The Second International Conference "Problems of Cybernetics and Informatics" September 10-12, 2008, Baku, Azerbaijan. Section #1 "Information and Communication Technologies" www.pci2008.science.az/1/38.pdf

RANKING ALGORITHM AND ITS APPLICATION FOR THE DOCUMENT SUMMARIZATION

Ramiz Aliguliyev

Institute of Information Technology of ANAS, Baku, Azerbaijan
a.ramiz@science.az

1. Introduction

Of the ranking algorithms based on the Web structure analysis the PageRank algorithm is the most popular one [1]. Despite the fact that the ranging algorithms achieved rather high level of accuracy in their early years later on their efficiency reduced; weak points in relation to unfair methods of rating manipulation (spam dexing) were revealed. The reason was that in those algorithms the page significance degree was determined by the number of entering hyperlinks and the pages subject affinity was not taken into account. To increase the efficiency in papers [2-5] some modifications of the algorithm PageRank were suggested that took into account the subject affinity of the linked pages. In those modifications a weight was given to every hyperlink that determines the affinity degree of the pages.

In the present paper a modification of the Page Rank algorithm is suggested that implies taking into account contribution of every Web-graph page when calculating the page rank. The suggested modification was applied to writing abstracts of documents and was used for ranking sentences.

2. Modification of the Page Rank algorithm

The Page Rank Algorithm is modeled by the following iteration process:

$$PR(v_j) = \frac{(1-d)}{n} + d \sum_{v_i \in B(v_j)} \frac{PR(v_i)}{|F(v_i)|}, \quad (1)$$

where $PR(v_i)$ is the Page Rank of page v_i, $B(v_i)$ is the set of pages which refer to page v_i, $F(v_i)$ is the set of pages to which page v_i refers, $d \in [0.8,1]$ is the damping coefficient. In algorithm (1) every outgoing reference in the page v_i is chosen with the same probability which is equal to $\frac{1}{|F(v_i)|}$, where $|F(v_i)|$ is the number of the outgoing references of the page v_i.

It is easy to see that in formula (1) the portion of the rank $PR(v_i)$ page v_i, that spreads to page v_j, directly depends on the chosen path that leads from vertex v_i to v_j. It is known that one can come from page v_i to page v_j by different paths and the rank contribution of page v_i in page v_j will differ depending on the path. To be able to record the rank contribution of vertex v_i in vertex v_j by all paths we suggest the following modification of the PageRank algorithm:

$$PR_{resist}(v_j) = \frac{(1-d)}{n} + d \sum_{\substack{v_i \in V \\ v_i \neq v_j}} \frac{PR(v_i)}{r_{ij}}, \quad (2)$$

where r_{ij} is the resistance distance between vertices v_i and v_j. The resistance distance between the graph vertices is determined similarly to how the resistance of an electric chain is calculated by the Kirchhoff's law [6,7].

To calculate the resistance distance first the Laplace matrix is formed. The Laplace matrix $\mathbf{L}$ of weighted nonoriented graph is determined like this [6]:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{3}$$

where $\mathbf{A}$ is symmetric matrix which elements are determined like this:

$$a_{ij} = \begin{cases} w_{ij}, \text{if vertices } v_i \text{ and } v_j \text{ are adjacent ones} \\ 0, \text{ otherwise} \end{cases}, \tag{4}$$

where $w_{ij} \geq 0$ is the weight of the edge that connects vertices v_i and v_i. The matrix $\mathbf{D}$ is the diagonal one which elements are determined by the formula $d_{ii} = [\mathbf{D}]_{ii} = a_{i\bullet} = \sum_{j=1}^{n} a_{ij}$. For the oriented graph the Laplace matrix is determined by the formula: $\mathbf{L} = \mathbf{A}\mathbf{A}^T$ where $\mathbf{A}^T$ is the transposition of the matrix $\mathbf{A}$.

