# Stiefel 다양체에서 곱셈의 업데이트를 이용한 비음수 행렬의 직교 분해
# (Orthogonal Nonnegative Matrix Factorization: Multiplicative Updates on Stiefel Manifolds)

유 지 호 †　　최 승 진 ††
(Jiho Yoo)　　(Seungjin Choi)

**요 약** 주어진 비음수 데이터를 두 개의 비음수 행렬의 곱의 형태로 표현하는 비음수 행렬 분해(Nonnegative Matrix Factorization) 는 비음수 데이터의 다변량 분석에서 폭넓게 사용되고 있는 방법이다. 비음수 행렬 분해는 집단화(Clustering), 특히 문서의 집단화에서 유용하게 쓰일 수 있다. 본 논문에서는 주어진 문서들로부터 구성된 단어-문서 행렬을 두 개의 비음수 행렬의 곱으로 분해할 때, 그 중 하나의 행렬에 직교 제한을 주는 비음수 행렬의 직교 분해(Orthogonal Nonnegative Matrix Factorization) 방법을 다룬다. 현존하는 비음수 행렬의 직교 분해 방법은 직교 제한과 관련된 항을 더해주는 방식을 사용하지만, 여기서는 Stiefel 다양체 위에서의 실제 기울기를 직접 구하여 곱셈의 업데이트 알고리즘을 유도하였다. 다양한 문서 데이터에 대한 실험을 통해 새롭게 유도된 비음수 행렬의 직교 분해 방법이 기존의 비음수 행렬 분해나 기존의 비음수 행렬의 직교 분해보다 문서 집단화에서 우수한 성능을 나타냄을 보였다.
**키워드** : 문서 집단화, 비음수 행렬 분해, Stiefel 다양체

***Abstract*** Nonnegative matrix factorization (NMF) is a popular method for multivariate analysis of nonnegative data, the goal of which is decompose a data matrix into a product of two factor matrices with all entries in factor matrices restricted to be nonnegative. NMF was shown to be useful in a task of clustering (especially document clustering). In this paper we present an algorithm for orthogonal nonnegative matrix factorization, where an orthogonality constraint is imposed on the nonnegative decomposition of a term-document matrix. We develop multiplicative updates directly from true gradient on Stiefel manifold, whereas existing algorithms consider additive orthogonality constraints. Experiments on several different document data sets show our orthogonal NMF algorithms perform better in a task of clustering, compared to the standard NMF and an existing orthogonal NMF.
**Key words** : document clustering, nonnegative matrix factorization, Stiefel manifold

## 1. 서 론

비음수 행렬 분해(Nonnegative Matrix Factorization)는 이미지, 스펙트로그램, 문서 등의 비음수 행렬 형태의 데이터로부터 의미 있는 표현을 얻어내는 데에 사용되는 다변량 데이터 분석 방법이다[1]. 비음수 행렬 분해는 주어진 데이터 행렬을 데이터의 기저를 나타내는 행렬과, 그 기저들의 활성화 정도를 나타내는 인코딩 행렬로 분해한다. 일반적인 비음수 행렬 분해에서는, 분해를 통해서 얻은 기저 행렬과 활성화 행렬의 원소들 역시 모두 비음수로 표현되어야 한다. 이렇게 비음수 기저 행렬의 비음수 활성화를 얻음으로써, 주어진 비음수 데이터를 기저들의 합에만 의존하는 조합으로 표현할 수 있으며, 결과적으로 데이터를 구성하는 각 부분들의

· 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원 사업의 연구결과로 수행되었음(IITA-2009-C1090-0902-0045)
· 이 논문은 제35회 추계학술대회에서 'Stiefel 다양체에서 곱셈의 업데이트를 이용한 비음수 행렬의 직교 분해'의 제목으로 발표된 논문을 확장한 것임
† 학생회원 : 포항공과대학교 컴퓨터공학과
zentasis@postech.ac.kr
†† 종신회원 : 포항공과대학교 컴퓨터공학과 교수
seungjin@postech.ac.kr
논문접수 : 2008년 12월 13일
심사완료 : 2009년 3월 12일

