CORBETTA Andrea
DE SIMONI Pietro
SESSI Michela
VERRIER Alexandra

Stockholm University

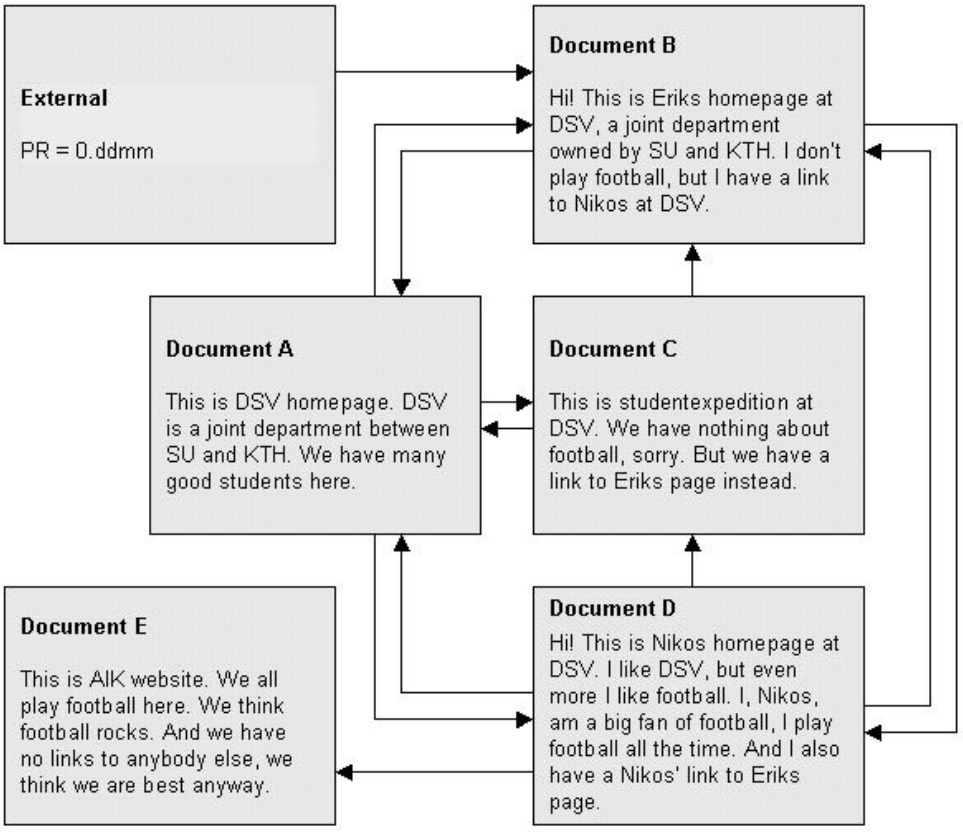
Report of Assignment 2 :

Impact of Page Rank

Table of Contents

Introduction

The aim of this assignment is to rank the following documents starting from a selected query using Text Similarity and PageRank algorithm.



We have chosen the following query:

***q:* Nikos DSV**

According to the birthday of the oldest member in our group, we have the following Page Rank value of the external document:

$PR_{external}$ = 0.25050

Text Similarity

We want to calculate the text similarity value between each document *d* and the query *q*. The text similarity formula is the following:

$$sim(q,d) = \frac{\sum_{i=1}^{n} q_i * d_i}{N_d}$$

According to our query we can simplify the formula as the following one:

$sim(q,d) = \frac{q_{Nikos} * d_{Nikos} + q_{DSV} * d_{DSV}}{N_d}$, with N_d the number of all words in the document d including the stop-word.

Using the previous formula we obtain the following values:

Document *d*	*d(DSV)*	*d(Nikos)*	N*d*	*sim(q,d)*
A	2	0	19	0,10526
B	2	1	28	0,10714
C	1	0	20	0,05000
D	2	3	40	0,12500
E	0	0	27	0,00000

Page Rank

The Page Rank value of each document has been set at zero. According to the Page Rank algorithm and the links we have computed the values with the following formulas:

$$PR(A) = (1-0,85) + 0,85 \cdot (PR(B)/2 + PR(C)/2 + PR(D)/4)$$
$$PR(B) = (1-0,85) + 0,85 \cdot (PR(A)/3 + PR(C)/2 + PR(D)/4 + PR(EXT)/1)$$
$$PR(C) = (1-0,85) + 0,85 \cdot (PR(A)/3 + PR(D)/4)$$
$$PR(D) = (1-0,85) + 0,85 \cdot (PR(A)/3 + PR(B)/2)$$
$$PR(E) = (1-0,85) + 0,85 \cdot (PR(D)/4)$$

We iterated those formulas until convergence to determinate the final Page Rank values.
The results (reported in the final table) were obtained after 76 iterations.
Here the drive sheet for the iterations: Page Rank Calculation (Part 1).

