The PageRank Citation Ranking

Page And Brin

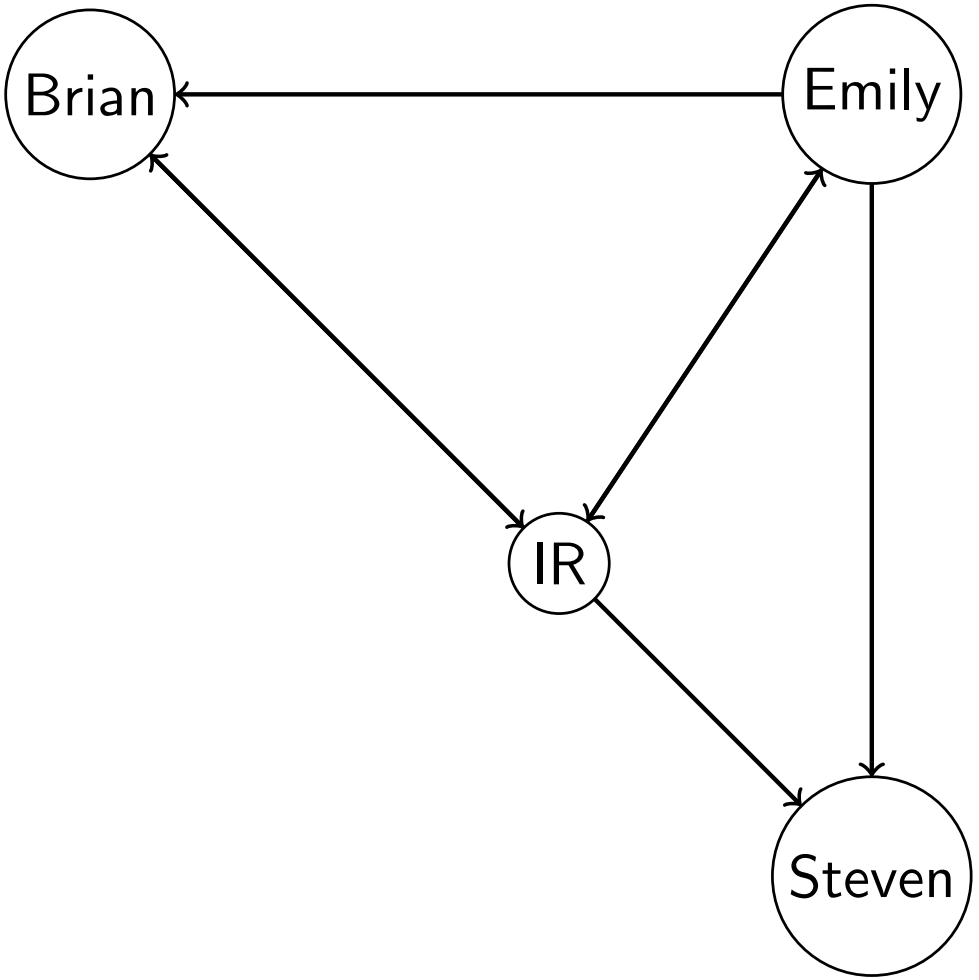
October 17, 2012

Main Idea - Page Rank

Paper Analysis

web page is important if it points to by other important web pages. *Note the recursive definition

IR - course web page, Brian home page, Emily home page, Steven home page



Algebra

- A *eigenvector* is simply a vector V which when multiple by matrix A map back to itself
- A *eigenvalue* is the value of this mapping $v * A = 1/2v$
- for eigenvalue of 1 $v * A = v$

First Idea - Page rank is distributed to out links

- If we know that page importance depends on the importance of pages pointing to it, we can represent it as set of linear equations.
- We can represent the linear equations as vector and matrix multiplication. This is much more useful when the graph is big
- In the matrix - rows are out links, columns are in links.
- The problem transformed into finding the eigenvector which would give us eigenvalue of 1. This eigenvector is the page rank vector.
- Is there a solution? Is there one solution? How long does it take to find the solution?
- What constrains can we impose on A in order to make sure that there is one consist solution which can be found efficiently.
- The condition is that A should be non negative matrix.

Second Idea - Dangling Nodes

- Dangling nodes represent row of zeros.which mean that the only page rank which solve the problem is 0.
- We can create a dangling nodes vector and multiple it by I and add it to the original matrix. d $=$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

- The resulting matrix emulate random walk over the graph. However, one more thing left

Third Idea - Rank Sinks

- Rank sinks are set of pages which link to each other without outer links.
- This can create situation with two solutions for the page rank vector.
- To solve this we introduce randomization.
- Add teleport function to the random walk. I.e. with probably α use existing links. and 1α jump to any other node.
$(1 - \lambda) * T + \lambda * (W + A)$

Forth Idea - Efficiency

- Finding the eignvector on web scale can take a lot of time.
- Instead use an iterative method, which involve mainly matrix by vector multiplication.
- Apparently converges in small number of iterations.

