線形代数と Google PageRank　補足資料

担当 TA： 横山 俊一（九州大学大学院数理学府：修士 2 年）

※ 小テストの内容とは一切関係ありませんが、興味のある方はご覧頂ければ幸いです。

小テストも残すところあと数回となりました。皆さんが学習した内容は線形代数の入り口であって、この先にはもっともっと広大なフィールドが待ち構えています。一生関わることのない方がほどんどかもしれませんが、何かの縁で再び数学を勉強する人もいるかもしれませんね。

この小文では**「線形代数って何の役に立つの？」「日常生活の何に活かされているの？」**といった素朴な疑問に対して、一つ有名な例を紹介したいと思います。

＊＊＊

さて、皆さんは毎日のようにインターネットを使っていると思いますが、そこで無くてはならないのが**「検索エンジン」**です。特に圧倒的なシェアを有するのが Google（グーグル[1]）であり、以下のように調べたいキーワードを入力するだけで、情報集めの手助けをしてくれます。

では、一つに「線形代数・参考書」と言っても数十万あるページの「順位」をどうやって決めているのでしょうか？ … 実はここに線形代数の基礎が活かされているのです !!

そもそも Google がこれ程までに大きな企業へと成長出来たのは、同様の検索エンジンを提供している他社に比べて卓越した検索技術を持っていたからに他なりません。これは **PageRank** と呼ばれているシステムで、同じキーワードでも「皆が知りたい・読みたいと思われる情報」をピックアップすることに非常に長けていることで知られています。実はこの PageRank システムは、線形代数の**固有値・固有ベクトル**の考え方を使って設計されているのです。皆さんのテキストでは 5 章と 6 章にあたるもので、今期の講義では詳しく勉強することは無いかもしれませんが、それほど難しいものでもありません。一度目を通してみるのもいいかもしれませんね。

[1]実は日本だけのシェアでは Yahoo!（ヤフー）が首位、グーグルは 2 位です。そういえば先日の「シルシルミシル」（テレビ朝日系列）という深夜番組でも面白い特集があっていました。ちなみに僕は堀君の大ファンです。

さて、少しだけ PageRank について紹介しておきましょう。ネット上のありとあらゆるページには「リンク」がはられ、ページ間を行き来出来るようになっています（ネットサーフィンをする人にとっては一つの楽しみかもしれませんね)。そこで、ページ間のリンクを次のように定式化してみます。

- サイト A から B へリンクがはられている時、B にはポイントが加算される。
- ここで更にサイト A の信頼性も評価しておき、それに応じてポイントを変化させておく。
- すると、多くの「良いページ」からリンクされているほど「良いページ」であると判断できる。
- この意味で「良いページ」の順番に（＝ポイントの高い順番に）表示していく。

これで順位の意味付けが出来ました。ではここからは簡単な例を使って説明しましょう。

＊＊＊

今、簡単の為ネット上には 5 つしかページが無いと仮定しましょう。ここで、ページ 1 からページ 5 からは次のようにリンクがはられている、という状態を考えます。

- ページ 1： ページ 2、ページ 4
- ページ 2： ページ 1、ページ 3、ページ 5
- ページ 3： ページ 1
- ページ 4： ページ 2、ページ 3
- ページ 5： ページ 1

また、ページ i（$1 \leq i \leq 5$）のポイントを p_i としておき、ページ i からリンクされることによって与えられるポイントは

$$\frac{\text{ページ } i \text{ のポイント}}{\text{ページ } i \text{ からのリンク数}}$$

で与えられるとしましょう。つまり、ページ 2 からリンクされれば $p_2/3$ ポイント、ページ 3 からリンクされればまるまる p_3 ポイントを得る、という具合です。この時、ページ i へのリンクによって与えられるポイントを P_i とすると、次の等式が成り立ちます。

$$\begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{3} & 1 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix}$$

さて、ここで重要なのは最初のポイント p_i とリンクによって与えられるポイント P_i との相対的関係が等しくなくてはならない、ということです。つまり、ポイント達の比が等しいということですから、定数 λ を使って次のように表現出来ることになりますね。

$$\begin{bmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \end{bmatrix} = \lambda \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{3} & 1 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix}$$

というわけで、残すは各ページのランク付け[2]、即ち $p_1, \cdots, p_5$ を求めれば良いことになります。そこで分かりやすいように

$$T = \begin{bmatrix} 0 & \frac{1}{3} & 1 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & 0 \end{bmatrix}, \quad u = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix}$$

とおけば、先程の式は単純明快に

$$Tu = \lambda u$$

と書けてしまいます。では、テキスト p.98 の中ほどをご覧ください。… まさに固有値と固有ベクトルそのものですね。つまり λ とは行列 T の固有値のこと、ベクトル u はその固有ベクトルのことなのです。

＊＊＊

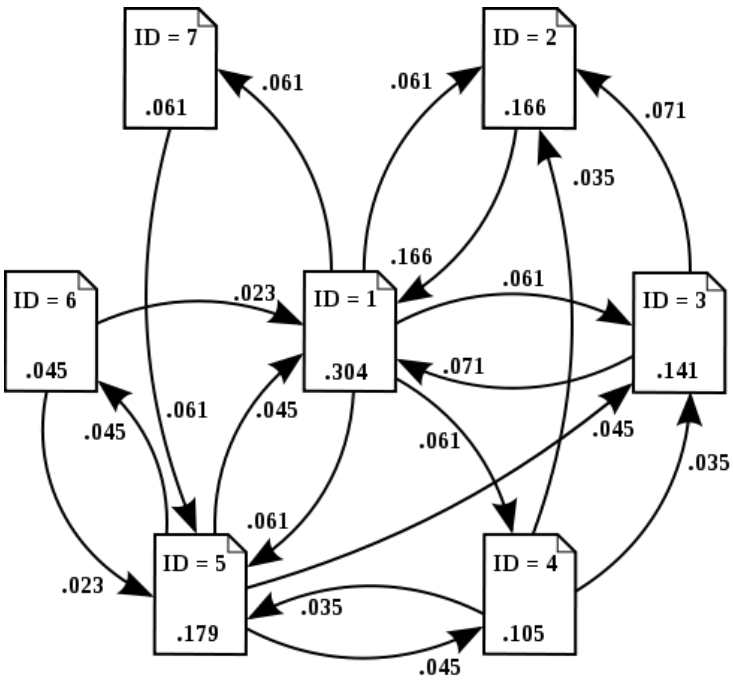
まとめると、Google の検索システムは

「ホームページの様子を行列で表し、その固有値・固有ベクトルを求める」

ということをしているに過ぎない、ということです。しかし「なぁんだ、1 年生のテキストに載ってる程度の知識しか使わないのかぁ。じゃあ大したことないじゃん」と思ったそこのアナタはちょっと甘い。Google はこれをバックボーンとして

- 超巨大行列（全世界のページをコントロールする）を高速に計算するアルゴリズム
- 各ページからキーワードを的確に検出する能力

を最大限に進化させた結果今の地位を築き上げたのですから、その労力ははかり知れません。しかしその根源にあるのは意外と素朴なアイデアですから、いつか皆さんが新しい発見をする日が来るかもしれませんね !!



[2] これが PageRank の名前の由来である。