Combining Text Similarity with Page Rank

We re-rank the documents using a different formula that uses both Text similarity and Page Rank values that we have calculated, as the following step:

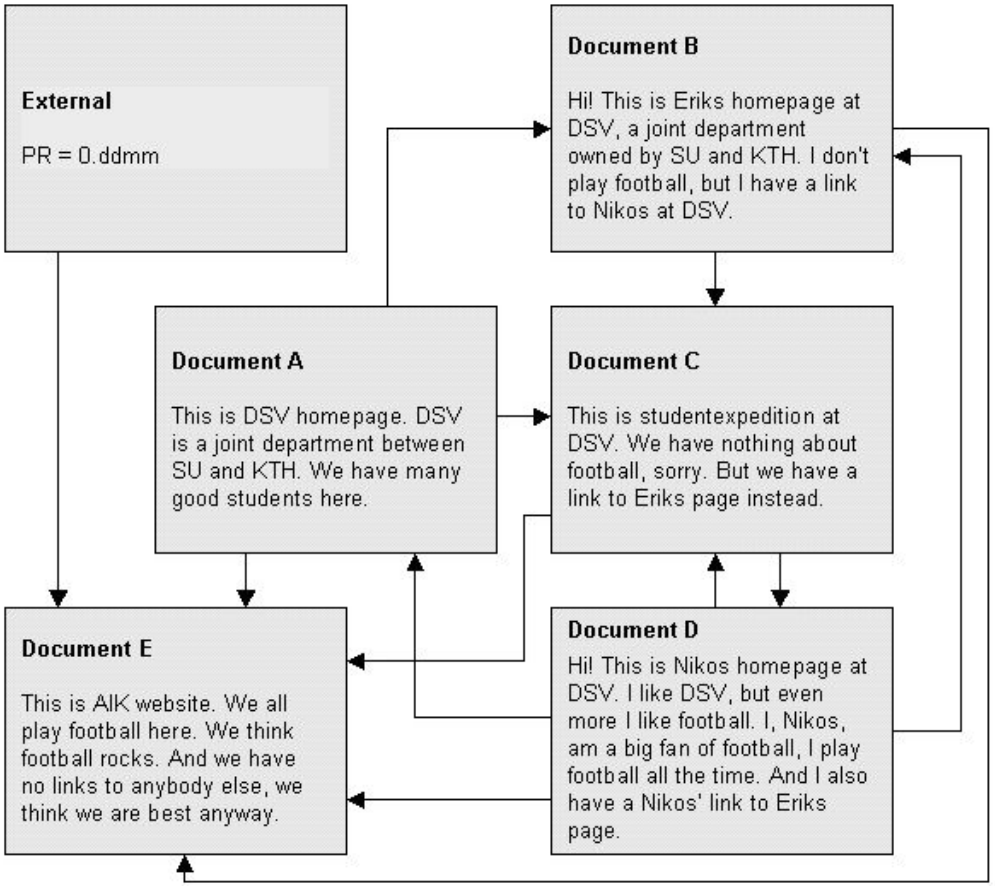
$$SIM_1(q,d) = sim(q,d) + 0,5 \cdot PR_{initial}(d)$$

Re-linking the Documents

We have re-linked the documents in order to move up the last two documents (C and E) in the document list. We start moving links following this rules:

- The total number of the links does not change.
- The source of the link does not change - every document has the same outgoing links.
- Only the destination of the links may change - the last two documents get more links.
- Still, each document (except External) must have at least one incoming link.

Finally we have the following new linking between our 6 documents:



According to the Page Rank algorithm and to the new outgoing links we have computed the values with the following formulas:

$$PR(A) = (1-0,85) + 0,85 \cdot (PR(D)/4)$$
$$PR(B) = (1-0,85) + 0,85 \cdot (PR(A)/3 + PR(D)/4)$$
$$PR(C) = (1-0,85) + 0,85 \cdot (PR(A)/3 + PR(B)/2 + PR(D)/4)$$
$$PR(D) = (1-0,85) + 0,85 \cdot (PR(C)/2)$$
$$PR(E) = (1-0,85) + 0,85 \cdot (PR(A)/3 + PR(B)/2 + PR(C)/2 + PR(D)/4 + PRexternal/1)$$

Using the same process as in the first part, the results (reported in the final table) were obtained after 22 iterations.
Here the drive sheet for the iterations: Page Rank Calculation (Part 2)

To re-rank the documents we use the following formula:

$$SIM_2(q,d) = sim(q,d) + 0,5 \cdot PR_{new}(d)$$

The following table provides a resume of our calculation and values using the previous formulas:

Rank	*Initial Page Rank*		*New Page Rank*		*sim(q, d)*		*SIM1(q, d)*		*SIM2(q, d)*	
	Doc. id	**PR value**	**Doc. id**	**PR value**	**Doc. id**	**Sim. value**	**Doc. id**	**Sim. value**	**Doc. id**	**Sim. value**
1	B	1,19272	E	0,78065	D	0,12500	B	0,70350	E	0,39033
2	A	1,15846	C	0,39840	B	0,10714	A	0,68450	D	0,28466
3	D	0,98514	D	0,31932	A	0,10526	D	0,61757	C	0,24920
4	C	0,68757	B	0,27958	C	0,05000	C	0,39379	B	0,24693
5	E	0,35934	A	0,21785	E	0,00000	E	0,17967	A	0,21419

Conclusion

From the first passage we checked the weights of the words in unique documents with the similarity; then we ranked these documents using incoming links with Page Rank algorithm. This two measures gave us different ranking, that could be well combined by the SIM1.

The original link structure returns high Page Rank for documents B and A, instead documents C and E have a low rank. Through the re-linking process we managed to bring documents E and C to top of the Page Rank scale.

The exercise shows how the link structure influences the documents ranking, in fact Page Rank algorithm is based on the incoming links as explained in the formulas.
We can conclude that if you want to improve the visibility of your page on a search engine results you should work on both aspects:

- relevant keywords in the text according to specific queries.
- valuable incoming links from high ranked pages.