Link analysis algorithms (HITS and PageRank) in the information retrieval context

Dario Smolčić, Deni Munjas

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
dario.smolcic@fer.hr, deni.munjas@fer.hr

Abstract

We implemented an information retrieval system that re-ranks the documents based on the analysis of between-document links with the link analysis algorithms PageRank and HITS. The information retrieval system was built using a binary vector-space retrieval model. Scores produced by vector space model were refined with the HITS and PageRank algorithms. The performance of the IR system both with and without PageRank and HITS was evaluated on the European Media Monitor test collection using standard IR metrics (R-Precision, Mean Average Precision, mean reciprocal rank).

1. Introduction

Information retrieval (IR) is the activity of obtaining information resources relevant to an user's information need from a collection of information resources. The focus of information retrieval in this paper are unstructured text documents. The classical problem in IR is the ad-hoc retrieval problem. In ad-hoc retrieval, the user enters a query describing the desired information. The system then returns a list of documents.

The biggest information retrieval systems are search engines like Google and Yahoo. Search engines are nowadays becoming necessity tools for information retrieval, learning and many other aspects of life. While different information retrieval systems may have different search models and principles, vector space model, PageRank and HITS are commonly used by most popular ones, and they can be combined together to improve the search performance.

2. Vector space model

In vector space model documents and queries are represented as vectors of index terms:

$$\mathbf{d_j} = [w_{1j}, w_{2j}, ..., w_{tj}]$$

$$\mathbf{q} = [w_{1q}, w_{2q}, ..., w_{tq}]$$

Where t is the number of index terms in the vocabulary. Since we are using binary vector space model, weights in vectors are binary numbers. Weights are computed according to the following formula:

$$w_{ij} = \begin{cases} 1, & \text{if document } \mathbf{d_j} \text{ contains term } i \\ 0, & \text{otherwise} \end{cases}$$

The relevance of the document for the query is estimated by computing some distance or similarity metric between the two vectors.

Some possible similarity metrics are cosine similarity and dice similarity. Some possible distance metrics are euclidean distance and Manhattan distance. In this project, we use cosine similarity and dice similarity.

$$Cosine(\mathbf{d_j}, \mathbf{q}) = \frac{\mathbf{d_j q}}{\|\mathbf{d_j}\| \|\mathbf{q}\|}$$

$$Dice(\mathbf{d_j}, \mathbf{q}) = \frac{2\|\mathbf{d_j q}\|}{\|\mathbf{d_j}\|^2 + \|\mathbf{q}\|^2}$$

The most relevant documents for the given query are the ones with the highest similarity score.

3. Link analysis algorithms

Information retrieval content scores are not efficient for web usage due to web's massive size, therefore popularity/importance scores are introduced. Popularity scores harness information from the immense graph defined by web's hyperlink structure. In our case we are not operating online on a website database, but we are performing information retrieval on a great number of documents and we did create links among them.

Hyperlinks among documents are interpreted as recommendations. Documents with more recommendations are more important than documents with less recommendations. When assessing importance of a document, it is necessary to take into account the importance of the recommender. Recommendations from more important recommenders are worth more while the overall number of recommendations issued by the recommender also matter.

3.1. PageRank

Let **H** be the row normalized Web graph adjacency matrix.

$$\mathbf{H_{ij}} = \begin{cases} 1/|P_i| & \text{if there is a link from page } P_i \text{ to page } P_j \\ 0, & \text{otherwise} \end{cases}$$

Assume there is a random surfer who surfs the web by following the hyperlink structure of the Web represented by **H**. When on a page containing no hyperlinks, surfer may hyperlink to any other page therefore a stochasticity adjustment needs to be applied to **H**:

$$\mathbf{S} = \mathbf{H} + \mathbf{a}(\frac{1}{n}\mathbf{e}^T)$$

Where $a_i = 1$ if page i has no hyperlinks and **e** is a vector of ones.

Sometimes, a random surfer abandons the hyperlink method of surfing, randomly choosing the next page. With that in regard a primitivity adjustment needs to be applied according to the following formula:

$$\mathbf{G} = \alpha \mathbf{s} + (1-\alpha)\mathbf{E}$$

Where α is a scalar between 0 and 1 determining the frequency of abandoning the hyperlink structure and $\mathbf{E}$ is the normalized matrix of ones.

Let $\boldsymbol{\pi}$ be the probability distribution of the surfer across the web pages (the vector of PageRank scores) then the following formula must be valid:

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \mathbf{G}$$

Thus if we were to compute the principal left eigenvector of the matrix $\mathbf{G}$ — the one with eigenvalue 1 — we would have computed the PageRank values.

