A Two-Level Learning Hierarchy of Nonnegative Matrix Factorization Based Topic Modeling for Main Topic Extraction

Hendri Murfi

Department of Mathematics, Universitas Indonesia
Depok 16424, Indonesia
hendri@ui.ac.id

Abstract. Topic modeling is a type of statistical model that has been proven successful for tasks including discovering topics and their trends over time. In many applications, documents may be accompanied by metadata that is manually created by their authors to describe the semantic content of documents, e.g. titles and tags. A proper way of incorporating this metadata to topic modeling should improve its performance. In this paper, we adapt a two-level learning hierarchy method for incorporating the metadata into nonnegative matrix factorization based topic modeling. Our experiments on extracting main topics show that the method improves the interpretability scores and also produces more interpretable topics than the baseline one-level learning hierarchy method.

Keywords: topic modeling, nonnegative matrix factorization, incorporating metadata, nonnegative least squares, main topic extraction

1 Introduction

As our collection of digital documents continues to be stored and gets huge, we simply do not have the human power to read all of the documents to provide thematic information. Therefore, we need automatic tools for extracting the thematic information from the collection. Topic modeling is a type of statistical model that has been proven successful for this task including discovering topics and their trends over time. Topic modeling is an unsupervised learning in the sense that it does not need labels of the documents. The topics are mined from textual contents of the documents. In other words, the general problem for topic modeling is to use the observed documents to infer the hidden topic structures. Moreover, with the discovered topics we can organize the collection for many purposes, e.g. indexing, summarization, dimensionality reduction, etc [5]

Latent Dirichlet allocation (LDA) [6] is a popular probabilistic topic model. It was developed to fix some issues with a previously developed topic model probabilistic latent semantic analysis (pLSA) [9]. LDA assumes that a document typically represents multiple topics which are modeled as distributions over a vocabulary. Each word in the document is generated by randomly choosing a topic from a distribution over topics, and then randomly choosing a word from a distribution over the vocabulary. The common methods to compute posterior of the model are approximate inference techniques. Unfortunately, the maximum likelihood approximations are NP-hard [2]. As a result, several researchers continue to design algorithms with provable guarantees for the problem of learning the topic models. These algorithms include nonnegative matrix factorization (NMF) [2, 4, 1].

In many applications, the documents may contain metadata that we might want to incorporate into topic modeling. Titles and tags are examples of the metadata that usually accompany the documents in many applications. This metadata is manually created by human to describe the thematic information of documents. It becomes important because not only reflects the main topics of documents but it also has a compact form. Therefore, a proper way to incorporate this metadata to topic modeling is expected to improve the performance of topic modeling. As far as we know, the methods that address the issue of incorporating these metadata into NMF-based topic models are still rare. The simple approach to incorporate the metadata into NMF-based topic modeling is by unifying the metadata and the textual contents of documents, and then extracting topics from this union set. The union of both textual data sets may use a fusion parameter reflecting the importance of each set. We call this method as an one-level learning hierarchy (OLLH) method. Another approach is a two-level learning hierarchy (TLLH) method that is originally proposed for tag recommendations [15]. This learning method extracts topics from the textual sources separately. At the lower level, topics and topic-entity structures are discovered by a NMF algorithm from tags. Having these topic-entity structures, the extracted topics are enriched by words existing in textual contents related to the entity using a NLS algorithm at higher level. Recently, a method called nonnegative multiple matrix factorization (NMMF) is proposed [17]. This method incorporates the metadata as an auxiliary matrix that shares column with the content matrix and then decomposes both matrices simultaneously. From technical point of view, this method is similar to OLLH which extracts topics from the contents and the metadata together. Moreover, this method is applicable only for a specific NMF algorithm, i.e. multiplicative update algorithm.

In this paper, we adapt the TLLH method for main topic extraction. First the method is extended to be applicable for general NMF algorithms. At the lower level, topics is discovered by a NMF algorithm from the contents (the metadata). Given the topics and the contents (the metadata), topic-content (topic-metadata) structures are approximated using a NLS algorithm. Having these topic-content (topic-metadata) structures, the extracted topics are enhanced by words existing in the metadata (the contents) using a NLS algorithm at higher level. In contrast with OLLH, TLLH combines the vocabularies from the contents and the metadata after the learning process. Therefore, TLLH is more efficient in adapting to the characteristic of both textual sources. For example, some online news portals share complete titles and only small part of contents, but other applications may share both titles and contents in a complete form. Our experiments on extracting main topics from online news show that incorporating the metadata into topic modeling improves interpretability or coherence scores of the extracted topics. Moreover, the experiments show that TLLH is not only more efficient but it also gives higher interpretability scores than OLLH. The trends of extracted main topics over a time period may be used as background information for other applications, e.g. sentiment analysis [8, 7].

