The Mathematics of Internet Search Engines

David Marshall

Department of Mathematics
Monmouth University

April 4, 2007

Outline

Mathematics of Information Retrieval

Information retrieval methods such as

- traditional vector space methods (keyword searches)
- Google's PageRank
- HITS (Ask.com)
- SALSA

all proveide

Mathematics of Information Retrieval

Information retrieval methods such as

- traditional vector space methods (keyword searches)
- Google's PageRank
- HITS (Ask.com)
- SALSA

all proveide

- relevant, modern application of mathematics

Mathematics of Information Retrieval

Information retrieval methods such as

- traditional vector space methods (keyword searches)
- Google's PageRank
- HITS (Ask.com)
- SALSA

all proveide

- relevant, modern application of mathematics
- aspects which are accessible early in the curriculum

Mathematics of Information Retrieval

Information retrieval methods such as

- traditional vector space methods (keyword searches)
- Google's PageRank
- HITS (Ask.com)
- SALSA

all proveide

- relevant, modern application of mathematics
- aspects which are accessible early in the curriculum
- other aspects which provide appropriate investigations at intermediate and advanced levels

Outline

The Olden Days - 1990's

The internet was still quite new to most people in 1994.

The Olden Days - 1990's

The internet was still quite new to most people in 1994.

- America Online gave many a “friendly” introduction.

The Olden Days - 1990's

The internet was still quite new to most people in 1994.

- America Online gave many a "friendly" introduction.
- Internet searching was basically a game of luck.

The Olden Days - 1990's

The internet was still quite new to most people in 1994.

- America Online gave many a "friendly" introduction.
- Internet searching was basically a game of luck.
- excite LYCOS

The Olden Days - 1990's

The internet was still quite new to most people in 1994.

- America Online gave many a "friendly" introduction.
- Internet searching was basically a game of luck.
- excite LYCOS
- altavista HOTBOT

The year 1996

At Stanford

At IBM's Almaden Research Center

The year 1996

At Stanford

- Larry Page and Sergey Brin, two CS graduate students

At IBM's Almaden Research Center

The year 1996

At Stanford

- Larry Page and Sergey Brin, two CS graduate students

- The result - PageRank and Google

At IBM's Almaden Research Center

The year 1996

At Stanford

- Larry Page and Sergey Brin, two CS graduate students

- The result - PageRank and Google

At IBM's Almaden Research Center

- Jon Kleinberg

The year 1996

At Stanford

- Larry Page and Sergey Brin, two CS graduate students

- The result - PageRank and Google

At IBM's Almaden Research Center

- Jon Kleinberg

- The result - HITS (Hyperlink-Induced Topic Search), Ask.com

Outline

Internet Keyword Searches

- Early internet search engines were based solely on keyword searches.
- Low reliability of search returns (why?).

Internet Keyword Searches

- Early internet search engines were based solely on keyword searches.
- Low reliability of search returns (why?).
- No *review* process.

Internet Keyword Searches

- Early internet search engines were based solely on keyword searches.
- Low reliability of search returns (why?).
- No *review* process.
- Nothing to guarantee the quality of a site.

Internet Keyword Searches

- Early internet search engines were based solely on keyword searches.
- Low reliability of search returns (why?).
- No *review* process.
- Nothing to guarantee the quality of a site.
- Very susceptible to manipulation (how?).

A Novel Idea

- Keyword search model viewed the web as a large library, with webpages acting as books on a shelf.
- Keyword search model ignores an integral aspect of the web.

A Novel Idea

- Keyword search model viewed the web as a large library, with webpages acting as books on a shelf.
- Keyword search model ignores an integral aspect of the web.
- The web consists of two sets: its pages *and* its links.

A Novel Idea

- Keyword search model viewed the web as a large library, with webpages acting as books on a shelf.
- Keyword search model ignores an integral aspect of the web.
- The web consists of two sets: its pages *and* its links.
- Page-Brin-Kleinberg: use the organic and social link structure of the web in order to *rank* pages.

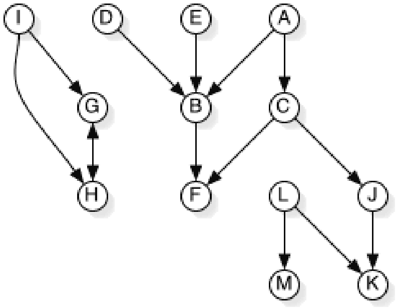
A Novel Idea

- Keyword search model viewed the web as a large library, with webpages acting as books on a shelf.
- Keyword search model ignores an integral aspect of the web.
- The web consists of two sets: its pages *and* its links.
- Page-Brin-Kleinberg: use the organic and social link structure of the web in order to *rank* pages.
- The web is modeled as a *directed graph*

Directed Graphs

A *directed graph* is a collection of nodes (the web pages) together with a collection of arrows pointing from one node to another (the links).

