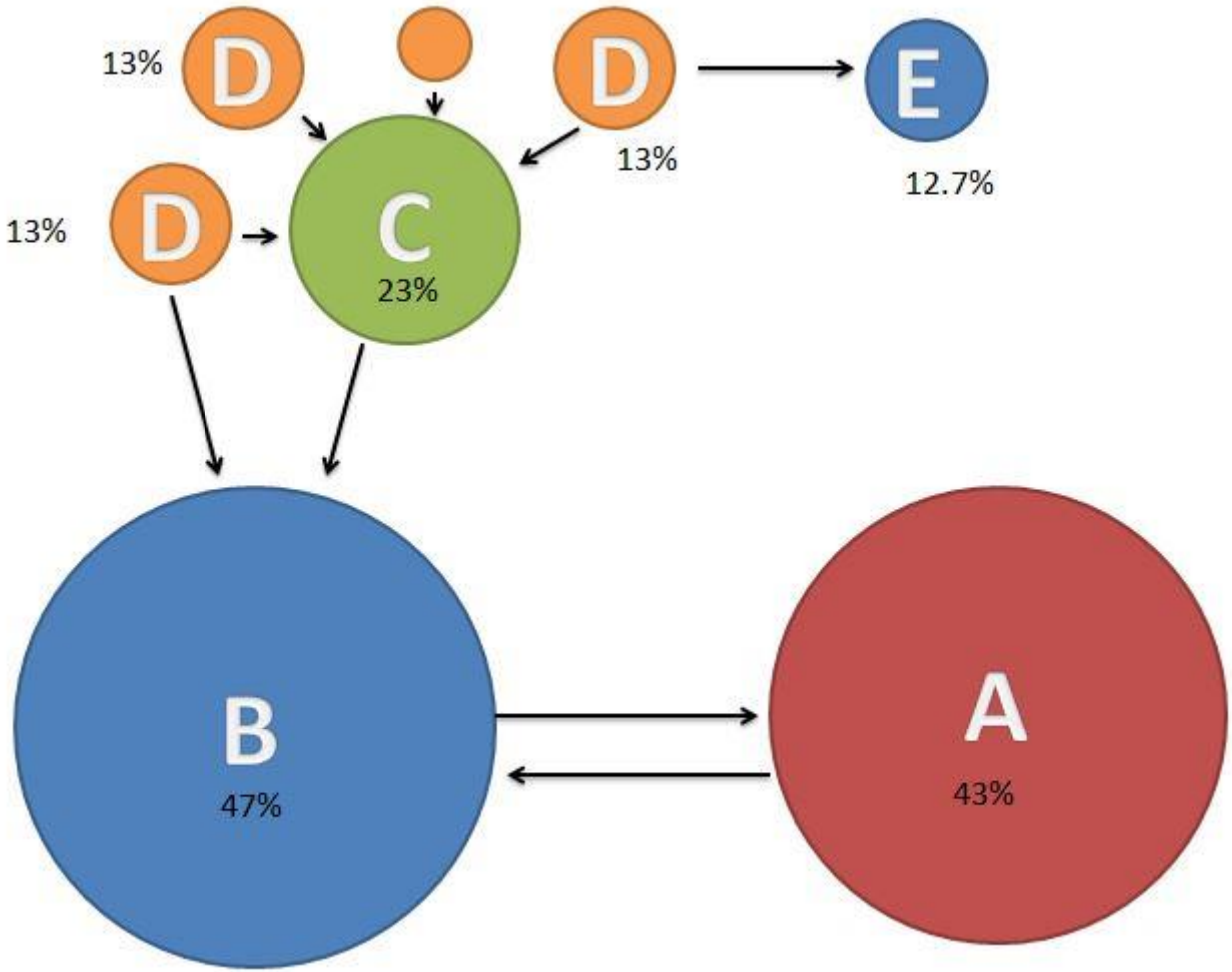
■ Page Rank



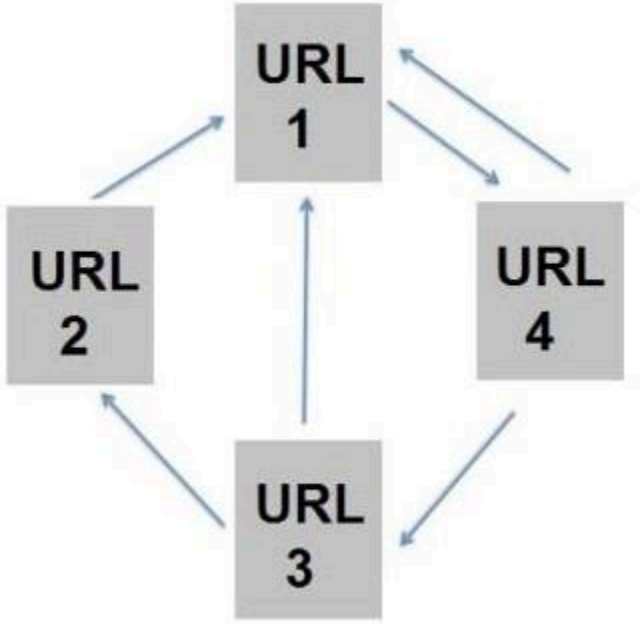
$$PageRank\ of\ site = \sum \frac{PageRank\ of\ inbound\ link}{Number\ of\ links\ on\ that\ page}$$