The rest of the paper is organized as follows: Section 2 discusses learning the topic model parameters using nonnegative matrix factorization. Section 3 describes our proposed two-level learning hierarchy method. In Section 4, we show our case study and results. We conclude and give a summary in Section 5.

2 Learning Model Parameters

Topic modeling has been used to various text analyzes, where the most common topic model currently in use is latent Dirichlet allocation (LDA) [6]. The intuition behind LDA is that all documents in the collection represent the same set of topics in different proportion, where each topic is a combination of words. Thus, each combination of topics is itself a distribution on words. LDA hypothesizes a Dirichlet distribution to generate the topic combinations. LDA addresses some issues with probabilistic latent semantic analysis (pLSA) [9] relating to the number of parameters to be estimated and how to deal with documents outside the training set. Both models decompose the collection of documents into groups of words representing the main topics and the new document representations indicate which topics each document has. Because of some limitation to learning the model parameters, e.g. NP-hard and getting stuck in a local minimal, several researchers continue the work to design algorithms with provable guarantees [2]. The problem for learning the topic model parameters is described in the following formulation:

> "There is an unknown topics matrix $A \in R^{n \times k}$ where $a_{ij} \geq 0$, and a stochastically generated unknown matrix $W \in R^{k \times m}$. Each column of AW is viewed as a probability distribution on rows, and for each column we are given $N << n$ iid samples from the associated distribution. The goal of this meta problem in topic modeling is to reconstruct A and parameters of the generating distribution for W" [2].

In literature, the problem of finding nonnegative matrix A, W when given the matrix AW is called nonnegative matrix factorization (NMF). Actually, research related to the NMF was initiated by Paatero and Tapper in 1994 under the name positive matrix factorization [16]. It became more widely known as nonnegative matrix factorization after Lee and Seung published some simple and useful algorithms called a multiplicative update algorithm [12]. The NMF problem is formulated as a constrained optimization problem. The algorithm had shown that the objective function value is non-increasing and claimed that the limit points of the sequence A, W is a stationary point which is a necessary condition for the local minimum [13]. However, this claim was later shown to be incorrect. A summary about the convergence of the multiplicative update algorithm is:

> ”When the algorithm has converged to a limit point in the interior of the feasible region, this point is a stationary point. This stationary point may or may not be a local minimum. When the limit point lies on the boundary of the feasible region, its stationary can not be determined” [3].

Due to the shortcoming related to convergence properties, another method called alternating nonnegative least squares (ANLS) algorithm is considered to be an alternative one. In this algorithm, a NLS step is followed by another NLS step in an alternating fashion. Let $X \in R^{n \times m} \approx AW$ be a word-document matrix and k be the number of topics, an ANLS algorithm for topic modeling is described in Algorithm 1.

Algorithm 1 ANLS algorithm

```
Given $X$ and $k$
Q = WordCooccurences(X)
Initialization $A$
while stopping criteria is not true do
    $S = \text{NLS}(Q, A) \equiv \min_{S \geq 0} \frac{1}{2} \|Q - AS\|_F^2$
    $A = \text{NLS}(Q, S) \equiv \min_{A \geq 0} \frac{1}{2} \|Q - AS\|_F^2$
end while
```

ANLS algorithm starts by forming the Gram matrix XX^T which is an empirical word-word covariance matrix. As the number of documents increases $\frac{1}{m}XX^T$ tends to a limit $Q = \frac{1}{m}E[AWW^TA^T]$, implying $Q = ARA^T$. Here, Q is a product of three nonnegative matrices. Therefore, a NMF algorithm can identify the topic matrix A if we consider Q as a product of two nonnegative matrices, A and $S = RA^T$. First, A is initialized and S is generated by NLS. Having S, A is updated by NLS in next step. Theses two processes are iterated until a stopping criteria is achieved.