정보과학회논문지: 소프트웨어 및 응용 제36권 제5호(2009.5)

조합으로 표현할 수 있다[1]. 지역성(locality)이나 직교성(orthogonality) 등의 추가적인 제약 조건을 부여하는 것이 더 지역적인 성분을 찾아내거나, 보다 성긴(sparse) 표현을 이끌어 냄으로써 순수한 비음수 행렬 분해보다 유용한 결과를 이끌어 낼 수 있다[2]. 집단화(clustering) 를 위한 비음수 행렬 분해의 적용 방법과, 직교 제한의 부여를 통한 성능의 개선 방법이 소개된 바 있다[3].

단어-문서 행렬의 분해를 통한 문서의 집단화는 비음수 행렬 분해의 유용한 적용 분야들 중 하나이다[4,5]. 여기서는 문서의 집단화를 위하여, 단어-문서 행렬의 분해에 직교 제한을 부여한 비음수 행렬의 직교 분해 방법을 소개한다. 기존의 비음수 행렬의 직교 분해 방법이 직교 제한을 나타내는 항을 더함으로써 직교 분해를 수행하는 것과는 달리, 여기서는 직교 행렬들로 구성된 Stiefel 다양체 위에서의 실제 기울기를 이용하여, 곱셈의 업데이트를 사용한 알고리즘을 개발하였다. 여러 개의 문서 데이터 모음에 대한 실험을 통하여 제안된 비음수 행렬의 직교 분해 방법이 기존의 비음수 행렬 분해 방법이나 현존하는 비음수 행렬의 직교 분해 방법보다 우수한 성능을 나타냄을 보였다.

## 2. 문서 집단화를 위한 비음수 행렬 분해

다수의 문서로 구성된 문서 데이터 집합을 벡터 공간에서 표현하기 위해서는, $m$개의 단어로 이루어진 사전이 문서 데이터 집합을 생성하는 데에 사용되었다고 할 때, $m$개의 단어를 각각의 벡터 성분으로 사용하여 각각의 문서를 $m$ 차원의 벡터 $x_t \in \mathbb{R}^m$ 로 구성하는 방법을 사용한다. 이 같은 방법을 사용하면 전체 문서 데이터 집합 안에 있는 총 N 개의 문서를 사용하여 하나의 단어-문서 행렬 $X \in \mathbb{R}^{m \times n}$ 를 구성할 수 있다. 이 때, 단어-문서 행렬 내의 하나의 원소 $X_{ij}$ 는 단어 $t_i$ 가 문서 $d_j$ 에서 얼마나 중요하게 사용되는지를 나타내는 값으로, 다음과 같이 계산된다.

$$X_{ij} = \mathrm{TF}_{ij} \log\left(\frac{N}{\mathrm{DF}_i}\right)$$

여기서 $\mathrm{TF}_{ij}$ 는 문서 $d_j$ 에서 전체 단어들의 총 사용 횟수에 비해 단어 $t_i$ 가 얼마의 비율로 사용되었는지를 나타내는 항이며, $\mathrm{DF}_i$ 는 단어 $t_i$ 가 사용된 문서의 개수를 나타내는 항이다. 이러한 방식으로 단어-문서 행렬을 구성하면, 각각의 원소 $X_{ij}$ 는 항상 비음수가 되며, 나타나지 않은 단어에 대해서 0의 값을 갖게 된다.

비음수 행렬 분해는 비음수 행렬 $X \in \mathbb{R}^{m \times N}$ 가 주어졌을 때 다음과 같은 형태로 분해되는 행렬 $U$ 와 $V$ 를 찾는다.

$$X \approx UV^T, \tag{1}$$

여기서 행렬 $U \in \mathbb{R}^{m \times K}$ 와 행렬 $V \in \mathbb{R}^{N \times K}$ 는 비음수 행렬이어야 한다. $K$ 는 집단화의 경우 집단의 개수를 나타내는 수이다. 일반적으로 비음수 행렬 분해를 통해 얻은 행렬 $U$ 와 행렬 $V$ 는 다음과 같이 해석된다.

