Authoritative Documents Identification Based On Nonnegative Matrix Factorization

Nacim Fateh Chikhi, Bernard Rothenburger, Nathalie Aussenac-Gilles
Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex
{chikhi,rothenburger,aussenac}@irit.fr

Abstract

Current techniques for authoritative documents identification (ADI) suffer two main drawbacks. On the one hand, results of several ADI algorithms cannot be interpreted in a straightforward manner. This symptom is observed for instance in the HITS family algorithms. On the other hand, accuracy of some ADI algorithms is poor. For instance, PHITS overcomes the interpretability issue of HITS at the price of a low accuracy.
In this paper, we propose a new ADI algorithm, namely NHITS, which experimentally outperforms both HITS and PHITS in terms of interpretability and accuracy.

1. Introduction

When seeking information on the web (or any large collection of documents), a user is most likely to be interested in a tiny part of the web (or the collection of documents). However, even if the user's request is very specific, many documents may be relevant to the query and hundreds of documents may be returned by a search engine. Unfortunately, approaches based solely on relevance proved in practice to be unsatisfactory. Thus, sophisticated search engines try to return documents which are relevant and *authoritative* [1]. Actually, the notion of "authority" has been borrowed from bibliometrics which addresses questions such as "who is the most authoritative author in a given community?" or "what is the most authoritative journal in a given discipline?".

In his seminal paper, Kleinberg [2] proposed the HITS algorithm for authoritative documents identification (ADI). Despite its simplicity, HITS was shown to be useful in some cases where the hyperlink structure between documents is simple. Many studies have reported some limitations of the initial version of HITS and various extensions have been proposed [1]. However, as we show in section 2, HITS based approaches suffer from the *interpretability* problem i.e. their results cannot be interpreted in straightforward manner. Therefore, Cohn and Chang [3] proposed PHITS, a probabilistic model for link analysis as a better interpretable alternative to HITS. Conceptually, PHITS is very different from Kleinberg's algorithm and cannot be considered as a simple extension. In the following of the paper, we show that whereas PHITS improves interpretability, it degrades at the same time the *accuracy* of the results.

To our knowledge, no authoritative documents identification algorithm presents these two properties (i.e. interpretability and accuracy). Thus, in this paper, we propose a new link analysis algorithm which has both a good interpretability and a good accuracy.

The remainder of this paper is organized as follows. In section 2, HITS and PHITS are reviewed and discussed from a factor model viewpoint. Section 3 describes the new algorithm that we propose for authoritative documents identification. Experimental results are reported in section 4 before concluding in section 5.

2. The HITS and PHITS factor models

Currently, we analyze two of the most popular algorithms for authoritative documents identification, namely HITS and PHITS. We analyze them from the factor models point of view. Factor models have been used in many fields such as text analysis, image analysis and collaborative filtering. Factor models have been shown to be able to capture the semantics underlying the observed data using a small number of *factors* [4]. For instance, a factor model which analyzes a collection of documents is able to explain the occurrence of words in documents using a set of latent concepts. Formally, a factor model is defined as matrix decomposition of the original data matrix. Hence, given a data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$, a factor model is defined as [4]:

$$\mathbf{X} = \mathbf{AY} + \mathbf{N}$$

$\mathbf{A} \in \mathbb{R}^{N \times K}$ (factors matrix) and $\mathbf{Y} \in \mathbb{R}^{K \times D}$ (loading coefficients) are matrices on which different conditions (such as orthogonality or sparsity) may be imposed. $\mathbf{N} \in \mathbb{R}^{K \times D}$ represents the noise model associated with the factorization process.

