XHITS: LEARNING TO RANK IN A HYPERLINKED STRUCTURE

Francisco Benjamim Filho, Raúl Pierre Rentería and Ruy Luiz Milidiú
Department of Computing, Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente, 225, Rio de Janeiro, Brazil

Keywords: Search engines, Keyword-based ranking, Link-based ranking.

Abstract: The explosive growth and the widespread accessibility of the Web has led to a surge of research activity in the area of information retrieval on the WWW. This is a huge and rich environment where the web pages can be viewed as a large community of elements that are connected through links due to several issues. The HITS approach introduces two basic concepts, hubs and authorities, which reveal some hidden semantic information from the links. In this paper, we review the XHITS, a generalization of HITS, which expands the model from two to several concepts and present a new Machine Learning algorithm to calibrate an XHITS model. The new learning algorithm uses latent feature concepts. Furthermore, we provide some illustrative examples and empirical tests. Our findings indicate that the new learning approach provides a more accurate XHITS model.

1 INTRODUCTION

Classification plays a vital role in many information management and retrieval tasks. On the Web, the link structure provides valuable information that can be used to improve information retrieval quality (Borodin et al., 2001),(Chakrabarti et al., 2001),(Lempel and Moran, 2001),(Ding et al., 2002a).

There are many different proposals for searching and ranking information on the WWW, (Mendelzon and Rafiei, 2000), (Cohn and Chang, 2000), (Giles et al., 2000), (yu Kao et al., 2003), (Fowler and Karadayi, 2002), (Ding et al., 2002b), (Agosti and Pretto, 2005), (Mizzaro and Robertson, 2007),(Lempel and Moran, 2001). Some proposals just improve the quality of existing ones by incorporating user behavior data, (Agichtein et al., 2006),(Craswell and Szummer, 2007).

In a seminal paper (Kleinberg, 1999), Jon Kleinberg introduced the notion of two fundamental categories of web pages: authorities and hubs. These categories have a mutual reinforcement relationship and to break it and classify the pages, Jon Kleinberg proposed the HITS algorithm.

However, the extended Kleinberg's approach, XHITS (Filho, 2005),(Filho et al., 2009) introduces new page categories and captures more individual judgment information from the hyperlinked environment improving the page ranking.

Furthermore, with the inclusion of new categories, emerged several parameters in the model to be calibrated. A learning process that uses a gradient descendent method has been previously proposed to calibrate these parameters (Filho et al., 2009).

This paper focuses on the learning process, proposing a novel algorithm and comparing it to the previous one. For this, we introduce a new objective function with two major components: one is the error due to the Singular Value Decomposition(SVD) of the influence matrix and the other is the average ranking error rescaled through a sigmoid function. With this new approach, we improved the ranking quality.

This paper is structured as follows. In section 2, we summarize the XHITS modeling approach. In section 3, we introduce a SVD Machine Learning procedure to calibrate the model. Next, in section 4, we examine the empirical behaviour of the SVDLP approach and compare it with Aproximate Learning Process (ALP). Finally, in section 5, we state some interesting consequences of our results and draw our conclusions.

2 PAGE ROLE CLASSIFICATION

In this section, we summarize the XHITS algorithm proposed in (Filho et al., 2009), extracting the main structures and definitions, necessary to understand the present paper.

Benjamim Filho F., Pierre Renteria R. and Luiz Milidiú R..
XHITS: LEARNING TO RANK IN A HYPERLINKED STRUCTURE.
DOI: 10.5220/0003632503770381
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 377-381
ISBN: 978-989-8425-79-9

The XHITS model extends *Hubs and Authorities Model* introducing k new categories, which are represented in each page with an addition of k weights. These weights are reinforced through the links and there are both forward and backward influences.

Furthermore, this approach presents the extended model in the matrix form and uses a learning process to calibrate the influence matrix M_σ, as can be seen in equation (1),where σ represents the query, B is the backward influence matrix, F is forward influence matrix and A_σ is the adjacent matrix of the web graph for the query σ.

The influence matrix reveals the combination of the two sources of mutual influence: link propagation and category reinforcement. The special case of symmetric reinforcement turns the matrix M_σ symmetric and the *Power Method* can be used for finding the largest eigenvalue and a corresponding eigenvector for M_σ. After sorting the eigenvector, we have the rank of the pages and can analyze the quality this rank.

$$M_\sigma = (B \otimes A_\sigma^T) + (F \otimes A_\sigma) \quad (1)$$

However, with the inclusion of new categories, emerged several parameters in the model to be calibrated that influences the quality of web pages ranked. These parameters are summarized in the matrix F, in the special symmetric case $B = F^T$ and $M_\sigma = M_\sigma^T$, as can be seen in the equation (2).

$$M_\sigma = (F \otimes A_\sigma)^T + (F \otimes A_\sigma) \quad (2)$$

Another interesting case of the XHITS model is if the matrices B and F are positive, the matrix M_σ is also positive and the authors called this *positive reinforcement*. In this case, the *Perron-Frobenius Theorem* asserts that the largest eingenvalue is positive and there is also a corresponding eigenvector with positive coordinates and this is enough to guarantee convergence of iteration.

