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 $\mathbf{H}$ 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 $\mathbf{H}$. When on a page containing no hyperlinks, surfer may hyperlink to any other page therefore a stochasticity adjustment needs to be applied to $\mathbf{H}$:

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

Where $a_i = 1$ if page i has no hyperlinks and $\mathbf{e}$ 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) is calculated.
5. PageRank and HITS scores are calculated.
6. In the last step vector space model scores are refined with PageRank or HITS scores.

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.

6. Results

In scientific papers, this section usually (but not necessarily) briefly describes the related research.

6.1. First subsection

This is a subsection of the second section.

6.2. Second subsection

This is the second subsection of the second section. Referencing the (sub)sections in text is performed as follows: “in Section 6.1. we have shown . . . ”.

6.2.1. Sub-subsection example

This is a sub-subsection. If possible, it is better to avoid sub-subsections.

7. Extent of the paper

The paper should have a minimum of 3 and a maximum of 5 pages, plus an additional page for references.

8. Figures and tables

8.1. Figures

Here is an example on how to include figures in the paper. Figures are included in LaTeX code immediately *after* the text in which these figures are referenced. Allow LaTeX to place the figure where it believes is best (usually on top of the page of at the position where you would not place the figure). Figures are referenced as follows: “Figure **??** shows . . . ”. Use tilde (˜) to prevent separation between the word “Figure” and its enumeration.

8.2. Tables

There are two types of tables: narrow tables that fit into one column and a wide table that spreads over both columns.

8.2.1. Narrow tables

Table 1 is an example of a narrow table. Do not use vertical lines in tables – vertical tables have no effect and they make tables visually less attractive.

8.3. Wide tables

Table 2 is an example of a wide table that spreads across both columns. The same can be done for wide figures that should spread across the whole width of the page.

Table 2: Wide-table caption

Heading1	Heading2	Heading3
A	A very long text, longer that the width of a single column	128
B	A very long text, longer that the width of a single column	3123
C	A very long text, longer that the width of a single column	−32

Table 1: This is the caption of the table. Table captions should be placed *above* the table.

Heading1	Heading2
One	First row text
Two	Second row text
Three	Third row text
	Fourth row text

9. Math expressions and formulas

Math expressions and formulas that appear within the sentence should be written inside the so-called *inline* math environment: $2 + 3$, $\sqrt{16}$, $h(x) = \mathbf{1}(\theta_1 x_1 + \theta_0 > 0)$. Larger expressions and formulas (e.g., equations) should be written in the so-called *displayed* math environment:

$$b_k^{(i)} = \begin{cases} 1 & \text{if } k = \text{argmin}_j \|\mathbf{x}^{(i)} - \mu_j\| \\ 0 & \text{otherwise} \end{cases}$$

Math expressions which you reference in the text should be written inside the *equation* environment:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} b_k^{(i)} \|\mathbf{x}^{(i)} - \mu_k\|^2 \tag{3}$$

Now you can reference equation (3). If the paragraph continues right after the formula

$$f(x) = x^2 + \varepsilon \tag{4}$$

like this one does, use the command *noindent* after the equation to remove the indentation of the row.

Multi-letter words in the math environment should be written inside the command *mathit*, otherwise LaTeX will insert spacing between the letters to denote the multiplication of values denoted by symbols. For example, compare $Consistent(h, \mathcal{D})$ and
$Consistent(h, \mathcal{D})$.

If you need a math symbol, but you don't know the corresponding LaTeX command that generates it, try *Detexify*.[1]

10. Referencing literature

References to other publications should be written in brackets with the last name of the first author and the year of publication, e.g., (**?**). Multiple references are written in sequence, one after another, separated by semicolon and without whitespaces in between, e.g., (**?**; **?**; **?**). References are typically written at the end of the sentence and necessarily before the sentence punctuation.

If the publication is authored by more than one author, only the name of the first author is written, after which abbreviation *et al.*, meaning *et alia*, i.e., and others is written as in (**?**). If the publication is authored by only two authors, then the last names of both authors are written (**?**).

If the name of the author is incorporated into the text of the sentence, it should not be in the brackets (only the year should be there). E.g., "**?**) suggested that ...". The difference is whether you reference the publication or the author who wrote it.

The list of all literature references is given alphabetically at the end of the paper. The form of the reference depends on the type of the bibliographic unit: conference papers, (**?**), books (**?**), journal articles (**?**), doctoral dissertations (**?**), and book chapters (**?**).

All of this is automatically produced when using BibTeX. Insert all the BibTeX entries into the file `tar2016.bib`, and then reference them via their symbolic names.

11. Conclusion

Conclusion is the last enumerated section of the paper. It should not exceed half of a column and is typically split into 2–3 paragraphs. No new information should be presented in the conclusion; this section only summarizes and concludes the paper.

Acknowledgements

If suitable, you can include the *Acknowledgements* section before inserting the literature references in order to thank those who helped you in any way to deliver the paper, but are not co-authors of the paper.

[1] `http://detexify.kirelabs.org/`