Regarding to the convergence of ANLS algorithm, contrary to the multiplicative update algorithm which still lacks convergence properties, ANLS algorithms has better optimization properties. A corollary about the convergence properties of this method is:

> ”Any limit point of the sequence A, W generated by the ANLS algorithm is a stationary point” [3].

Some NLS algorithms that properly enforce non-negativity have also been proposed, e.g. projected gradient method [14], active set method [10].

Besides only convergence to stationary points, another difficulty in using NMF in practice is that the constrained optimization problem is NP-hard [18].

Therefore, additional assumptions on the data are needed to compute NMF in practice. Under a separability assumption, Arora, Ge and Moitra (AGM) present a provable algorithm than runs in polynomial time [2]. In general, the AGM algorithm works in two steps: firstly, the algorithm chooses anchor words for each topic; and then in recovery step, it reconstructs topic distribution given anchor words (Algorithm 2).

Algorithm 2 AGM algorithm

```
Given $X$ and the number of anchors $k$
Q = WordCooccurences(X)
S = AnchorWords(Q,k)
A = Recover(Q,S)
```

When we are given the exact value of ARA^T and k anchor words, we can permute the rows of A so that the anchor words appear in the first k rows and columns. Therefore, $A^T = (D, U^T)$ where D is a diagonal matrix. The key idea to the recover step is that the row sums of $DR\mathbf{1}$ and $DRA^T\mathbf{1}$ are the same because A is topic-term matrix and its columns sum up to 1, that is, $A^T\mathbf{1} = \mathbf{1}$. Having these vectors, A and R can be discovered as shown in Algorithm 3.

Algorithm 3 AGM Recover algorithm [2]

```
Permute the rows and columns of Q so that the anchor words are the first $k$ words
Compute $DRA^T\mathbf{1}$
$DR\mathbf{1} = DRA^T\mathbf{1}$
Solve for $\boldsymbol{z} : DRD\boldsymbol{z} = DR\mathbf{1}$
$A = ((DRDDiag(z))^{-1}DRA^T)^T$
```

Recently, some approaches have also been published to improve the performance of the AGM algorithm, e.g. [4], [1].

3 Two-Level Learning Hierarchy

In many applications, the documents may be accompanied by metadata that describes thematic information of the documents, e.g. titles and tags. This metadata is important because not only reflects the main topic of a document but it also has a compact form. Moreover, this metadata becomes more important when the documents contain only limited amounts of textual contents, e.g. most of online news articles that are shared in RSS formats have titles and only some first sentences of the contents. Therefore, a proper way to incorporate this metadata to NMF-based topic modeling is needed to reach the best performance of topic modeling.

A simple approach to incorporate the metadata into topic modeling is by unifying the contents and the metadata, and then extracting topics from this union. We call the method as an one-level learning hierarchy (OLLH) method. The union of both textual data sets may use a fusion parameter reflecting the importance of each set. The parameter needs to be determined before the learning phase that is usually time consuming. To obtained the optimal fusion parameter, we need to repeat the learning process for every selected parameter. Therefore, this approach is not so efficient in adapting the models to the characteristic of both textual sources.

To overcome the weakness of OLLH, we adapt a two-level learning hierarchy (TLLH) method which extracts topics from the contents and the metadata separately [15]. At the lower level, topics are discovered by a NMF algorithm from the contents (the metadata) and topic-document structures are estimated by a NLS algorithm using the extracted topics and the contents (the metadata). Having these structures, the extracted topics may be enriched by words existing in the metadata (the contents) using NLS algorithms at higher level. Using this mechanism, the fusion parameter is optimized after the learning process. In other words, the learning process is executed only one time before the optimization of the fusion parameter. Due to the time consuming of the learning process, TLLH becomes more efficient in adapting the models to both textual data than OLLH.