3.2. HITS

HITS defines hubs and authorities. A page is considered a hub if it contains many hyperlinks. On the other hand it is considered an authority if many hyperlinks point to it. A page is a good hub if it points to good authorities, and a good authority if it is pointed to by good hubs. We need to assign two scores to every document: an authority score and a hub score. We can update these scores iteratively according to the following formulae:

$$\begin{aligned} \mathbf{h} &\leftarrow \mathbf{A}\mathbf{a} \\ \mathbf{a} &\leftarrow \mathbf{A}^T\mathbf{h} \end{aligned} \tag{1}$$

Where $\mathbf{h}$ and $\mathbf{a}$ are the vectors of hub and authority scores and $\mathbf{A}$ is adjacency matrix. A_ij is 1 if there is a hyperlink from page i to page j, and 0 otherwise. Now the right hand side of each line of equation 1 is a vector that is the left hand side of the other line of equation. Substituting these into one another, we may rewrite equation 1 as:

$$\begin{aligned} \mathbf{h} &\leftarrow \mathbf{A}\mathbf{A}^T\mathbf{h} \\ \mathbf{a} &\leftarrow \mathbf{A}^T\mathbf{A}\mathbf{a} \end{aligned} \tag{2}$$

Now we can compute vector of hub scores $\mathbf{h}$ as principal eigenvector of $\mathbf{A}\mathbf{A}^T$ and vector of authority scores $\mathbf{a}$ as principal eigenvector of $\mathbf{A}^T\mathbf{A}$.

In this paper the links were created in that way that adjacency matrix $\mathbf{A}$ is always symetric. That means that hub and authority scores turn out to be identical, so in future references we will denote this score as single HITS score.

4. Pipeline

In this section we present the pipeline of the entire information retrieval system.

1. Each document is preprocessed with a Porter stemmer and stopword removal.

2. Vocabulary is constructed from the entire collection.

3. Every document is converted to a binary vector.

4. For each document and a query, vector space model score (similarity score) $S_vsm(d_j, q)$ is calculated.

5. PageRank and HITS scores are calculated ($S_{la}(d_j)$).

6. In the last step vector space model scores are refined with PageRank or HITS scores according to the following formula:

$$S(d_j, q) = wS_vsm(d_j, q) + \frac{(1-w)S_{la}(d_j)}{log(r(d_j, q)) + log5} \tag{3}$$

   Where $r(d_j, q)$ is the rank of documnt d_j according to the vector space model and w is the number between 0 and 1 that determines the importance of vector space model score in relation to the link analysis score. This formula was taken from (**?**).

Table 1: Mean average precision.

VSM	VSM + PageRank	VSM + HITS
0.264	0.266	0

5. Test collection

We worked with European Media Monitor test collection which contains 1742 documents and a set of 48 queries with relevance judgments. As mentioned in section 3. we created the hyperlinks among documents ourselves. Two different methods were used in hyperlink creation:

1. Each document came with a number. We created links between documents whose number distance was smaller than 200. This method produced more or less random links among documents.

2. We calculated cosine similarity for each pair of documents. Documents with a similarity greater than 0.8 were linked. In reality it is expected that similar documents would probably link to each other.

6. Results

Information retrieval system was evaluated using mean average precision, mean reciprocal rank and mean R-precision. Three different setups were evaluated:

- Vector space model score.
- Vector space model score refined with PageRank.
- Vector space model score refined with HITS.

Our results are presented in tables 1, 2 and 3. Each table represents a different evaluation method. We find that mean reciprocal rank is the most important evaluation method for interpretability.

Random link generation only caused a decrease in all precisions. With that in regard we decided to generate links using cosine similarity as described in section **??**. Among earlier mentioned similarity metrics we decided to use cosine similarity due to dice similarity poor results.

Table 2: Mean reciprocal rank.

VSM	VSM + PageRank	VSM + HITS
0.502	0.524	0

Table 3: Mean R-precision.

VSM	VSM + PageRank	VSM + HITS
0.246	0.245	0

Mean reciprocal rank in each setup is around 0.5 which means that on average the first relevant document will be found on the second rank in the retrieved document list.

Refining vector space model scores with PageRank shows an improvement in mean average precision and mean reciprocal rank according to tables 1 and 2. However refining vector space model with HITS doesn't show any improvement.

7. Conclusion

For this project we implemented an information retrieval system based on binary vector space model in combination with link analysis algorithms PageRank and HITS. It was shown that refining vector space model scores with PageRank improves system precision while HITS made no noticable difference.

In the future the retrieval system could be improved. An obvious idea would be to use tf-idf weighting scheme instead of binary. We could also try further parameter adjustment to achieve better results. Also the system could be tested on a dataset that already contains links among documents.

References

Judith Butcher. 1981. *Copy-editing*. Cambridge University Press, 2nd edition.

K. E. Chave. 1964. Skeletal Durability and Preservation. In J. Imbrie and N. Newel, editors, *Approaches to paleoecology*, pages 377–87, New York. Wiley.

N. Chomsky. 1973. Conditions on Transformations. In S. R. Anderson and P. Kiparsky, editors, *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

W. B. Croft. 1978. *Organizing and searching large files of document descriptions*. Ph.D. thesis, Cambridge University.

F. Feigl, 1958. *Spot Tests in Organic Analysis*, chapter 6. Publisher publisher, 5th edition.

W. W. Howells. 1951. Factors of human physique. *American Journal of Physical Anthropology*, 9:159–192.

G. B. Johnson and W. W. Howells. 1974. Title title title title title title title title title title. *Journal journal journal*.

G. B. Johnson, W. W. Howells, and A. N. Other. 1976. Title title title title title title title title title title. *Journal journal journal*.