The adjacency matrix $\mathbf{A}$ of the oriented graph $G = (V, E)$ with $n \times m$ elements (n is the number of vertices, m is the number of edges) is determined like this:

$$(\mathbf{A})_{ie} = \begin{cases} -1, \text{ if the vertex } v_i \text{ is the end of the edge } e \\ 1, \text{ if the vertex } v_i \text{ is the start of the edge } e \\ 0, \text{ if the vertex } v_i \text{ and the edge } e \text{ are not adjacent} \end{cases}. \tag{5}$$

It follows from the (3)-(5) that for any graph $\sum_{i=1}^{n} (\mathbf{L})_{ij} = \sum_{j=1}^{n} (\mathbf{L})_{ij} = 0$, i.e. the matrix $\mathbf{L}$ is singular one. In [6] determination of the resistance distance is reduced to the calculation of the Moore-Penrose pseudo reversion $\mathbf{L}^+$ of the Laplace singular matrix $\mathbf{L}$:

$$r_{ij} = (\mathbf{L}^+)_{ii} + (\mathbf{L}^+)_{jj} - (\mathbf{L}^+)_{ij} - (\mathbf{L}^+)_{ji}. \tag{6}$$

In the formula (6) $(\mathbf{L}^+)_{ij}$ designates ij element of the pseudo reversion $\mathbf{L}^+$. The calculation of the Moore-Penrose pseudo reversion $\mathbf{L}^+$ [6,7] so in the paper [7] the following formula is suggested for calculation of the distance r_{ij}:

$$r_{ij} = \frac{\det \mathbf{L}(i, j)}{\det \mathbf{L}(i)},$$

where the submatrix $\mathbf{L}(i)$ is obtained from the matrix $\mathbf{L}$ by removing i row and i column and the submatrix $\mathbf{L}(i, j)$ is obtained by removing i row и j column, $i \neq j$.

The suggested algorithm is currently applied to writing abstracts of documents. Particularly it is applied for the sentence ranking. The rank of the sentence determines degree of its informativeness.

3. Application of the ranking algorithm to writing abstracts of documents

Automatic abstract writing is the process of forming an abridged version of the original document that reflects its basic content. There are to classes of approaches to writing abstracts of the text documents [8]. The first approach called abstracting generates abstract by generating new sentences. The second approach is called extracting. It called abstract by extracting informative sentences from the original text. The most papers deal with extracting informative sentences from original document. It is directly related with difficulties of forming grammatical correct sentences.

Various methods were suggested for extracting informative fragments from documents. The methods suggested in the papers [9-11] based on determining the relevance of the sentences. These methods are efficient if the document deals with one subject matter. It the document is related with several subject matters these methods do not give desirable results. To solve this

problem the methods were suggested in the papers [12-15] that determine subject fields and informative sentences in the documents. The suggested methods include two stages. In the first stage subject fields are determined in the document by clustering sentences. After clustering informative sentences are found in every cluster (subject field) that reflect its (the cluster's) content. In this paper the document is summarized by ranking sentences. For ranking the sentences the algorithm is used that was described in the previous column

Let the document $D = \{S_1,...,S_n\}$ be given where n is the number of sentences in the document. To apply the ranking algorithm described above the document is presented in the form of weighted graph. Let $T = \{t_1,...,t_m\}$ is the set of words in the documents. By using the Vector Space Model let us present the sentence S_i as a vector of a m - numerical space $S_i = \{w_{i1},...,w_{im}\}$ where w_{ik} is the weight of the word t_k in the sentence S_i; the weight is determined by the pattern TF-IDF, $w_{ik} = f_{ik} \log\left(\frac{n}{n_k}\right)$. Then the proximity measure between the couples S_i and S_j is calculated which is determined by the metrics of the cosine:

$$w_{ij} = sim(S_i, S_j) = \cos(S_i, S_j) = \frac{\sum_{k=1}^{m} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{m} w_{ik}^2 \cdot \sum_{k=1}^{m} w_{jk}^2}}, \quad i, j = 1,...,n.$$

If the proximity measure $w_{ij} > 0$ then the vertices S_i and S_j are connected by edge. Thus the document is presented as a weighted nonoriented graph. The edge weight is equal to the proximity measure of the vertices (sentences).