OR

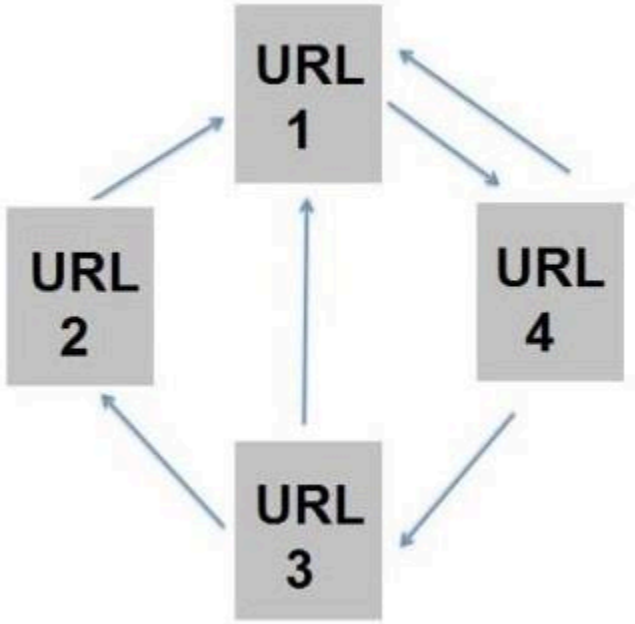
$$PR(u) = (1-d) + d \times \sum \frac{PR(v)}{N(v)}$$

d: damping factor; Random Surfer Model (experimentally 0.85)

url_1	url_4
url_2	url_1
url_3	url_2
url_3	url_1
url_4	url_3
url_4	url_1



url_1	url_4
url_2	url_1
url_3	url_2
url_3	url_1
url_4	url_3
url_4	url_1



$$A^{12}v = \begin{pmatrix} 0.366 \\ 0.089 \\ 0.184 \\ 0.360 \end{pmatrix}$$

iterative

$$A = \begin{pmatrix} 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad v = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.125 \\ 0.125 \\ 0.25 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.5 \\ 0.125 \\ 0.125 \\ 0.25 \end{pmatrix} = \begin{pmatrix} 0.312 \\ 0.062 \\ 0.125 \\ 0.5 \end{pmatrix}$$

■ Collecting Links

```
def getUrls(url, params=None, num_retries=2):
    headers = {"User-agent":"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_5) AppleWebKit/537.36

    resp = requests.get(url, params=params, headers=headers)

    if 500 <= resp.status_code < 600 and num_retries>0:
        print("error:", resp.status_code, resp.reason)
        return getUrls(url=url, params=params, num_retries=(num_retries-1))

    html = BeautifulSoup(resp.content, "lxml")
    links = html.select("h3.r > a")

    return [link["href"] for link in links if link.has_attr("href") == True]



seed = "https://www.google.co.kr/search"
params = {
    "q":"한글",
    "ie":"utf-8"
}

urlList = getUrls(seed, params)
print(urlList)
```

■ Following Links

```
while queue:
    url = queue.pop()
    urlList = getUrls(url)
    print(url, len(urlList))
    queue.extend(urlList)
```

○ 예제

```
def getUrls(url, params=None, select="a", num_retries=2):
    headers = {"User-agent":"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_5) AppleWebKit/537.36 (KHTML, like Gecko)

    resp = requests.get(url, params=params, headers=headers)

    if 500 <= resp.status_code < 600 and num_retries>0:
        print("error:", resp.status_code, resp.reason)
        return getUrls(url=url, params=params, num_retries=(num_retries-1))

    html = BeautifulSoup(resp.content, "lxml")
    links = html.select(select)

    return [link.get("href") for link in links if link.has_attr("href") == True]

def checkUrl(url):
    dest = parse.urlparse(url)

    if len(dest.scheme) > 0 and dest.scheme in ["http", "https"]:
        return True
    else:
        return False

seed = "https://www.google.com/search"
params = {
    "q":"한글",
    "ie":"utf-8"
}

queue = getUrls(seed, params, "h3.r > a")
result = list()

while queue:
    url = queue.pop()

    while checkUrl(url) is False:
        url = queue.pop()

    urlList = getUrls(url)

    result.append(url)
    result.extend(urlList)

    print(url, len(urlList), end="\n\n")
```

■ Circular Links

도메인,
방문이력