- 데이터 행렬 $X$ 에서 각각의 열이 $m$ 차원 벡터 공간에 존재하는 각각의 데이터를 표현하는 데에 사용되었을 경우, 행렬 $U$ 의 열들이 기저 벡터를 나타내고, 행렬 $V$ 의 행들은 각각의 기저 벡터가 데이터를 재구성하는데 얼마만큼의 비율로 쓰이는지를 나타내는 인코딩 값이 된다.
- 위와는 달리, 데이터 행렬 $X$의 각각의 행이 $N$차원 위에 존재하는 각각의 데이터를 나타내고 있는 경우에는, 행렬 $V$의 열이 기저 벡터가 되고, 행렬 $U$의 각각의 행이 인코딩을 나타낸다.

단어-문서 행렬에 비음수 행렬 분해 방법을 적용함으로써 문서의 집단화를 수행할 수 있다. 일반적으로 단어-문서 행렬 $X$의 각각의 열이 $m$차원 공간에서 하나의 문서를 나타내는 형태를 갖는다. 이러한 경우, (1) 식과 같은 형태의 비음수 행렬 분해는 다음과 같이 해석될 수 있다.

- 원소 $U_{ij}$ 는 단어 $t_i$ 가 집단 $c_j$ 에서 어느 정도의 중요도를 가지고 있는지를 나타낸다. 행렬 $U$ 의 $j$ 번째 열 $u_j$ 는 집단 $c_j$ 의 표준, 혹은 중심을 나타내는 벡터가 된다.
- 원소 $V_{ij}$ 는 문서 $d_i$ 가 집단 $c_j$ 에 속하는 정도를 나타낸다. 적절한 정규화 과정을 거치면 $V_{ij}$ 의 값은 문서 $d_i$ 가 주어졌을 때 집단 $c_j$ 의 사후 확률을 나타내게 된다.

### 2.1 비음수 행렬 분해를 위한 곱셈의 업데이트

유클리디안 거리의 제곱을 데이터 행렬 $X$ 와 모델 $UV^T$ 의 차이를 측정하기 위하여 사용하면, 다음과 같은 최소 자승 오차 함수를 얻을 수 있다.

$$\varepsilon = \frac{1}{2} \| X - UV^T \|^2 \tag{2}$$

이 때 비음수 행렬 분해는 다음 최적화 문제로 표현할 수 있다.

$$\operatorname{arg\,min}_{U \geq 0,\, V \geq 0} \frac{1}{2} \| X - UV^T \|^2 \tag{3}$$

덧셈의 업데이트를 사용한 최대 기울기 방향 감소 학습을 이용하여 (3)의 최적화 식의 해답을 구할 수 있다. 하지만 행렬 $U$ 와 행렬 $V$ 의 비음수 제약 조건이 업데이트 과정에서 지켜진다는 보장이 없으므로, 제약 조건의 준수를 위해서는 추가적인 장치가 필요하다.

이러한 추가적인 장치 없이 (3)의 최적화 식에 대한 해답을 구할 수 있는 방법으로 곱셈에 기반한 방법이 제시되었다 [6]. 여기에서는 [6]에서 제시된 방법과 다른 접근을 통하여 같은 곱셈의 업데이트 방법을 유도하는

방법을 소개한다. 오차 함수의 최대 감소 방향 식은 다음과 같이 양수의 부분과 음수의 부분으로 분해하여 표시할 수 있다.

$$\nabla\varepsilon = [\nabla\varepsilon]^{+} - [\nabla\varepsilon]^{-}, \tag{4}$$

위의 식에서 각각의 항은 $[\nabla\varepsilon]^{+} > 0$ 과 $[\nabla\varepsilon]^{-} > 0$의 조건을 만족하여야 한다. 이 때, 인수 $\Theta$에 대한 곱셈의 업데이트 식은 다음과 같은 형태로 계산할 수 있다.

$$\Theta \leftarrow \Theta \odot \left( \frac{[\nabla\varepsilon]^{-}}{[\nabla\varepsilon]^{+}} \right)^{.\eta}, \tag{5}$$