Example



Application for the Classroom - 1

Graph theory has become a common topic introduced in

- mathematics survey courses for non-science, non-engineering majors.
- discrete mathematics courses.

Typical focus is on routing and scheduling problems:

The Web provides another very relevent, useful model utilizing directed graphs.

Application for the Classroom - 1

Graph theory has become a common topic introduced in

- mathematics survey courses for non-science, non-engineering majors.
- discrete mathematics courses.

Typical focus is on routing and scheduling problems:

- Euler circuits and paths

The Web provides another very relevent, useful model utilizing directed graphs.

Application for the Classroom - 1

Graph theory has become a common topic introduced in

- mathematics survey courses for non-science, non-engineering majors.
- discrete mathematics courses.

Typical focus is on routing and scheduling problems:

- Euler circuits and paths
- traveling salesman problems

The Web provides another very relevent, useful model utilizing directed graphs.

Application for the Classroom - 1

Graph theory has become a common topic introduced in

- mathematics survey courses for non-science, non-engineering majors.
- discrete mathematics courses.

Typical focus is on routing and scheduling problems:

- Euler circuits and paths
- traveling salesman problems
- networks and critical paths

The Web provides another very relevent, useful model utilizing directed graphs.

Application for the Classroom - 1

Graph theory has become a common topic introduced in

- mathematics survey courses for non-science, non-engineering majors.
- discrete mathematics courses.

Typical focus is on routing and scheduling problems:

- Euler circuits and paths
- traveling salesman problems
- networks and critical paths

The Web provides another very relevent, useful model utilizing directed graphs.

- diameter of the WWW

Application for the Classroom - 1

Graph theory has become a common topic introduced in

- mathematics survey courses for non-science, non-engineering majors.
- discrete mathematics courses.

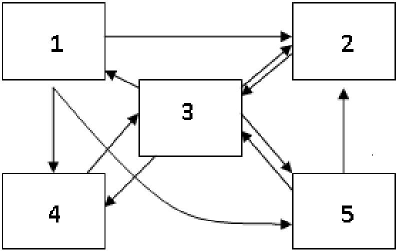
Typical focus is on routing and scheduling problems:

- Euler circuits and paths
- traveling salesman problems
- networks and critical paths

The Web provides another very relevent, useful model utilizing directed graphs.

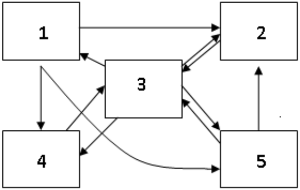
- diameter of the WWW
- search engine page rankings

The Mathematics of Google's PageRank



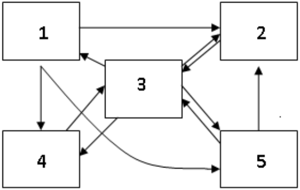
Each page is assigned a *rank*, which is a numerical value between 0 and 1. Then, when a keyword search is performed, results are returned according to their ranks, with higher ranks returned first.

PageRank Example



Each page inherits its rank from those sites linking to it. So, for example, Page 2 gets:

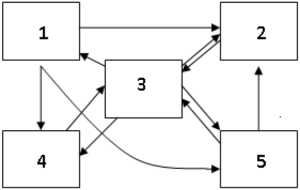
PageRank Example



Each page inherits its rank from those sites linking to it. So, for example, Page 2 gets:

- $\frac{1}{3}$ of Page 1's rank

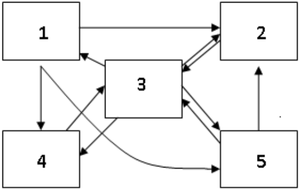
PageRank Example



Each page inherits its rank from those sites linking to it. So, for example, Page 2 gets:

- $\frac{1}{3}$ of Page 1's rank
- $\frac{1}{4}$ of Page 3's rank

PageRank Example



Each page inherits its rank from those sites linking to it. So, for example, Page 2 gets:

- $\frac{1}{3}$ of Page 1's rank
- $\frac{1}{4}$ of Page 3's rank
- $\frac{1}{2}$ of Page 5's rank

PageRank Example

Letting r_i denote the rank of page i, $i = 1, \ldots, 5$, we obtain the following system of linear equations:

PageRank Example

Letting r_i denote the rank of page i, $i = 1, \ldots, 5$, we obtain the following system of linear equations:

$$
\begin{aligned}
0r_1 + 0r_2 + 1/4r_3 + 0r_4 + 0r_5 &= r_1 \\
1/3r_1 + 0r_2 + 1/4r_3 + 0r_4 + 1/2r_5 &= r_2 \\
0r_1 + 1r_2 + 0r_3 + 1r_4 + 1/2r_5 &= r_3 \\
1/3r_1 + 0r_2 + 1/4r_3 + 0r_4 + 0r_5 &= r_4 \\
1/3r_1 + 0r_2 + 1/4r_3 + 0r_4 + 0r_5 &= r_5
\end{aligned}
$$

Application for the Classroom - 2

Solutions to systems of linear equations are often covered in

Google's PageRank algorithm provides a relevent and modern example of a problem whose solution is given by such a system.

Application for the Classroom - 2

Solutions to systems of linear equations are often covered in

- college algebra

Google's PageRank algorithm provides a relevent and modern example of a problem whose solution is given by such a system.

Application for the Classroom - 2

Solutions to systems of linear equations are often covered in

- college algebra
- finite mathematics

Google's PageRank algorithm provides a relevent and modern example of a problem whose solution is given by such a system.

Application for the Classroom - 2

Solutions to systems of linear equations are often covered in

- college algebra
- finite mathematics
- mathematical modeling (in, for example, the social or biological sciences)

Google's PageRank algorithm provides a relevent and modern example of a problem whose solution is given by such a system.

Application for the Classroom - 2

Solutions to systems of linear equations are often covered in

- college algebra
- finite mathematics
- mathematical modeling (in, for example, the social or biological sciences)

Google's PageRank algorithm provides a relevent and modern example of a problem whose solution is given by such a system.

- too big to solve by hand

Application for the Classroom - 2

Solutions to systems of linear equations are often covered in

- college algebra
- finite mathematics
- mathematical modeling (in, for example, the social or biological sciences)

Google's PageRank algorithm provides a relevent and modern example of a problem whose solution is given by such a system.

- too big to solve by hand
- good opportunity to discuss the role of technology

PageRank Example

We may rewrite the previous system of linear equations in matrix form.

$$\begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/2 \\ 0 & 1 & 0 & 1 & 1/2 \\ 1/3 & 0 & 1/4 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \end{bmatrix}$$

PageRank Example

We may rewrite the previous system of linear equations in matrix form.

$$\begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/2 \\ 0 & 1 & 0 & 1 & 1/2 \\ 1/3 & 0 & 1/4 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \end{bmatrix}$$

In this example, the system has a 1-parameter family of solutions. There is a unique solution whose entries are positive and sum to 1. This vector is called the *pagerank vector*, and is given by

$$\begin{bmatrix} .103 & .207 & .414 & .138 & .138 \end{bmatrix}$$

Application for the Classroom - 3

Linear Algebra

Application for the Classroom - 3

Linear Algebra

- Writing the system as $Hx = x$, we clearly have an eigenvector problem.

Application for the Classroom - 3

Linear Algebra

- Writing the system as $Hx = x$, we clearly have an eigenvector problem.
- At this point, we encounter several "why"'s.

Application for the Classroom - 3

Linear Algebra

- Writing the system as $Hx = x$, we clearly have an eigenvector problem.
- At this point, we encounter several "why"'s.
- But at a minimum, the PageRank algorithm gives a new example of a problem whose solution is provided by an eigenvector of a matrix.

The Hyperlink Matrix

The matrix obtained on the previous slide is called the *hyperlink matrix* of the web:

$$H = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/2 \\ 0 & 1 & 0 & 1 & 1/2 \\ 1/3 & 0 & 1/4 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 \end{bmatrix}$$

The Hyperlink Matrix

The matrix obtained on the previous slide is called the *hyperlink matrix* of the web:

$$H = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/2 \\ 0 & 1 & 0 & 1 & 1/2 \\ 1/3 & 0 & 1/4 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 \end{bmatrix}$$

- completely describes the link structure of the web.

The Hyperlink Matrix

The matrix obtained on the previous slide is called the *hyperlink matrix* of the web:

$$H = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/2 \\ 0 & 1 & 0 & 1 & 1/2 \\ 1/3 & 0 & 1/4 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 \end{bmatrix}$$

- completely describes the link structure of the web.
- The pagerank vector is a particular eigenvector associated to the eigenvalue 1.