2.1. HITS

HITS [2] is an algorithm for web community identification. It has been proposed by Kleinberg to identify hubs and authorities. Starting from a user's query, HITS constructs a citation graph which is represented by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. A Singular Value Decomposition is then performed on **A**, yielding three matrices $\mathbf{U} \in \mathbb{R}^{N \times K}$, $\mathbf{\Sigma} \in \mathbb{R}^{K \times K}$ and $\mathbf{V} \in \mathbb{R}^{K \times N}$ such that:

$$\mathbf{A}_+ = \mathbf{U}_\pm \mathbf{\Sigma}_+ \mathbf{V}_\pm + \mathbf{N}_1$$

where $\mathbf{N}_1$ is a Gausian noise model

According to the factor model syntax, the above equation can be rewritten as:

$$\mathbf{A}_+ = (\mathbf{U}\mathbf{\Sigma})_\pm \mathbf{V}_\pm + \mathbf{N}_1 = \mathbf{C}_\pm \mathbf{V}_\pm + \mathbf{N}_1$$

Usually, matrix **A** contains 0/1 values indicating the presence or not of a link between two documents. In the HITS' terminology, matrix **U** is known as the hub matrix. It corresponds to the eigenvectors of the bibliographic coupling matrix $\mathbf{A}\mathbf{A}^\mathbf{T}$. Respectively, matrix **V** is called the authority matrix. It represents the eigenvectors of the co-citation matrix $\mathbf{A}^\mathbf{T}\mathbf{A}$. $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values. The "+" (respectively "±") sign means that the matrix contains only positive values (respectively contains both positive and negative values).

The major drawback of the HITS algorithm is related to the *interpretability issue* of the discovered classes of documents, that can be interpreted as communities [3][5]. More precisely, it is well-known that the dominant eigenvector found by HITS can be easily interpreted because, according to Perron–Frobenius theorem, the left and right eigenvectors of a positive matrix contain only positive values. However, the interpretation of the non principal singular vectors is more difficult since they contain both positive and negative values. To bypass this problem, Kleinberg suggests an empirical rule to identify the communities in such situations. His heuristic consists in manually examining the positive and negative parts of each hub or authority vector. In fact, this rule is based on the observation that the relevant communities are in some cases present in the positive part, and in other cases they are found in the negative part. Clearly, the Kleinberg's rule imposes a serious limitation since we cannot automate the community identification task.

2.2. PHITS

PHITS [3] is an ADI algorithm based on the Probabilistic Latent Semantic Analysis (PLSA) model [6]. PLSA is a factor model which was initially proposed for text analysis. Basically, the PLSA's principle is that the relationship between documents and words can be explained by a small number of factors called topics. This model has been transposed by Cohn and Chang to the case of citation analysis by replacing words with citations.

Similarly to HITS, PHITS can be formulated as a matrix decomposition where an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ is decomposed as

$$\mathbf{A}_+ = \mathbf{F}_+ \Omega_+ \mathbf{G}_+ + \mathbf{N}_2$$

where $\mathbf{N}_2$ is a multinomial noise model

Notice that the above equation can be rewritten as follows according to the syntax of factor models:

$$\mathbf{A}_+ = (\mathbf{F}\Omega)_+ \mathbf{G}_+ + \mathbf{N}_2 = \mathbf{B}_+ \mathbf{G}_+ + \mathbf{N}_2$$

$\mathbf{F} \in \mathbb{R}^{N \times K}$ contains the hub probabilities, $\mathbf{G} \in \mathbb{R}^{K \times N}$ contains the authority probabilities, and $\Omega \in \mathbb{R}^{K \times K}$ is a diagonal matrix containing the probability of each community. Since matrices **F** and **G** correspond to probabilities, they are positive by definition. Thus, PHITS' results are highly interpretable unlike HITS' results which are composed of mixed sign values [3].

Although PLSA has been successfully applied to text analysis [6] and was shown to be superior to the well-known Latent Semantic Analysis through many experiments, no comparative evaluation has been carried out to validate the performances of PHITS over other ADI algorithms. A notable exception is [7] where authors compare classification accuracy of PHITS to PLSI (i.e. link versus content analysis) in many configurations. Authors report a significant superiority of PLSI over PHITS. PHITS' poor performances are actually due to its unsuitability for citation analysis. Citation (bibliographic and web) data are particular and different from other data such as texts. Especially, citation data are characterized by their large *sparsity* [1]. To illustrate this specificity, let's consider the Citeseer dataset used in Section 4. The dataset is composed of 3000 documents and 2000 unique words (after elimination of stop words and very frequent words). While the total number of word occurrences is very large (100 000), the total number of links between documents is rather small (4 000).

