

# Linear Algebra behind Google: notes

Algirdas Grybas

October 25, 2006

General functions of a web search engine:

1. Locate web pages with public access
2. Index the data to make it searchable
3. Rate the importance of each page in the database

Google's way of rating - **PageRank**: ranks the importance of web pages according to an eigenvector of a weighted link matrix.

**Definition 1** A number  $x_k \in \mathbb{R}$ ,  $x_k \geq 0$  that quantitatively rates importance of a web page  $k$  is called the **importance score**. It is derived from the links made to the page from other web pages.

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j}$$

**Definition 2 Backlink** of a page is a link to that page from another one.

**Definition 3** A **graph** is a set of vertices and a set of edges. A **directed graph** is a graph where each edge has a direction, i.e. a starting and an ending vertex.

Problem: *Find  $x_k$  for a given directed graph  $\Rightarrow$  find an eigenvector with eigenvalue 1 for a square matrix.*

**Definition 4 Dangling node** - a page with no outgoing links.

**Definition 5** A square matrix is called a **column-stochastic** matrix if all of its entries are non-negative and the entries in each column sum to 1. A matrix is **column-substochastic** if the column sums are all less than or equal to 1.

**Proposition 1** Every column-stochastic matrix has 1 as an eigenvalue.

Notation:  $V_\lambda(A)$  is the eigenspace for the eigenvalue  $\lambda$  of a column-stochastic matrix  $A$ .

Two problems with finding eigenvectors:

1. Nonunique rankings -  $\dim(V_1(A))$  may not be equal to 1, i.e. the eigenvector  $x$  with  $\sum_i x_i = 1$  is not unique.

2. Dangling nodes -  $A$  is column-substochastic, and so may not have 1 as an eigenvalue. Solution: use positive eigenvalue  $\lambda \leq 1$  and a corresponding eigenvector  $x$  (Perron eigenvector).

If a web is strongly connected (one can get from any page to any other page in a finite number of steps),  $V_1(A) = 1$ . If a web  $W$ , considered as an undirected graph, is not strongly connected, it consists of  $r$  disconnected subwebs  $W_1, \dots, W_r$ .

**Proposition 2** For a disconnected graph  $W$  as above,  $\dim(V_1(A)) \geq r$ .

We need a modified link matrix that would ensure unambiguous importance scores for an  $n$ -page web with no dangling nodes, but with possibly multiple subwebs.

Replace the link matrix  $A$  with the matrix

$$M = (1 - m)A + mS$$

where  $0 \leq m \leq 1$  (Google:  $m = 0.15$ ) and  $S$  is an  $n \times n$  matrix with all entries  $1/n$ .  $M$  is obviously column-stochastic for  $m \in [0, 1]$ . Using  $M$  instead of  $A$  gives a web page with no backlinks (dangling node) the importance score of  $\frac{m}{n}$ , and the matrix  $M$  is column-substochastic for any  $m < 1$  since the matrix  $A$  is column-substochastic. Matrix  $M$  allows us to compare pages in different subwebs. Now the equation  $Mx = x$  can also be written as

$$x = (1 - m)Ax + ms$$

where  $s$  is a column vector with all entries  $1/n$ .

We are left to prove that  $\dim(V_1(M)) = 1$  if  $m \in (0, 1]$ .

**Definition 6** A matrix  $M$  is **positive** if  $M_{ij} > 0$  for all  $i, j$ .

**Proposition 3** If  $M$  is positive and column-stochastic, then any eigenvector  $x \in V_1(M)$  has all positive or all negative components.

**Proposition 4** Let  $v, w \in \mathbb{R}^m$ ,  $m \geq 2$ , be linearly independent vectors. Then for some values  $s, t \in \mathbb{R}$ , the vector  $x = sv + tw$  has both positive and negative components.

**Proposition 5** If  $M$  is positive and column-stochastic, then  $\dim(V_1(M)) = 1$ .

Thus, we are guaranteed the existence of a unique eigenvector  $x \in V_1(M)$  with positive components such that  $\sum_i x_i = 1$ .

Google uses **power method** to calculate the importance score eigenvector. We prove that the method converges for any starting value.

**Proposition 6** Let  $M$  be a positive column-stochastic  $n \times n$  matrix and let  $V$  denote the subspace of  $\mathbb{R}^n$  consisting of vectors  $v$  such that  $\sum_j v_j = 0$ . Then  $Mv \in V$  for any  $v \in V$ , and

$$\|Mv\|_1 \leq c \|v\|_1$$

for any  $v \in V$ , where  $c = \max_{1 \leq j \leq n} |1 - 2 \min_{1 \leq i \leq n} M_{ij}| < 1$ .

**Proposition 7** Every positive column-stochastic matrix  $M$  has a unique vector  $q$  with positive components such that  $Mq = q$  with  $\|q\|_1 = 1$ . The vector  $q$  can be computed as  $q = \lim_{k \rightarrow \infty} M^k x_0$  for any initial guess  $x_0$  with positive components such that  $\|x_0\|_1 = 1$ .