위의 식에서 기호 $\odot$는 Hadamard (행렬 원소 기준) 곱셈 을 나타내며, 기호 $(\bullet)^{.\eta}$ 역시 행렬의 원소를 기준으로 하는 거듭 제곱을 나타낸다. $\eta(0 < \eta \le 1)$의 값은 학습량으로 생각할 수 있다. 여기서 곱셈의 업데이트식 (5)가 인수 $\Theta$의 비음수 제약 조건을 유지시킴과 동시에 수렴하는 상황에서 $\nabla\varepsilon = 0$을 만족시킴을 알 수 있다.

오차 함수 (2)를 행렬 $V$를 고정시킨 채로 행렬 $U$를 미분한 식과, 행렬 $U$를 고정시킨 채로 행렬 $V$에 대해 미분한 식은 다음과 같이 계산된다.

$$\nabla_U\varepsilon = [\nabla_U\varepsilon]^{+} - [\nabla_U\varepsilon]^{-} = UV^TV - XV, \tag{6}$$

$$\nabla_V\varepsilon = [\nabla_V\varepsilon]^{+} - [\nabla_V\varepsilon]^{-} = VU^TU - XU. \tag{7}$$

이 결과를 곱셈의 업데이트식 (5)에 학습량 $\eta = 1$로 하여 대입하면 잘 알려진 비음수 행렬 분해의 업데이트 식 [6]을 얻을 수 있다.

$$U \leftarrow U \odot \frac{XV}{UV^TV}, \tag{8}$$

$$V \leftarrow V \odot \frac{X^TU}{VU^TU}, \tag{9}$$

위의 식에서 분수로 표시된 나누기 역시 행렬의 원소별 나누기를 의미한다.

### 2.2 확률 표현과 정규화

비음수 행렬 분해의 확률적인 해석과, I-divergence를 기준으로 한 비음수 행렬 분해와 확률적 잠재 의미 색인화(Probabilistic latent semantic indexing)와의 동치 관계가 [7]에서 소개된 바 있다. 단어와 문서의 공통 확률 $p(t_i, d_j)$는 다음과 같이 분해하여 표현할 수 있다.

$$\begin{aligned} p(t_i, d_j) &= \sum_k p(t_i, d_j | c_k) p(c_k) \\ &= \sum_k p(t_i | c_k) p(d_j | c_k) p(c_k) \end{aligned} \tag{10}$$

여기서 $p(c_k)$ 는 집단 $c_k$의 선행 확률이 된다. 단어-문서 행렬의 각각의 원소 $X_{ij}$는 본질적으로는 단어와 문서의 공통된 발생 정도로 볼 수 있다. 따라서 확률로서의 조건 $\Sigma_i\Sigma_j X_{ij} = 1$를 만족하도록 정규화를 시켜주면, 정규화된 $X_{ij}$의 값이 공통 확률 $p(t_i, d_j)$ 를 나타낸다고 볼 수 있다. 즉, 1로 구성된 벡터 $1 = [1, \dots 1]^T$를 사용하여 $1^TX1$의 형태로 전체 원소들의 합을 구한 후, 각각의 원소를 구한 합의 값으로 나누어 주는 방식으로 주어진 단어-문서 행렬이 공통 확률을 표현하도록 정규화 할 수 있다.

식 (10)의 확률 분해식을 식 (1)의 행렬 분해식과 연관지어 생각해 볼 수 있다. (1) 식에서의 원소 $U_{ik}$는 단어 $t_i$가 집단 $c_k$에서 갖는 중요도를 나타내는 값으로, 적절한 정규화를 거쳐 확률 $p(t_i|c_k)$를 표현하게 만들 수 있다. 행렬 $U$의 각각의 열의 합이 1이 되도록 $D_U \equiv \mathrm{diag}(1^TU)$행렬을 사용하여 $UD_U^{-1}$ 형태로 정규화를 해 주면, 다음과 같은 정확한 관계를 얻을 수 있다.