The cost function optimized by PHITS is known as the Kullback-Leibler divergence and is defined as [6]

$$J_{PHITS} = KL\left(\mathbf{A} \,||\, \mathbf{BG}\right) = \sum_{ij} \mathbf{A}_{ij} \log \frac{\mathbf{A}_{ij}}{\left(\mathbf{BG}\right)_{ij}} - \mathbf{A}_{ij} + \left(\mathbf{BG}\right)_{ij}$$

Minimizing the quantity J_{PHITS} is equivalent to maximizing data likelihood. However, as it is well established in the discrete data analysis community, very sparse contingency tables poses many problems to model fitting by the maximum likelihood estimation principle [8].

3. Nonnegative HITS

Thus, to avoid the interpretability issue of HITS, we propose to impose an additional constraint on the HITS' factor model which forces the model to factorize the adjacency matrix into positive matrices. Moreover, this constraint corresponds to the nature of the adjacency matrix which is always positive. According to the new specification of NHITS, the desired factor model is the one which decomposes an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ into two matrices $\mathbf{W} \in \mathbb{R}^{N \times K}$ and $\mathbf{H} \in \mathbb{R}^{K \times N}$ such that

$$\mathbf{A}_+ = \mathbf{W}_+ \mathbf{H}_+ + \mathbf{N}_3$$

where $\mathbf{N}_3$ is a Gausian noise model

Since $\mathbf{N}_3$ is Gaussian, factorization of **A** reduces to minimizing the sum of squared error between **A** and **WH** [9]. In other terms, to obtain **W** and **H** one has to minimize the following objective function:

$$J_{NHITS} = \frac{1}{2}\left\|\mathbf{A} - \mathbf{WH}\right\|^2 \quad \text{s.t. } \mathbf{W} \geq 0,\ \mathbf{H} \geq 0$$

Optimizing the problem expressed in the above equation is known as Nonnegative Matrix Factorization (NMF) [9]. It has received much of attention from the machine learning community and many algorithms have been proposed to solve it [10][11].

In NHITS, we adopt the multiplicative update rules proposed by Lee and Seung [9]. Even if the convergence of these update rules has been criticized recently by many researchers (e.g. [12]), we selected them because of their good performances and simplicity. Furthermore, in our experiments, we have tested the update rules suggested by Lin [12], but similar results to those with Lee and Seung's update rules were obtained.

3.1. Multiplicative update rules in NHITS

Using the definition $\left\|\mathbf{X}\right\| = \sqrt{tr(\mathbf{X}\mathbf{X}^{\mathbf{T}})}$, J_{NHITS} can be written as:

$$\begin{aligned} J_{NHITS} &= \frac{1}{2} tr\left(\left(\mathbf{A} - \mathbf{WH}\right)\left(\mathbf{A} - \mathbf{WH}\right)^T\right) \\ &= \frac{1}{2}\left(tr\left(\mathbf{A}\mathbf{A}^T\right) - 2tr\left(\mathbf{A}\mathbf{H}^T\mathbf{W}^T\right) + tr\left(\mathbf{W}\mathbf{H}\mathbf{H}^T\mathbf{W}^T\right)\right) \\ &\quad \text{s.t. } \mathbf{W} \geq 0,\ \mathbf{H} \geq 0 \end{aligned}$$

As indicated by Lee and Seung, the above problem has no closed-form solution. Therefore, we resort to optimization techniques and use the Lagrange multipliers method. The Lagrangian of J_{NHITS} is

$$L(J_{NHITS}) = J_{NHITS} + tr(\boldsymbol{\lambda}\mathbf{W}^T) + tr(\boldsymbol{\mu}\mathbf{H}^T)$$

where $\boldsymbol{\lambda} = \left[\lambda_{ik}\right]$ and $\boldsymbol{\mu} = \left[\mu_{kj}\right]$ are the Lagrange multipliers. The Karush-Kuhn-Tucker (KKT) necessary conditions for our optimization problem are:

$$\begin{aligned} &\left[\nabla_{\mathbf{W}} J_{NHITS}\right]_{ik} = \lambda_{ik},\ \left[\nabla_{\mathbf{H}} J_{NHITS}\right]_{kj} = \mu_{kj} \\ &\lambda_{ik} w_{ik} = 0,\ \mu_{kj} h_{kj} = 0 \\ &\lambda_{ik} \geq 0,\ \mu_{kj} \geq 0,\ w_{ik} \geq 0,\ h_{kj} \geq 0 \end{aligned}$$

where $\nabla_{\mathbf{W}} J_{NHITS}$ and $\nabla_{\mathbf{H}} J_{NHITS}$ are the gradients of J_{NHITS} with respect to W and H, respectively

$$\nabla_{\mathbf{W}} J_{NHITS} = \frac{\partial J_{NHITS}}{\partial \mathbf{W}} = -\mathbf{A}\mathbf{H}^T + \mathbf{W}\mathbf{H}\mathbf{H}^T$$

$$\nabla_{\mathbf{H}} J_{NHITS} = \frac{\partial J_{NHITS}}{\partial \mathbf{H}} = -\mathbf{W}^T\mathbf{A} + \mathbf{W}^T\mathbf{W}\mathbf{H}$$

Using the KKT conditions, we obtain

$$\begin{aligned} (\mathbf{A}\mathbf{H}^T)_{ik} w_{ik} &= (\mathbf{W}\mathbf{H}\mathbf{H}^T)_{ik} w_{ik} \\ (\mathbf{W}^T\mathbf{A})_{kj} h_{kj} &= (\mathbf{W}^T\mathbf{W}\mathbf{H})_{kj} h_{kj} \end{aligned}$$

By solving iteratively the two above equations, we are led to the following update rules

$$w_{ik} \leftarrow w_{ik} \frac{(\mathbf{A}\mathbf{H}^T)_{ik}}{(\mathbf{W}\mathbf{H}\mathbf{H}^T)_{ik}} \qquad h_{kj} \leftarrow h_{kj} \frac{(\mathbf{W}^T\mathbf{A})_{kj}}{(\mathbf{W}^T\mathbf{W}\mathbf{H})_{kj}}$$

3.2. NHITS algorithm

The complete NHITS algorithm is given in Table 1. In Step 7, a normalization is performed for two reasons. Firstly, it solves a well-known problem concerning NMF where, if **W** and **H** are solutions to the NMF problem then, matrices **WA** and **BH** s.t. **AB** = ***I***, **WA** ≥ 0 and **BH** ≥ 0 are also solutions to the NMF problem. Hence, normalizing **W** and **H** avoids this problem [10][13]. Secondly, the positive vector **M** (Step 6) containing the magnitude of each factor may be very useful to order different communities.

An example of convergence criterion (Step 5) is when the objective decrease is below a certain threshold or if a maximum number of iterations is reached.

We note also that the small constant value $\zeta = 10^{-10}$ is used in steps 2 and 3 to avoid division by zero.

Algorithm: Nonnegative HITS (NHITS).

Input: An adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ and the number of communities K.

Output: Hub matrix $\mathbf{W} \in \mathbb{R}^{N \times K}$, authority matrix $\mathbf{H} \in \mathbb{R}^{K \times N}$ and magnitude vector $\mathbf{M} \in \mathbb{R}^{K \times 1}$.

Steps:

1. Initialization: initialize $\mathbf{W}$ and $\mathbf{H}$ with random positive values, $t \leftarrow 0$;
2. Update authority value for every document in each community
   For k=1 to K
   For j=1 to N
$$\mathbf{H}_{kj}^{(t+1)} = \mathbf{H}_{kj}^{(t)} \frac{\left(\mathbf{W}^{T}\mathbf{A}\right)_{kj}}{\left(\mathbf{W}^{T}\mathbf{W}\mathbf{H}\right)_{kj} + \zeta};$$
3. Update hub value for every document in each community
   For i=1 to N
   For k=1 to K
$$\mathbf{W}_{ik}^{(t+1)} = \mathbf{W}_{ik}^{(t)} \frac{\left(\mathbf{A}\mathbf{H}^{T}\right)_{ik}}{\left(\mathbf{W}\mathbf{H}\mathbf{H}^{T}\right)_{ik} + \zeta};$$
4. $t \leftarrow t + 1$;
5. *If* a convergence criterion is not met *then* go to step 2;
6. Compute magnitude of each community
   For k=1 to K
$$\mathbf{M}_{k} = \sum_{i=1}^{N} \mathbf{W}_{ik} \times \sum_{j=1}^{N} \mathbf{H}_{kj}$$
7. Normalize columns of $\mathbf{W}$ and rows of $\mathbf{H}$ to have unit L1 norm;