In the way of finding the matrix F that optimizes the rank quality, the authors proposed an *Approximate Learning Process (ALP)* that minimizes training function (3), where C_σ is the cost function, $Y_{\sigma j}$ is the reference rank and $H_j(F,A_\sigma)$ is the XHITS function.

The cost function C_σ is defined as a quadratic distance between the reference rank $Y_{\sigma j}$ and the generated rank $H_j(F,A_\sigma$, which is differentiable and simple. But, during the differentiation process, the differential of the $H_j(F,A_\sigma$ is a high computational cost function and was adapted by the authors.

$$E_{train} = \frac{1}{pq} \sum_{\sigma=1}^{q} \sum_{j=1}^{p} (C_\sigma(Y_{\sigma j}, H_j(F,A_\sigma))) \quad (3)$$

Finally, the minimization process is made using the gradient descendent method. In the present work we rebuilt the above training function (3) replacing the approximate term and improve the quality of the results.

3 XHITS LEARNING WITH SVDLP

XHITS provides the vector r_σ with all page ranks, by computing and sorting the eigenvector associated with the largest eigenvalue of M_σ. Since M_σ depends on F, as we change the values of F, the eigenvector value also modifies, as well as the corresponding ranks.

Hence, we develop a Machine Learning algorithm to find the value of F that maximizes ranking quality. The required matrix must be query independent, generalizing what is found in the training set.

3.1 Learning Goal

Assume that we are given a training set $T = \{(j,\sigma,o_{j\sigma}) | j = 1,\cdots,p \text{ and } \sigma = 1,\cdots,q\}$, where j is a page-id, σ is a query-id and $o_{j\sigma}$ is the correct rank of page j for query σ.

Let $r_{\sigma j}$ be the *XHITS rank* of page j for query σ. Our learning goal here is to find F that minimizes the training error E, given by

$$E = \frac{1}{q} \sum_{\sigma=1}^{q} \left\{ \begin{array}{c} \left[\frac{1}{p} \sum_{j=1}^{p} \left(\frac{1}{1+e^{o_{\sigma j}}} - \frac{1}{1+e^{r_{\sigma j}}} \right)^2 \right] + \\ + \left[\frac{\alpha}{(cp)^2} \sum_{i=1}^{cp} \sum_{j=1}^{cp} (M_{\sigma ij} - b_{\sigma i}.r_{\sigma j})^2 \right] \end{array} \right\} \quad (4)$$

Observe that the term

$$\left[\frac{1}{p} \sum_{j=1}^{p} \left(\frac{1}{1+e^{-o_{\sigma j}}} - \frac{1}{1+e^{-r_{\sigma j}}} \right)^2 \right] \quad (5)$$

corresponds to the average ranking error due to page σ, where we rescale the rank values through a sigmoid function.

Additionally, the term

$$\left[\frac{\alpha}{(cp)^2} \sum_{i=1}^{cp} \sum_{j=1}^{cp} (M_{\sigma ij} - b_{\sigma i}.r_{\sigma j})^2 \right] \quad (6)$$

corresponds to the average largest eingenvalue estimate error contribution, as in the Singular Value Decomposition(SVD) method (Brand, 2002),(Langville et al., 2005), (Gorrell, 2006).

3.2 Gradient descent learning

To minimize E, we apply the gradient descent method. Adapting the method for our purposes, than:

$$b_\sigma^{m+1} \leftarrow b_\sigma^m - \mu_1 \frac{\partial E}{\partial b_\sigma} \tag{7}$$

$$r_\sigma^{m+1} \leftarrow r_\sigma^m - \mu_2 \frac{\partial E}{\partial r_\sigma} \tag{8}$$

$$F^{m+1} \leftarrow F^m - \mu_3 \frac{\partial E}{\partial F} \tag{9}$$

Next, we show the partial derivatives of E with respect to r_σ, b_σ and F, that is

$$\frac{\partial E}{\partial r_\sigma} = \frac{1}{q}\sum_{\sigma=1}^{q} \left\{ \begin{array}{l} \left[\frac{2}{p}\sum_{j=1}^{p}\left(\frac{1}{1+e^{o_{\sigma j}}} - \frac{1}{1+e^{r_{\sigma j}}}\right)\left(\frac{e^{r_{\sigma j}}}{(1+e^{r_{\sigma j}})^2}\right)\right] - \\ -\left[\frac{2\alpha}{(cp)^2}\sum_{i=1}^{cp}\sum_{j=1}^{cp}(M_{\sigma,i,j} - b_{\sigma i}.r_{\sigma j}).b_{\sigma i}\right] \end{array} \right\} \tag{10}$$