Let $X \in R^{a \times m}$ be a word-content (word-metadata) matrix and k be the number of topics. Given X and k, firstly TLLH executes a NMF algorithm that produces a topic matrix $A \in R^{a \times k}$, that is:

$$A = \text{NMF}(X) \tag{1}$$

Next, the topic-document structure matrix $W \in R^{k \times m}$ can be expected using a NLS algorithm as described in the following Equation:

$$W = \text{NLS}(X, A) \tag{2}$$

Having the topic-document structure matrix W, TLLH examines vocabularies from a word-metadata (word-content) matrix $Y \in R^{b \times m}$ to enrich the extracted topics using the NLS algorithm, that is:

$$B = \text{NLS}(Y, W) \tag{3}$$

Therefore, the final topics matrix $T \in R^{c \times k}$ can be constructed by the following equation:

$$T = (1 - \alpha)\tilde{A} + \alpha\tilde{B} \tag{4}$$

where c is the number of vocabularies which are a union of vocabularies derived from the metadata and the contents, $\tilde{A} \in R^{c \times k}$ is matrix A related to the new vocabulary set, $\tilde{B} \in R^{c \times k}$ is matrix B related to the new vocabulary set, and α is a fusion parameter.

4 A Case Study of Main Topic Extraction from Online News

The development of information and communication infrastructures has encouraged an increasing number of internet users in Indonesia. At the end of 2013 it was noted that 28% of population, around 71 million people, have been using the internet. The increasing number of internet users is very significant when compared to internet users in 2000, which was only about 2 million users. The increasing number of Indonesian who accesses the internet is followed the development of applications and internet contents related to Indonesian. One of the internet contents that is widely used today is online news. All famous news agencies have published their articles online on the internet.

The news portals submit large number of articles in order to provide the most actual news. For example, the popular news portal Detikcom submits about 200 articles per day. As the digital news articles continue to be released, we simply do not have the human power to read each of them to provide the main topics and their trend on a given time period. Therefore, we need to develop automatic tools for this task.

NMF-based topic modeling is one of the automatic tools that we consider for the task. We assume that the most frequent topics on a given time period would be the main topics on that period. Given the topic-document structure matrix $W \in R^{k \times m}$, the most frequent topic is:

$$\max_{i \in \{1..k\}} \sum_{j=1}^{m} w_{ij} \tag{5}$$

To provide the main topics of Indonesian news on a given time period, we analyze digital news articles that are shared online through RSS feeds by nine Indonesian news portals that are widely known in Indonesia, i.e. Antara (antaranews.com), Detik (detik.com), Inilah (inilah.com), Kompas (kompas.com), Okezone (okezone.com), Republika (republika.co.id), Rakyat Merdeka (rmol.co), Tempo (tempo.co) and Viva (viva.co.id). The news articles contain published dates, titles and some first sentences of contents.

For the simulations, we use news articles from three time periods, i.e. January 2013, February 2014 and March 2014. The number of news articles is an average of 43000 articles per month. For creating the word-content matrices, the word-title matrices, and the word-union matrices, contents, titles and unions are parsed and vocabularies are created using a standard tokenization method. The non alphabet characters are removed and standard stop words of Indonesian are applied. Finally, the matrices are weighted by a term frequency inversed document frequency (TFIDF) weighting scheme. The statistics of the experimental data are described in Table 1.

Pointwise mutual information (PMI) between the topic words is used for estimating the interpretability or coherence score of a topic [11]. Let $t = \{w_1, w_2, ..., w_r\}$ be a topic having r words, the PMI score of t is given in Equation 6. In our ex-

Table 1. The statistics of the experimental data. The data were collected monthly from January, 2014 to March, 2014

Period	**Articles**	**Contens Vocabularies**	**Title Vocabularies**	**Union Vocabularies**
January	50304	40884	23692	44605
February	46834	39934	23378	43738
March	31855	34064	19797	37381

periment, articles that published from August 2013 to March 2014 are used as a reference corpus.

$$\mathrm{PMI}(t) = \sum_{j=2}^{r} \sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_i)P(w_j)} \tag{6}$$

We fit a 100-topic NMF model to each experimental data, that is, contents, titles and union of contents and titles. For OLLH, the model extracts topics from the union of contents and titles. Two types of TLLH are considered: TLLH-content - the model extracts topics from the contents and then they are enriched by vocabularies of the titles, and TLLH-title - the model extracts topics from the titles and then they are enhanced by vocabularies of the contents.

In TLLH, first we need to optimize the fusion parameter α. The parameter reflects the importance of contents and titles as sources of the topics vocabularies. α equals to zero indicates that the vocabularies of contents are not considered in constructing topics. In other words, the topics are built only by the vocabularies of titles. While α equals to one means that only the vocabularies of content that make up the topics.