Table 1 - NHITS algorithm

4. Experiments

We have compared NHITS with HITS and PHITS using two evaluation methods. On the one hand, we assessed the accuracy of each model i.e. the ability of the model to cluster citation data and to capture the embodied communities. This evaluation was performed using traditional clustering assessment measures. On the other hand, we evaluated the interpretability of factors returned by each algorithm.

4.1. Accuracy evaluation

The various algorithms we study can be regarded as unsupervised learning techniques. Therefore, many evaluation measures can be used for assessing the clustering performance of each algorithm.

4.1.1. Datasets. The first corpus we have used in our experiments is a subset of the WebKb corpus [14]. WebKb is a collection of web pages crawled from various computer science department websites.

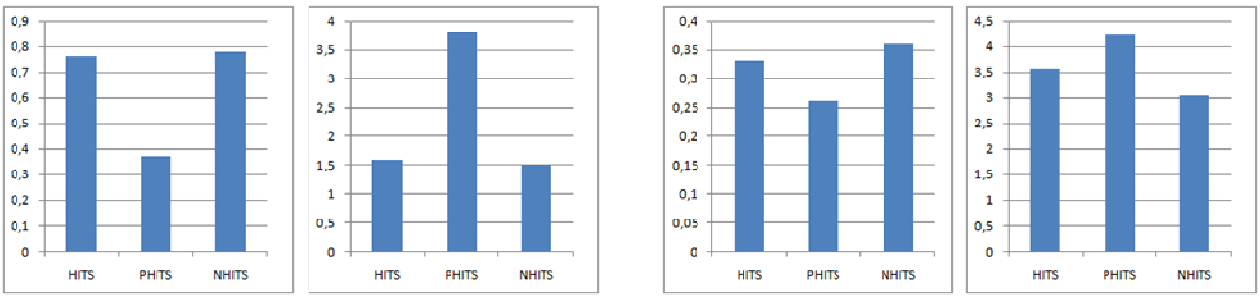
The second dataset we have used is a collection of scientific papers taken from the Citeseer database [15]. Table 2 summarizes properties of the two datasets.

4.1.2. Classification evaluation measures. To assess the clustering performance of HITS, PHITS and NHITS, we have used the classical F-measure and the more recent Variation of Information (VI) criterion. VI is an information based distance defined as [16]

$$VI(\Phi,\Pi) = H(\Phi) + H(\Pi) - 2I(\Phi,\Pi)$$

where Φ and Π are two clusterings, $H(\Phi)$ is the entropy of clustering Φ, and $I(\Phi,\Pi)$ is the mutual information between clusterings Φ and Π.

According to our evaluation measures, a perfect clustering would have an F-measure value of 1 and a VI value of 0.



(a) (b) (c) (d)

Figure 1 – F-measure (a) and Variation of Information (b) on WebKb; F-measure (c) and Variation of Information (d) on Citeseer

Dataset	*Documents*	*Links*	*Categories*	*Average links per document*	*Documents with no inlinks*
WebKb	4083	10420	4	2.55	57
Citeseer	2994	4277	5	1.43	1760
Armstrong	1503	3624	-	2.41	619

Table 2- Datasets and their properties

4.1.3. Results. We have applied HITS, PHITS and NHITS on Citeseer and WebKb using five and four factors respectively. Results (averaged over ten runs) of each algorithm are depicted in Figure 1. Notice that HITS's results do not correspond to cluster indicators because they contain both negative and positive values. Therefore, to obtain cluster indicators we apply K-means on the authority matrix returned by HITS.

We observe on Figure 1 that HITS and NHITS have almost the same performances with a slight advance for NHITS. However, results of PHITS are surprisingly very poor comparatively to HITS and NHITS. This observation is better emphasized by the VI distance values of PHITS which are very high particularly with the WebKb dataset.