$$\frac{\partial E}{\partial b_\sigma} = \frac{1}{q}\sum_{\sigma=1}^{q} -\left[\frac{2\alpha}{(cp)^2}\sum_{i=1}^{cp}\sum_{j=1}^{cp}(M_{\sigma ij} - b_{\sigma i}.r_{\sigma j}).r_{\sigma i}\right] \tag{11}$$

$$\frac{\partial E}{\partial F} = \frac{1}{q}\sum_{\sigma=1}^{q} +\left[\frac{2\alpha}{(cp)^2}\sum_{i=1}^{cp}\sum_{j=1}^{cp}(M_{\sigma ij} - b_{\sigma i}.r_{\sigma j}).\frac{\partial M_{\sigma ij}}{\partial F}\right] \tag{12}$$

Finally, in the next section, we describe the algorithm for the approach discussed here.

3.3 Algorithm

Now, we describe the approximate gradient descendent learning algorithm as follows:

```
Begin
   1 Initiates r_σ, b_σ and F
     with a random value.
   2 Calculate E_train for every item of
     the training set and if it is small
     enough stop, else continue
   3 Calculate ∂E_train/∂r_σ, ∂E_train/∂b_σ and ∂E_train/∂F
   4 Calculate b_σ^{m+1}, r_σ^{m+1} and F^{m+1}
   5 Go back to step 2
End
```

4 EXPERIMENTAL RESULTS

In this section, we replicate the environment proposed in (Filho et al., 2009) and compare the results with the new approach.

4.1 Test Goal

Our major performance measure is *ranking quality*. As a first instance, we examine the XHITS model with two more categories.

Our goal is to show that the SVD learning process (SVDLP) provides a remarkable improvement over approximate learning process (ALP).

4.2 Test Environment

We adopt the same scheme proposed in (Filho et al., 2009) to build our benchmark. First, we fix a set of queries. There are 400 queries in the set with no overlaps, derived from the most Google's twenty-searched topic for each day in a period of third days.

Cross-validation is a mainstay for measuring performance and progress in machine learning. Consequently, we kept the strategy and validated the ALP and SVDLP algorithms using the 10-fold cross validation.

As in any learning process, we have to split up the benchmark in two subsets and use one for training and other for tests. We randomly generate these two subsets to avoid any interference or vicious on the learning process.

For the reference rank needed to the learning process, we made the same choice and are using an *artificial expert* (as called in (Filho et al., 2009)): Google. This choice was made for benchmarking compatibility purposes. But we already working in the new benchmark to substitute the artificial expert and make comparisons with other state of the art algorithms.

Finally, we considered the first ten pages returned by the expert as the relevant ones and we use the training set for fine-tuning two matrices F using ALP and SVDLP. After the training step, the matrices F chosen are applied in the test set.

4.3 Test Results

There are lots of different metrics that can be used for recommendation systems and information retrieval. We believe that if we can return good pages in the first ten ones, it is a good way to analyze our results and the metric P@10 reflects it. This metric represents the precision of ten first pages of results displayed. Summarizing, we considered these as the relevant pages.

We summarize the test results in table 1 . We can observe that there is already a considerable gain of XHITS over the HITS, regardless of the learning process (LP) in terms of P@10. The gain of XHITS with the approximate learning process (ALP) is approximately 260%, and with the SVD learning pro-

Table 1: Precision at 10, HITS, XHITS ALP and XHITS (SVDLP).

Algorithm	Precision at Ten (P@10)
HITS	0.143542
XHITS (ALP)	0.372455
XHITS (SVDLP)	0.519875

cess (SVDLP) is approximately 400%, both with respect to the HITS. Another important fact that can be drawn is the increased performance of XHITS with the new approach of machine learning. We obtained a 40% improvement with the new approach.

Looking inside the rankings, the best and worst case of the proximity of the ranks produced by XHITS SVDLP and Google was observed for query *oprah* and the minimum for query *michele bachmann*. The corresponding values were 0.9 and 0.1 in P@10. During the period we selected the queries, *oprah*, the star, was about to reveal something involving her family and nine of the ten first pages matched with Google's first ones. You can see the result in table 2.