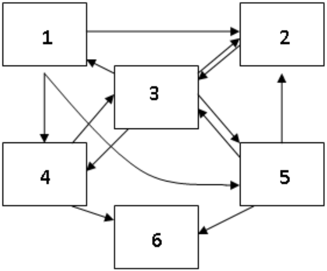
Application for the Classroom - 4

Why does the the system $Hx = x$ have a solution?

Application for the Classroom - 4

Why does the the system $Hx = x$ have a solution?

Example



The Dangling Node

The hyperlink matrix for the web on the previous slide is

$$H = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/3 & 0 \\ 0 & 1 & 0 & 1/2 & 1/3 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/3 & 0 \end{bmatrix}$$

The Dangling Node

The hyperlink matrix for the web on the previous slide is

$$H = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/3 & 0 \\ 0 & 1 & 0 & 1/2 & 1/3 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/3 & 0 \end{bmatrix}$$

- lack of outlinks from Site 6 corresponds to the column of 0's.

The Dangling Node

The hyperlink matrix for the web on the previous slide is

$$H = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/3 & 0 \\ 0 & 1 & 0 & 1/2 & 1/3 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/3 & 0 \end{bmatrix}$$

- lack of outlinks from Site 6 corresponds to the column of 0's.
- this effects on the nature of solutions to the equation $Hx = x$.
- matrix is no longer column stochastic, for example.

The Dangling Node

The hyperlink matrix for the web on the previous slide is

$$H = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/3 & 0 \\ 0 & 1 & 0 & 1/2 & 1/3 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/3 & 0 \end{bmatrix}$$

- lack of outlinks from Site 6 corresponds to the column of 0's.
- this effects on the nature of solutions to the equation $Hx = x$.
- matrix is no longer column stochastic, for example.
- Google's solution:

The Dangling Node

The hyperlink matrix for the web on the previous slide is

$$H = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/3 & 0 \\ 0 & 1 & 0 & 1/2 & 1/3 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/3 & 0 \end{bmatrix}$$

- lack of outlinks from Site 6 corresponds to the column of 0's.
- this effects on the nature of solutions to the equation $Hx = x$.
- matrix is no longer column stochastic, for example.
- Google's solution:
  - view H as a matrix of probabilities

The Dangling Node

The hyperlink matrix for the web on the previous slide is

$$H = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 1/3 & 0 \\ 0 & 1 & 0 & 1/2 & 1/3 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/3 & 0 \end{bmatrix}$$

- lack of outlinks from Site 6 corresponds to the column of 0's.
- this effects on the nature of solutions to the equation $Hx = x$.
- matrix is no longer column stochastic, for example.
- Google's solution:
  - view H as a matrix of probabilities
  - consider not just the action of following links, but also of directly typing in a URL.

The Dangling Node

Google's PageRank algorithm replaces the previous H with a "tweaked" matrix

$$S = \begin{bmatrix} 0 & 0 & 1/4 & 0 & 0 & 1/6 \\ 1/3 & 0 & 1/4 & 0 & 1/3 & 1/6 \\ 0 & 1 & 0 & 1/2 & 1/3 & 1/6 \\ 1/3 & 0 & 1/4 & 0 & 0 & 1/6 \\ 1/3 & 0 & 1/4 & 0 & 0 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/3 & 1/6 \end{bmatrix}$$

where each $1/6$ in the last column represents the probability of randomly visiting one of the 6 pages by directly typing in its URL.

The Google Matrix

The Google Matrix

- Let E denote the $n \times n$ matrix all of whose entries are $\frac{1}{n}$.

The Google Matrix

- Let E denote the $n \times n$ matrix all of whose entries are $\frac{1}{n}$.
- Let $0 \leq \alpha \leq 1$.

The Google Matrix

- Let E denote the $n \times n$ matrix all of whose entries are $\frac{1}{n}$.
- Let $0 \leq \alpha \leq 1$.
- Then define
$$G = \alpha S + (1 - \alpha)E.$$

The Google Matrix

- Let E denote the $n \times n$ matrix all of whose entries are $\frac{1}{n}$.
- Let $0 \leq \alpha \leq 1$.
- Then define
$$G = \alpha S + (1 - \alpha) E.$$
- Google's last public disclosure was $\alpha = .85$.

The Google Matrix

- Let E denote the $n \times n$ matrix all of whose entries are $\frac{1}{n}$.
- Let $0 \leq \alpha \leq 1$.
- Then define
$$G = \alpha S + (1 - \alpha)E.$$
- Google's last public disclosure was $\alpha = .85$.
- G is a positive, column stochastic matrix (this is good).