Paper Section

- Introduction and Motivation
- Ranking Web pages
- Implementation
- Convergence Properties
- Application
- Conclusion

Introduction and Motivation

- Web is the largest document collection which is used by non professional users.
- Web can be seen as a huge graph composed of hyper links.We want to take advantage of the link structure.
- Documents are very diverse and have very low quality.Hence traditional citation analysis do not work
- Spam as large percentage of pages.

Ranking web pages

- The paper develop the above ideas.
- Lack more mathematical formalism.
- Does not mention explicitly Markov chain theory

Applying Page Rank

- Page rank is better for under specified query. This is to be affected since content scoring are close.
- Title search
- Suggest usage for authority and trust (like HITS).
- Suggest personalize search engine by making the rank source matrix, point to a single or few web pages based on the user bookmark.
- Estimating web traffic - Notice interesting cases.
- Page rank proxy

Summary

- Seminal paper that changed IR and the web.
- Lack of mathematical formality.
- Led the way to more application in various area (Text summarization,etc).

Page Rank without Hyper Links

Kurland and Lee

October 17, 2012

Context

- This is re-ranking problem. Apply on post retrieval.
- Documents does not contain hyper links.
- We are working on the first X documents retrieved. For example first 50 document
- We want to rerank the documents in the initial results in order to increase relevance. Suggest 6 re ranking methods all based on centrality measures.
- Key issue: How to add centrality measure with no explicit links? I.e. how to create the initial document graph and how we compute centrality from this graph.
- Key Thesis: Adding centrality will help re ranking.

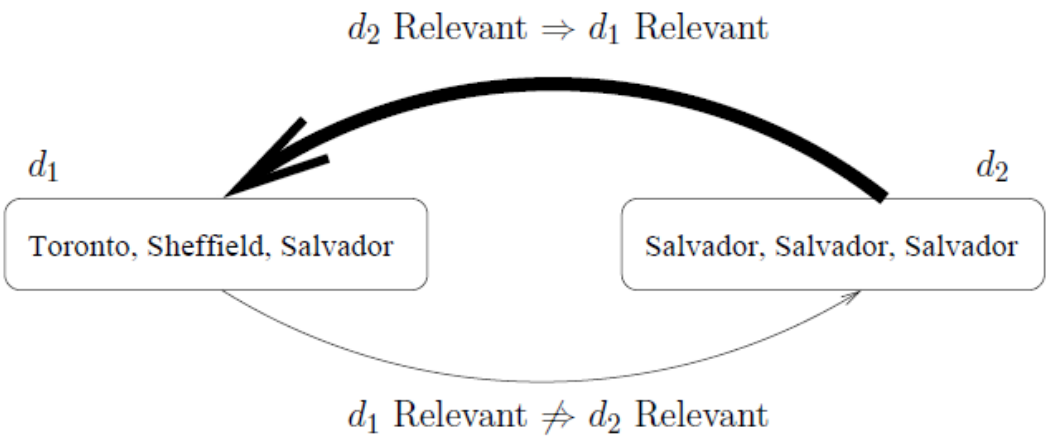
Definitions

- For each document we compute a unigram LM.

$$\widetilde{p}_x^{MLE}(t) \stackrel{def}{=} \frac{tf(t \in x)}{\sum_{t'} tf(t' \in x)}.$$

$$p_x^{MLE}(t_1 t_2 \cdots t_n) \stackrel{def}{=} \prod_{j=1}^{n} \widetilde{p}_x^{MLE}(t_j);$$

- We define a generation value of document o by LM of document g as $P_g(o)$,
- This value denotes how much centrality is transferred from o to g.



Applying PageRank

- Definition: TopGen(d) - set of documents that yield the highest $P_g(d)$
- Definition: Offspring of document d are documents for which d is part of TopGen(o)
- Centrality is transferred from the offspring to the generator as follow: We define weights between between documents as follow

$$wt_U(o \to g) = \begin{cases} 1 & \text{if } g \in TopGen(o), \\ 0 & \text{otherwise;} \end{cases}$$

$$wtw(o \to g) = \begin{cases} p_g(o) & \text{if } g \in TopGen(o), \\ 0 & \text{otherwise.} \end{cases}$$

- To make it solvable (non negative matrix),we add the smoothed version of page rank (random walk)

$$wt^{[\lambda]}(o \to g) = \lambda \cdot \frac{1}{|\mathcal{D}_{\text{init}}|} + (1-\lambda) \cdot \frac{wt(o \to g)}{\sum_{g' \in \mathcal{D}_{\text{init}}} wt(o \to g')}.$$

Centrality Measures

$$Cen_I(d;G) \stackrel{def}{=} \sum_{o \in \mathcal{D}_{\text{init}}} wt(o \rightarrow d).$$

- Unigram or Weighted Influx

$$Cen_{RI}(d;G) \stackrel{def}{=} \sum_{o \in \mathcal{D}_{\text{init}}} wt(o \rightarrow d) \cdot Cen_{RI}(o;G),$$