Table 2: The first ten links returned by XHITS engine after the training

Position	URL
1	http://www.oprah.com/
2	http://www.oprah.com/omagazine.html
3	http://www.imdb.com/name/nm0001856/
4	http://www.tmz.com/person/oprah-winfrey/
5	http://www.nydailynews.com/topics/Oprah+Winfrey
6	http://oprahsangelnetwork.org
7	http://www.livingoprah.com/
8	http://bossip.com/category/celeb-directory/oprah/
9	http://www.myspace.com/everything/oprah-winfrey
10	http://www.quotationspage.com/quotes/OprahWinfrey/

5 CONCLUSIONS AND FUTURE WORK

We explored the fact that XHITS model provides a powerful approach and rebuild the part of the model that is an open problem: how to find the set of parameters that best fit to a given data set (Filho et al., 2009). In the way to improve the model, a new learning process using SDV for the XHITS model has been presented. Previous analysis and empirical results have shown that SVDLP performs well in XHITS model. SVDLP learns an higher precision XHITS model, when compared to ALP. This approach has its own benefits, as follows:

- the SVDLP approach has no more approximate steps;
- the training function is fully differentiable;

For testing the new approach, we chose Google as our ranking expert, because we kept the compatibility with the previous learning process, and compared the performance of HITS, XHITS ALP and XHITS SVDLP in relation with each other. The gains of XHITS SVD' model over HITS' are substantial as shown in the experimental result, over 400 % gain of quality or proximity of the Google's ranking. We are not affirming that this gain reflects necessarily the quality of the ranking, but it shows that we can learn well a judged set of pages.

For future work, we are changing the benchmark to the ClueWeb09 collection and comparing the performance with other ranking algorithms already explored and reported in the literature.

REFERENCES

Agichtein, E., Brill, E., and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA. ACM.

Agosti, M. and Pretto, L. (2005). A theoretical study of a generalized version of kleinberg's hits algorithm. *Inf. Retr.*, 8(2):219–243.

Borodin, A., Roberts, G. O., Rosenthal, J. S., and Tsaparas, P. (2001). Finding authorities and hubs from link structures on the world wide web. In *Tenth International World Wide Web Conference*.

Brand, M. (2002). Incremental singular value decomposition of uncertain data with missing values. In *Proceedings of the 7th European Conference on Computer Vision-Part I*, ECCV '02, pages 707–720, London, UK, UK. Springer-Verlag.

Chakrabarti, S., Joshi, M., and Tawde, V. (2001). Enhanced topic distillation using text, markup tags, and hyperlinks. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 208–216.

Cohn, D. and Chang, H. (2000). Learning to probabilistically identify authoritative documents. http://citeseer.ist.psu.edu/438414.html; http://www.andrew.cmu.edu/∼huan/phits.ps.gz.

Craswell, N. and Szummer, M. (2007). Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 239–246, New York, NY, USA. ACM.

Ding, C., He, X., Husbands, P., Zha, H., and Simon, H. D. (2002a). Pagerank, HITS and a unified framework for

link analysis. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Poster session, pages 353–354.

Ding, C., Zha, H., Simon, H., and He, X. (2002b). Link analysis: Hubs and authorities on the world wide web. http://citeseer.ist.psu.edu/546869.html; http://www.nersc.gov/research/SCG/cding/papers_ps/hits3.ps.

Filho, F. B. (2005). Xhits: Extending the hits algorithm for distillation of broad search topic on www. Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil.

Filho, F. B., Rentería, R. P., and Milidiú, R. L. (2009). Xhits - multiple roles in a hyperlinked structure. In Fred, A. L. N., editor, *KDIR*, pages 189–195. INSTICC Press.

Fowler, R. H. and Karadayi, T. (2002). Visualizing the web as hubs and authorities richard H. fowler and tarkan karadayi. http://citeseer.ist.psu.edu/551939.html; http://bahia.cs.panam.edu/TR/TR_CS_02_27.pdf.

Giles, C. L., Flake, G. W., and Lawrence, S. (2000). Efficient identification of web communities. http://citeseer.ist.psu.edu/347042.html; http://www.neci.nec.com/∼lawrence/papers/web-kdd00/web-kdd00.ps.gz.

Gorrell, G. (2006). Generalized hebbian algorithm for incremental latent semantic analysis. In *Proceedings of Interspeech*.

Kleinberg, J. M. (1999). Hubs, authorities, and communities. *ACM Computing Surveys (CSUR)*, 31(4es):5.

Langville, A. N., Carl, and Meyer, D. (2005). A survey of eigenvector methods of web information retrieval. *SIAM Rev*.

Lempel, R. and Moran, S. (2001). SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, 19(2):131–160.

Mendelzon, A. O. and Rafiei, D. (2000). What is this page known for? computing web page reputations. http://citeseer.ist.psu.edu/295882.html; ftp://ftp.db.toronto.edu/pub/papers/www9.ps.gz.

Mizzaro, S. and Robertson, S. (2007). Hits hits trec: exploring ir evaluation results with network analysis. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 479–486, New York, NY, USA. ACM.

yu Kao, H., ming Ho, J., syan Chen, M., and hua Lin, S. (2003). Entropy-based link analysis for mining web informative structures. http://citeseer.ist.psu.edu/572554.html; http://kp05.iis.sinica.edu.tw/shlin/paper/CIKM02.pdf.