Figure 1 and Figure 2 give the average PMI score of the top 10 most frequent topics for various α values. The PMI score of a topic is calculated based on the top 10 most frequent words of the topic. From Figure 1, the optimal fusion parameters of TLLH-Content can be selected for each time period. They are 0.0, 0.3 and 0.1 for January, February and March, respectively. These parameters show that the optimal interpretability scores of extracted topics are highly influenced by the content's words. In other words, the enrichment of extracted topics with title's words does not gives significant impact to the interpretability scores. From Figure 2, we see that the optimal fusion parameters of TLLH-Title are 1.0, 0.8 and 0.7. These results show that the words of the contents also give better interpretability scores for TLLH-Title. An explanation for these results is that each content has larger number of words than its corresponding title. Therefore, the TFIDF weighting is more informative to represent the importance of the content's words than the title's words.

After getting the optimal fusion parameter, we compare the interpretability of TLLH-Content and TLLH-Title with other approaches, i.e. Content, Title and OLLH. The similar procedure is applied to get the optimal fusion parameter of OLLH. Table 2 shows the comparison of the average PMI score for the top 10

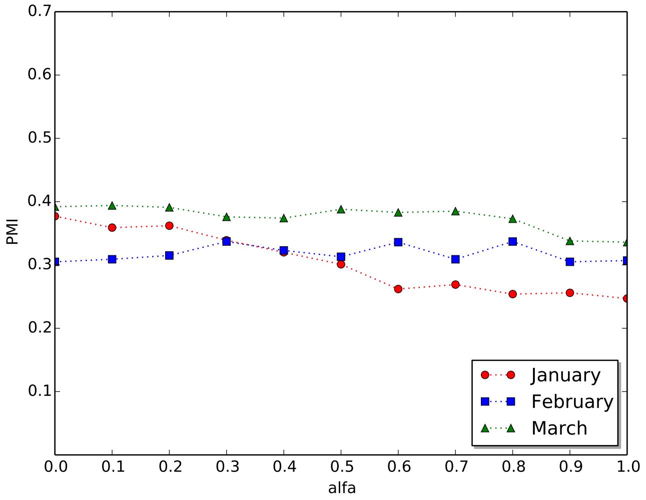

Fig. 1. The average PMI score of the top 10 most frequent topics extracted by TLLH-Content for various α values

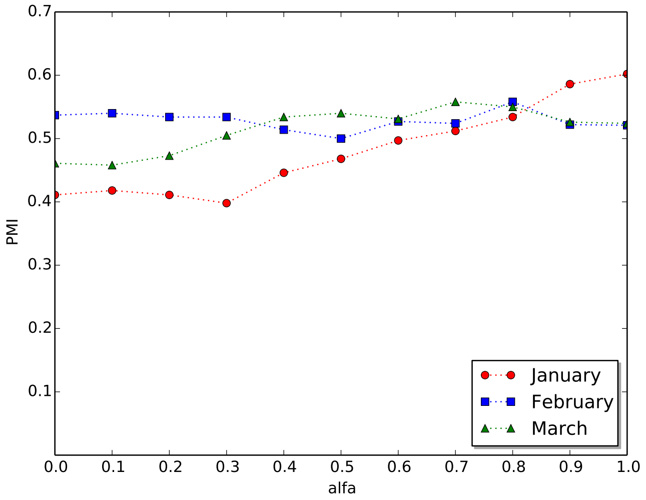

Fig. 2. The average PMI score of the top 10 most frequent topics extracted by TLLH-Title for various α values

most frequent topics extracted by each approach. All scores are calculated based on the top 10 most frequent words for each topic. From these results we see that OLLH and TLLH improve the interpretability scores of the Content approach. It means that incorporating metadata is a potential approach to improve the interpretability score. Among three incorporating methods, TLLH-Title gives the best results. It produces topics that have the interpretability score an average of 62% better than topics extracted from only the contents. Its scores are also an average of 18% higher than OLLH and 56% higher than TLLH-Content. An possible reason for these results is that TLLH-Title extracts the topics from the titles that is manually created by their authors to describe the thematic contents and in a compact form, while other methods examine the topics also from the contents that are usually not in a complete form.