- Recursive Unigram or Weighted Influx

$$Cen(d;G) \cdot p_d(q),$$

- Unigram or Weighted Influx + LM
- Recursive Unigram or Weighted Influx + LM

Results

	AP89			AP			WSJ			TREC8		
	prec@5	prec@10	MRR	prec@5	prec@10	MRR	prec@5	prec@10	MRR	prec@5	prec@10	MRR
upper bound	63.7	53.1	75.5	87.6	78.8	93.0	89.6	80.0	100.0	94.4	85.0	98.0
init. ranking	28.3	26.5	52.3	45.7	43.2	59.6	54.8	48.4	76.2	50.0	45.6	69.1
opt. baselines	30.0	27.4	**54.3**	46.5	43.9	63.5	56.0	49.4	77.2	51.2	46.4	**69.6**
U-In	29.6	*27.8*	39.5 $_o$	*50.9*	*49.0* $^{i}_{o}$	**66.3**	50.0	46.6	66.7	50.0	45.0	62.0
W-In	*31.3*	*29.6*	46.8	*51.3*	*48.7* ^{i}	*64.4*	52.0	47.8	63.3 $_o$	49.2	43.4	63.7
U-In+LM	**33.5**	27.0	46.5	*51.3* ^{i}	**49.4** $^{i}_{o}$	63.2	*56.4*	49.2	73.6	*52.8*	**52.0** $^{i}_{o}$	66.6
W-In+LM	*31.7*	*27.6*	48.4	*51.1* ^{i}	*48.4* $^{i}_{o}$	63.0	*57.2*	*50.0*	77.2	*51.6*	*49.6* ^{i}	64.5
R-U-In	*31.3*	*28.9*	46.4	*51.5*	*48.9* ^{i}	63.4	53.6	*49.6*	68.5	*52.0*	44.6	66.5
R-W-In	*32.2*	*29.6*	40.5 $_o$	*52.1* ^{i}	*49.1* $^{i}_{o}$	*63.9*	54.0	49.2	70.2	*52.4*	44.6	66.5
R-U-In+LM	*33.0*	*29.3*	45.8	*52.1* $^{i}_{o}$	*49.2* $^{i}_{o}$	*64.3*	**58.8** ^{i}	**51.0** ^{i}	**78.6**	*55.6*	46.0	68.4
R-W-In+LM	**33.5**	**29.8**	46.0	**52.9** $^{i}_{o}$	*49.0* $^{i}_{o}$	62.6	**58.8** ^{i}	*50.6*	**78.6**	**56.0**	45.8	67.6

Comparing LM to Vector space score

		U-In	W-In	U-In+LM	W-In+LM	R-U-In	R-W-In	R-U-In+LM	R-W-In+LM
AP89	prec @5	□	□				□		
	prec @10		♦						♦
	MRR	□	□	□		□	□	□	
AP	prec @5	♦	♦	♦	♦	♦	♦	♦	♦
	prec @10	♦	♦	♦	♦	♦	♦	♦	♦
	MRR	♦	♦		♦		♦	♦	
WSJ	prec @5						♦		
	prec @10					♦	♦		
	MRR		□						
TREC8	prec @5	♦	♦			♦	♦	♦	♦
	prec @10	♦	♦		♦		♦		
	MRR	□			□				

Re rank by using non structural heuristics

	AP89			AP			WSJ			TREC8		
	prec@5	prec@10	MRR	prec@5	prec@10	MRR	prec@5	prec@10	MRR	prec@5	prec@10	MRR
uniform (= init)	28.3	26.5	52.3	45.7	43.2	59.6	54.8	48.4	76.2	50.0	45.6	69.1
W-ln	*31.7*	*27.6*	48.4	*51.1∗*	*48.4∗*	**63.0**	*57.2*	*50.0*	*77.2*	*51.6*	**49.6∗**	64.5
R-W-ln	**33.5**	**29.8**	46.0	**52.9∗**	**49.0∗**	*62.6*	**58.8∗**	**50.6**	**78.6**	**56.0**	*45.8*	67.6
length	*29.1*	24.3	50.8	41.6	41.4	55.3	44.4∗	42.4∗	64.6∗	47.2	41.4	64.2
log(length)	*30.4*	*27.0*	*52.5*	45.3	43.2	*60.6*	*57.2*	*49.0*	69.8∗	49.6	*46.8*	*69.2*
entropy	*30.0*	26.5	**52.6**	*46.1*	42.5	*60.8*	*56.8*	*48.6*	71.1∗	49.6	*46.8*	**71.7∗**
uniqTerms	27.4	24.8	52.3	42.0	41.3	56.2	50.0	44.6	68.8	49.2	44.2	*71.2*
log(uniqTerms)	*30.4*	*27.0*	*52.5*	*45.9*	42.3	*60.8*	*57.2*	*49.0*	70.0∗	49.6	*47.2*	*70.0*

Summary and comments

- The Thesis proved to be correct. It is possible to get better precession by exploiting structural centrality measures
- Trying to apply it to full retrieval did not work as well.
- The proposed methods was better than HITS.
- Overall seems like a well researched paper supported by empirical results.