4.2. Interpretability evaluation

While the accuracy of an ADI model can be assessed using quantitative measures, evaluation of the model interpretability is generally based on qualitative criteria. These criteria depend mostly on the studied application.
For our authoritative documents identification task, we define the interpretability of an ADI algorithm's results as follows: "Results of an ADI algorithm A are said to be easily interpretable if each factor returned by A corresponds to a unique community".

4.2.1. Dataset. Following the procedure suggested by Kleinberg [2], we have constructed a citation graph using the keyword "Armstrong". Statistics about this dataset are given in Table 2.

4.2.2. Results. In this set of experiments, we tried to identify the three largest components (i.e. communities) underlying the Armstrong dataset. Factors computed using HITS, PHITS and NHITS are listed in Table 3, Table 4 and Table 5 respectively.

The third factor in Table 3 illustrates the interpretability deficiency of HITS. Whereas the negative end corresponds to a community about Lance Armstrong, the positive one do not correspond to a valid community. In the second factor, the two parts (i.e. positive and negative) denote a correct community. Let's notice also that some communities seem to be repeated like the one about Lance (positive end of factor 2 and negative end of factor 3).

We observe from Table 4 that PHITS identifies communities about Louis (Factor 1) and Lance (Factor 2). Unfortunately, these two factors are mixed with other pages which are not about the main topic of the community. This undesired mixing (especially in Factor 3) was previously indicated by Cohn and Chang [3] who explained partly this phenomenon by the non orthogonality of factors. Thanks to accuracy evaluation of PHITS (Section 4.1), we can explain this mixing by the non adequacy of PHITS to citation data analysis.

Table 5 reveals that NHITS extracts correctly three homogeneous communities about respectively, the jazzman Louis Armstrong, the cyclist Lance Armstrong and a Toyota car dealer.

5. Conclusion

In this paper, we addressed an application of link mining which consists of identifying authoritative documents. We have shown that the interpretability issue of HITS can be solved simply by adding a nonnegativity constraint on the HITS' factor model. Furthermore, we found that discarding the orthogonality constraint imposed by HITS allows overlapping communities. This overlapping property is important because in practice a document may be authoritative in several communities.

Our experiments have revealed the unsuitability of PHITS to analyze sparse citation networks. Indeed, PHITS solves the interpretability problem while lowering unexpectedly the accuracy. In contrast to PHITS, the proposed algorithm gives interpretable results without decreasing accuracy.

Due to its iterative nature, NHITS (and even PHITS) performance is highly dependent on the initialization step. This suggests an improvement of NHITS which consists of using a more elaborated initialization technique instead of the random initialization currently employed.

As part of our future work, we plan to apply NHITS to applications such as the identification of authoritative social actors or the analysis of influential weblogs.

Acknowledgment

This work was supported in part by the INRIA under Grant 200601075.

References

[1] B. Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, 2006.

[2] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632, 1999.

[3] D. Cohn, and H. Chang. Learning to probabilistically identify authoritative documents. In Proc. of the 17th ICML, 2000.

[4] D. J. Bartholomew. Latent variable models and factor analysis. New York: Oxford University Press, 1987.

[5] N. F. Chikhi, B. Rothenburger, and N. Aussenac-Gilles. A comparison of dimensionality reduction techniques for web structure mining. In proc. of the IEEE/WIC/ACM international conference on Web Intelligence., Silicon Valley, 2007.

[6] T. Hofmann. Probabilistic latent semantic analysis. In Proc. of the 15th UAI Conference, 1999.

[7] M. Fisher, and R. Everson. When Are Links Useful? Experiments in Text Classification. Proceedings of the 25th European Conference on IR Research (ECIR'2003), LNCS 2633. Pisa, Italy, pp. 41—56, 2003.

[8] A. Agresti. An Introduction to Categorical Data Analysis, 2nd Edition, Wiley: New York, 2007.

[9] D. Lee, and H. Seung. Algorithms for non-negative matrix factorization. In Proc. of NIPS, pages 556–562, 2000.