Table 2. The comparison of the average PMI score for the top 10 most frequent topics. Each topic is represented by the top 10 most frequent words

Methods	**January**	**February**	**March**
Content	0.377	0.305	0.392
Title	0.411	0.537	0.461
OLLH	0.465	0.537	0.461
TLLH-Content	0.377	0.337	0.394
TLLH-Title	0.602	0.558	0.558

Finally, we visualize the trends of the top 10 most frequent topics over March 2013 time period. Figure 3 is the trends of topics produced by TLLH-Topic. From this figure, we see that the method recognizes Topic 1 as the main topic of this time period. Its theme is about *campaign in Indonesia's election 2014*. This theme is manually interpreted from the top 10 most frequent words of the topic, i.e. *kampanye* (campaign), *terbuka* (open), *partai* (party), *cuti* (time off work), *parpol* (political party), *pemilu* (election), *jadwal* (schedule), *berkampanye* (campaign), *juru* (campaigners), *politik* (politics).

From Figure 3, we can also spot some extraordinary topics on some specific days, e.g. Topic 3 on March 9, 2014. It is about *the missing of airplane MH370.* This interpretation is concluded based on its following words: malaysia, airlines, *pesawat* (plane), *hilang* (disappeared), *hilangnya* (disappearance), *penumpang* (passenger), *pencarian* (search), kuala, lumpur, *penerbangan* (flight). Most of the news portals released articles about this topic that made the topic to be the most frequent topic on that day and some days after.

The trends of extracted main topics may be used as background information for other applications, e.g. sentiment analysis [8, 7], to draw relationships among the main topics of news and the sentiments of entities.

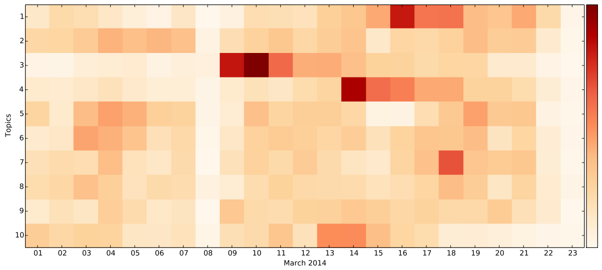

Fig. 3. Trends of the top 10 most frequent topics extracted by TLLH-Title over March 2014 time period

5 Conclusion

In this paper, we examine the problem of incorporating metadata into NMF-based topic modeling. Besides a simple one-level learning hierarchy method, we adapt a two-level learning hierarchy method for this task. Our experiments on the problem of main topic extraction show that these methods improve interpretability scores of the extracted topics. Moreover, the two-level learning hierarchy methods can achieve higher scores than the one-level learning hierarchy version.

Acknowledgment

This work was supported by Universitas Indonesia under PUPT 2013 grant no. 2784/H2.R12/HKP.05.00/2013. Any opinions, findings and conclusions or recommendations are the authors' and do not necessarily reflect those of the sponsor.

References

1. S. Arora, R. Ge, Y. Halpern, D. Mimno, and A. Moitra. A practical algorithm for topic modeling with provable guarantees. In *proceeding of the 30th International Conference on Machine Learning*, 2013.
2. S. Arora, R. Ge, and A. Moitra. Learning topic models-going beyond svd. In *Proceeding of the IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10, 2012.
3. 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 and Data Analysis*, 15(1):155–173, 2007.
4. V. Bittorf, B. Recht, C. Re, and J. A. Tropp. Factoring nonnegative matrices with linear programs. In *Neural Information Processing Systems*, 2012.

5. D. M. Blei. Probabilistic topic models. *Communication of the ACM*, 55(4):77–84, 2012.
6. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
7. E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.
8. R. Feldman. Techniques and applications for sentiment analysis. *Communication of the ACM*, 56(4):82–89, 2013.
9. T. Hofmann. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence*, pages 289–296, 1999.
10. H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, 30(2):713–730, 2008.
11. J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014.
12. D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.
13. D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing System*, pages 556–562, 2001.
14. C.-J. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19:2756–2779, 2007.
15. H. Murfi and K. Obermayer. A two-level learning hierarchy of concept based keyword extraction for tag recommendations. In *Proceedings of the ECML PKDD Discovery Challenge 2009*, pages 201–214, 2009.
16. P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.
17. K. Takeuchi, K. Ishiguro, A. Kimura, and H. Sawada. Non-negative multiple matrix factorization. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 1713–1720, 2013.
18. S. A. Vavasis. On the complexity on nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.