Application for the Classroom - 5

Some obvious (and important) questions:

Application for the Classroom - 5

Some obvious (and important) questions:

- Why does $Gx = x$ have a solution?

Application for the Classroom - 5

Some obvious (and important) questions:

- Why does $Gx = x$ have a solution?
- Why does it have a relevant solution?

Application for the Classroom - 5

Some obvious (and important) questions:

- Why does $Gx = x$ have a solution?
- Why does it have a relevant solution?
- Why does it have a unique relevant solution?

Application cont'd

Theorem (Perron-Frobenius)

Let $A > 0$ *with* $r = \rho(A)$. *Then the following are true.*

Application cont'd

Theorem (Perron-Frobenius)

Let $A > 0$ with $r = \rho(A)$. Then the following are true.

- $r > 0$.

Application cont'd

Theorem (Perron-Frobenius)

Let $A > 0$ *with* $r = \rho(A)$. *Then the following are true.*

- $r > 0$.
- $r \in \sigma(A)$.

Application cont'd

Theorem (Perron-Frobenius)

Let $A > 0$ with $r = \rho(A)$. Then the following are true.

- $r > 0$.
- $r \in \sigma(A)$.
- *r has algebraic and geometric multiplicities 1.*

Application cont'd

Theorem (Perron-Frobenius)

Let $A > 0$ with $r = \rho(A)$. Then the following are true.

- $r > 0$.
- $r \in \sigma(A)$.
- *r has algebraic and geometric multiplicities 1.*
- *There exists an eigenvector $x > 0$ such that $Ax = rx$.*

Application cont'd

Theorem (Perron-Frobenius)

Let $A > 0$ with $r = \rho(A)$. Then the following are true.

- $r > 0$.
- $r \in \sigma(A)$.
- *r has algebraic and geometric multiplicities 1.*
- *There exists an eigenvector $x > 0$ such that $Ax = rx$.*
- *There exists a unique vector p (the Perron vector) defined by*

Application cont'd

Theorem (Perron-Frobenius)

Let $A > 0$ with $r = \rho(A)$. Then the following are true.

- $r > 0$.
- $r \in \sigma(A)$.
- r *has algebraic and geometric multiplicities 1.*
- *There exists an eigenvector* $x > 0$ *such that* $Ax = rx$.
- *There exists a unique vector* p *(the Perron vector) defined by*
  - $Ap = rp$

Application cont'd

Theorem (Perron-Frobenius)

Let $A > 0$ with $r = \rho(A)$. Then the following are true.

- $r > 0$.
- $r \in \sigma(A)$.
- *r has algebraic and geometric multiplicities 1.*
- *There exists an eigenvector $x > 0$ such that $Ax = rx$.*
- *There exists a unique vector p (the Perron vector) defined by*
  - $Ap = rp$
  - $p > 0$

Application cont'd

Theorem (Perron-Frobenius)

Let $A > 0$ with $r = \rho(A)$. Then the following are true.

- $r > 0$.
- $r \in \sigma(A)$.
- *r has algebraic and geometric multiplicities 1.*
- *There exists an eigenvector $x > 0$ such that $Ax = rx$.*
- *There exists a unique vector p (the Perron vector) defined by*
  - $Ap = rp$
  - $p > 0$
  - $||p||_1 = 1$

Application cont'd

Theorem (Perron-Frobenius)

Let $A > 0$ with $r = \rho(A)$. Then the following are true.

- $r > 0$.
- $r \in \sigma(A)$.
- *r has algebraic and geometric multiplicities 1.*
- *There exists an eigenvector $x > 0$ such that $Ax = rx$.*
- *There exists a unique vector p (the Perron vector) defined by*
  - $Ap = rp$
  - $p > 0$
  - $||p||_1 = 1$
- *Note: When A is stochastic, $r = 1$.*

Application for the Classroom - 6

How does Google implement this process with over 11 billion websites?

Application for the Classroom - 6

How does Google implement this process with over 11 billion websites?

- The power method provides an iterative technique for computing a dominant eigenpair of G.

Application for the Classroom - 6

How does Google implement this process with over 11 billion websites?

- The power method provides an iterative technique for computing a dominant eigenpair of G.
- Typically converges in less than 20 iterations.
- Nice example for Numerical Analysis.

Further Reading

1. *Understanding Search Engines*, by Berry and Browne, 2005

2. *Google's PageRank and Beyond*, by Langville and Meyer, 2006