INGENJÖRSHÖGSKOLAN
Matematik
Fredrik Abrahamsson

Google och egenvektorer

Alla inser idag betydelsen av att ha tillgång till en bra sökmetod för att hitta websidor. När vi söker efter en websida vill vi få träffar som är relevanta för vårt sökord. Vad betyder detta? Naturligtvis måste websidan på något sätt innehålla vårt sökord men om det är många sidor som gör det, hur skall vi gradera dem? I ett tidigt skede så räknade sökmotorerna helt enkelt bara hur många gånger det aktuella sökordet förekommer, men denna metod ger ofta mycket dåliga resultat. Vi vill på något sätt kunna avgöra *hur* pass viktig en websida är. En idé är att beräkna hur många *andra* websidor som länkar till den aktuella websidan, men denna metod missar många sidor som kan vara väldigt viktiga utan att för den skull ha ett stort antal länkar till sig. En mer sofistikerad variant av denna idé används av sökmotorn Google (www.google.se). Google mäter graden av viktighet genom att använda ett system som kallas för PageRank och utvecklades av grundarna till Google. På googles hemsida kan man läsa att

> PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important."

Lite löst kan man alltså säga att en websida är viktig om andra viktiga websidor har länkar till den. Detta låter som en cirkeldefinition men låt oss försöka se vad denna tanke kan användas till.

Viktade viktigheter

Antag att vi numrerar alla websidor från 1 till n och låter låter x_k vara ett mått på hur pass viktig websida nummer k är. Tanken med PageRank ovan är då att talet x_k är proportionellt mot summan av alla x_i sådana att sida i har en länk till sida k. Alltså har vi ett system av ekvationer som kan se ut ungefär som

$$\begin{cases} x_1 = \alpha(x_{23} + x_{341} + x_{4525}) \\ x_2 = \alpha(x_3 + x_{634} + x_{346347} + x_{8286352}) \\ \vdots \end{cases}$$

Detta är ett (gigantiskt) linjärt ekvationssystem! Låt nu

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

och definiera $A = (a_{ij})$ som den $n \times n-$matris som uppfyller att

$$a_{ij} = \begin{cases} 1 & \text{om sida } i \text{ länkar till sida } j \\ 0 & \text{annars} \end{cases}$$

Då kan ekvationssystemet ovan skrivas som

$$X = \alpha AX$$

dvs. ekvationerna beskriver ett egenvärdesproblem! Vektorn X som vi vill ta reda på är tydligen en egenvektor till matrisen A hörande till egenvärdet $\lambda = 1/\alpha$. För att beräkna X kan man ta reda på samtliga egenvektorer till A och sedan välja X som en egenvektor med enbart positiva element. Om vi har beräknat X så är det sedan lätt att rangordna websidorna i fråga om viktighet; den viktigaste websidan är den sida k som hör ihop med det största talet x_k i vektorn X. På så sätt kan Google ordna sidorna vid en sökning så att den som letar får hjälp med att rangordna sidorna och därmed sparar tid.

Är HV bäst?

Denna idé för att rangordna objekt med hjälp av egenvektorer går tillbaka till 50-talet och har fått allt större betydelse i takt med den ökade graden av webapplikationer. Det finns dock många andra situationer där det kan vara intressant att göra en "intelligent" ranking enligt metoden ovan. Ta t.ex. en turnering där ett antal lag har mött varandra ett antal gånger. Om x_k är styrkan hos lag k och om vi antar att styrkan hos ett lag k är proportionellt mot styrkan i de lag som lag k har besegrat så återfår vi ett ekvationssystem av formen

$$X = \alpha A X$$

där $A = (a_{ij})$ och a_{ij} är antalet gånger som lag i besegrat lag j. Låt oss ta ett litet exempel. Antag att vi har 5 lag som har kämpat i en serie där alla mött varandra två gånger - en hemmamatch och en bortamatch. Matrisen A fick då följande utseende;

$$A = \begin{pmatrix} 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 2 & 1 & 0 & 0 & 0 \\ 2 & 0 & 2 & 0 & 1 \\ 1 & 1 & 2 & 1 & 0 \end{pmatrix}$$

Här ser vi t.ex. att lag 1 har vunnit två gånger mot lag 2 och 1 gång mot lag 5 osv. Beräkning av egenvektorerna till A ger att

$$X = \begin{pmatrix} 0.38 \\ 0.49 \\ 0.32 \\ 0.50 \\ 0.52 \end{pmatrix}.$$

Alltså är lag 5 det lag som rankas högst och därefter följer i fallande ordning lag 4, lag 2, lag 1 och sist lag 3. Observera att både lag 5 och lag 4 vunnit lika många matcher, men lag 5 har vunnit mot starkare lag jämfört med lag 4. Likaså har lag 1 och lag 3 vunnit lika många matcher men eftersom lag 1 vann en match mot vinnaren lag 5 så rankas lag 1 högre.