[10] T. Li, and C. Ding. The Relationships Among Various Nonnegative Matrix Factorization Methods for Clustering. In Proc. of the ICDM, 362-371, 2006.

[11] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis 52(1): 155-173, 2007.

[12] C. J. Lin. On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization. IEEE Transactions on Neural Networks 18(6): 1589-1596, 2007.

[13] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual intl. ACM SIGIR conf. on Research and development in informaion retrieval. Toronto, Canada, pp. 267-273, 2003.

[14] http://www.cs.cmu.edu/~webkb/

[15] http://citeseer.ist.psu.edu/

[16] M. Meila. Comparing clusterings--an information based distance. Journal of Multivariate Analysis 98(5): 873-895, 2007.

Table 3 – HITS results (URLs in bold are spurious links)

Factor	Score	URL of page
Factor 1 (positive end)	0.402	www.satchmo.net
	0.360	www.redhotjazz.com/louie.html
	0.259	**www.lancearmstrong.com**
Magnitude: 15.54	0.225	www.rockhall.../louis-armstrong
	0.211	en.wikipedia… /Louis_Armstrong
Factor 2 (positive end)	0.420	www.lancearmstrong.com
	0.168	www.lancearmstrongfanclub.com
	0.142	www.laf.org
Magnitude: 12.72	**0.139**	**www.armstronggardens.com**
	0.128	www.thepaceline.com
Factor 2 (negative end)	-0.459	www.satchmo.net
	-0.267	www.redhotjazz.com/louie.html
Magnitude: 12.72	-0.246	www.satchography.com
	-0.163	pbskids.org/jazz/nowthen/louis.htm
	-0.109	www.npg.si.edu/exh/armstrong
Factor 3 (positive end)	**0.148**	**www.armstronggardens.com**
	0.144	**www.rockhall.../louis-armstrong**
	0.135	**www.armstrongcounty.com**
Magnitude: 11.43	**0.123**	**www.armstrongblue.com**
	0.122	**en.wikipedia…/Neil_Armstrong**
Factor 3 (negative end)	-0.549	www.lancearmstrong.com
	-0.347	www.lancearmstrongfanclub.com
	-0.281	www.laf.org
Magnitude: 11.43	-0.224	www.askmen.com/men/sports/..
	-0.179	**www.satchmo.net**

Table 4 – PHITS results (URLs in bold are spurious links)

Factor	Score	URL of page
Factor 1	0.081	www.satchmo.net
	0.064	www.redhotjazz.com/louie.html
	0.031	www.npg.si.edu/exh/armstrong
Magnitude: 0.36	**0.027**	**en.wikipedia.org/wiki/Neil_Ar…**
	0.026	**www.ohiohistory.org/places/ar...**
Factor 2	**0.095**	**www.armstrongscion.com**
	0.086	www.lancearmstrong.com
	0.058	**www.armstrongcounty.com**
Magnitude: 0.32	0.035	www.askmen.com/men/sports/...
	0.035	www.lancearmstrongfanclub.com
Factor 3	**0.037**	**www.dooce.com**
	0.029	**www.newpuzzles.com**
	0.029	**www.armstrong.com**
Magnitude: 0.31	**0.027**	**www.oldpuzzles.com**
	0.026	**www.randomhouse.com/moder...**

Table 5 – NHITS results

Factor	Score	URL of page
Factor 1	0.198	www.satchmo.net
	0.150	www.redhotjazz.com/louie.html
	0.051	www.npg.si.edu/exh/armstrong
Magnitude: 338.53	0.046	ww.satchmo.com/louisarmstrong
	0.044	www.satchography.com
Factor 2	0.262	www.lancearmstrong.com
	0.109	www.lancearmstrongfanclub.com
	0.094	www.laf.org
Magnitude: 236.46	0.080	www.askmen.com/men/sports/...
	0.059	www.thepaceline.com
Factor 3	0.960	www.armstrongscion.com
	0.013	www.toyotaofhomestead.com
	0.001	www.armstrongtoyota.com/inter...
Magnitude: 38.39	0.001	www.armstrongtoyota.com/Infor...
	0.001	www.armstrongtoyota.com/Defau...