After this presentation the algorithm described above is applied to the sentence ranking. Then the second stage is executed – including the sentences with ranks higher than the specified threshold in the abstract. The threshold is the controlled parameter that depends on the compression ratio. The compression ratio is determined as the ratio of the sizes of the abstract and documents, i.e., it determines the compression degree of the document.

4.Conclusion

In the PageRank algorithm the page rank depends on the ranks of referring pages. It is known that one can get from initial page to destination page by different paths and page contribution will differ depending on the particular path. A modification of the PageRank algorithm is suggested that enables one to record the effect of path on the page rank. In the suggested modification the page contribution that spreads to other pages depends on the resistance distance. The resistance distance determines the relation degree of vertices which is very important for revealing informative sentences.

References

[1] Brin S., Page L. The anatomy of a large-scale hyper-textual web search engine //Computer Networks and ISDN systems. 1998. Vol. 30. №№ 1–7. P. 107–117.

[2] Diligenti M., Gori M., Maggini M. A unified probabilistic framework for web page scoring systems //IEEE Transactions on Knowledge and Data Engineering. 2004. Vol.16. № 1. P. 4-16.

[3] Ingongngam P., Rungsawang A. Topic-centric algorithm: a novel approach to Web link analysis //Proceedings of the 18th International Conference on Advanced Information Networking and Applications (AINA'04). Fukuoka. Japan. 2004. Vo1.2. P. 299-301.

[4] Aliguliyev R.M. Optimization model for ranging Web pages. //Management systems and information technologies. Moscow, 2006, N 3(25), p. 4-7 (In Russian)

[5] Alguliyev R.M., Aliguliyev R.M. Ranking Web pages by using of the reciprocal information between the hyperlinks. //Problems of management. Moscow, 2007, N 4, p. 24-29. (In Russian)
[6] Klein D.J. Resistance-distance sum rules //Croatica Chemica Acta. 2002. Vol. 75. № 2. P. 633-649.
[7] Bapat R.B., Gutman I., Xiao W. A simple method for computing resistance distance //Z. Naturforschung – A Journal of Physical Sciences. 2003. Vol. 58a. №№ 9-10. P. 494-498.
[8] Jones K. S. Automatic summarizing: the state of the art //Information Processing and Management. 2007. Vol. 43. № 6. P. 1449-1481.
[9] Alguliev R.M., Aliguliyev R.M. Effective summarization method of text documents //Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). 19-22 September 2005. France. P. 264-271.
[10] Alguliyev R.M., Aliguliyev R.M. A new method for summarizing text documents and estimation of the classification results from three standpoints. //Telekommunikatsii, Moscow, 2006, N 3, p. 7-16. (In Russian)
[11] Aliguliyev R.M. Using the F-measure as similarity measure for automatic text summarization //Vichislitelnie Tehnologii, Novosibirsk, 2008, v. 13, N 3, p. 5-14.
[12] Aliguliyev R.M. Automatic document summarization by sentence extraction // Vichislitelnie Tehnologii, Novosibirsk, 2007. Том 12. № 5. С. 5-15.
[13] Alguliev R.M., Aliguliyev R.M. Summarization of text-based documents with a determination of latent topical sections and information-rich sentences //Automatic Control and Computer Sciences. 2007. Vol. 41. № 3. P. 132-140.
[14] Aliguliyev R.M. A Novel Partitioning-Based Clustering Method and Generic Document Summarization //Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops) (WI-IATW'06). Hong Kong. 18-22 December 2006. P. 626-629.
[15] Alguliev R.M., Aliguliyev R.M., Bagirov A.M. Global optimization in the summarization of text documents //Automatic Control and Computer Sciences. 2005. Vol. 39. № 6. P. 42-47.