ساخت.

البته خود گوگل کاملا به Page Rank خود برای مرتب‌سازی سایت‌ها وفادار نیست و بعضی از سایت‌ها را با دید تبلیغاتی و با گرفتن دستمزد در میان نتایج جستجو نشان می‌دهد، هم‌چنین گوگل از روش‌های دیگری نیز برای بهبود و ارتقاء نتایج جستجو استفاده می‌کند، به ویژه روش‌هایی برای تشخیص مزرعه‌های پیوندی[۹] (مجموعه سایت‌هایی که به طور بیش از حد و غیرعادی و احتمالا عمدی به هم پیوند دارند) و خرید و فروش پیوند. در ضمن به این نکته باید اشاره کرد که گوگل تنها به این الگوریتم اکتفا نکرده و از روش‌های بسیار متنوعی در علوم مختلف برای بهبود نتایج جستجو استفاده می‌کند و خیلی از این روش‌ها تجربی است، در ضمن برخی از راه‌کارهای گوگل مخصوصا قسمت‌های علمی و مربوط به پیاده‌سازی و مبارزه با تقلب، محرمانه است.

[۹] Link Farms

مراجع

[1] Taher Haveliwala and Sepandar Kamvar, The Second Eigenvalue of the Google Matrix, Stanford University Techical Report:7056, March 2003.

[2] Atish Das Sarma, Anisur Rahaman Molla, Gopal Pandurangan and Eli Upfal, Fast Distributed PageRank Computation, 2012.

[3] Gianna M. Del Corso, Antonio Gull and Francesco Romani, Fast PageRank Computation via a Sparse Linear System, Internet Mathematics, Lecture notes in Computer Science 2(3):118, 2005.

[4] S. Brin and L. Page, The Anatomy of Large-Scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems, ISSN 0169-7552, 1998.

[5] David Rise and Mark Malseed, The Google Story, ISBN 0553-80457-x, 2005.

ترجمه‌ی فارسی این کتاب با نام سرگذشت شگفت‌انگیز گوگل، ترجمه‌ی سینا قربانلو توسط انتشارات مبلغان به چاپ رسیده است.

[6] L. Page, S. Brin, R. Motwani and T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, 1999.