$$[UD_U^{-1}]_{ik} = p(t_i|c_k).$$

행렬 $X$가 확률 조건 $\Sigma_i\Sigma_j X_{ij} = 1$을 만족하도록 정규화가 되어 있다고 하면, 비율 조정 행렬 $D_V \equiv \mathrm{diag}(1^TV)$를 사용하여 (1) 식의 행렬 분해를 다음과 같이 표현할 수 있다.

$$X = (UD_U^{-1})(D_UD_V)(VD_V^{-1})^T. \tag{11}$$

(11) 식을 (10) 식의 확률 표현과 비교하면, 행렬 $D \equiv D_UD_V$ 에서의 각각의 대각 원소가 집단의 선행 확률 $p(c_k)$를 나타내게 됨을 알 수 있다.

집단화 문제를 풀기 위해서는, 문서에 대한 집단의 사후 확률 $p(c_k|d_j)$를 계산할 수 있어야 한다. 베이즈 공식을 사용하면 사후 확률을 문서의 우도와 집단의 선행 확률을 사용하여 표현할 수 있다. 즉, 사후 확률 $p(c_k|d_j)$를 다음과 같이 구할 수 있다.

$$\begin{aligned} p(c_k|d_j) &\propto p(d_j|c_k)p(c_k) \\ &= [D(VD_V^{-1})^T]_{kj} \\ &= [(D_UD_V)(D_V^{-1}V^T)]_{kj} \\ &= [D_UV^T]_{kj} \end{aligned} \tag{12}$$

(12) 식을 통해 대각 행렬 $D_u$를 사용하여 정규화된 행렬 $(VD_U)^T$가 문서에 대한 집단의 사후 확률을 표현하게 됨을 알 수 있다. 따라서, 문서 $d_j$는 다음과 같은 조건을 만족할 때 집단 $k^*$로 집단화 할 수 있다.

$$k^* = \mathrm{argmax}_k [VD_U]_{jk}.$$

비음수 행렬 분해를 이용한 문서 집단화는 [4]에서 처음 소개된 바 있으나, 여기서는 다른 형태의 정규화 방법을 사용한다. 전체 알고리즘은 아래와 같이 정리된다.

**알고리즘: 비음수 행렬 분해를 이용한 문서 집단화**

1. 단어-문서 행렬 $X$를 만든다.
2. 비음수 행렬 분해를 이용, $X = UV^T$로 분해한다.
3. 행렬 $U$와 행렬 $V$를 $D_u = 1^TU$를 이용하여 다음과 같이 정규화한다:

$U \leftarrow UD_U^{-1}$
$V \leftarrow VD_U$

4. 문서 $d_j$를 다음 조건을 만족하는 집단 $k^*$에 속하도록 집단화 한다.

$k^*=\text{argmax}_k[VD_U]_{jk}$

## 3. 문서 집단화를 위한 비음수 행렬의 직교 분해

비음수 행렬의 직교 분해(Orthogonal Nonnegative Matrix Factorization)는 비음수 행렬 분해에서처럼 식 (1) 형태의 분해를 찾으나, 행렬 $U$혹은 행렬 $V$가 직교 제한 ($U^TU$=1 혹은 $V^TV$=1)를 만족하도록 하는 것이다 [8]. $V^TV$=1의 조건을 만족하는 경우, 비음수 행렬의 직교 분해의 최적화 대상 함수가 일반적인 $k$-평균 집단화 (k-means clustering)와 같은 형태를 갖게 되므로[9], 분해된 행렬로부터 재구성 했을 때의 오차를 최소화하는 것을 목적으로 하는 비음수 행렬 분해보다 집단화에 적합한 결과를 얻을 수 있게 된다. 따라서, 여기서는 행렬 $V$가 직교 제한을 만족시키는 경우를 다룬다. 이 장에서, 비음수 행렬의 직교 분해를 위한 새로운 알고리즘을 Stiefel 다양체 위에서의 최대 감소 방향을 이용한 곱셈의 업데이트 형태로 구하는 방법을 소개한다.

직교 조건 $V^TV$=1을 만족하는 비음수 행렬의 직교 분해는 다음과 같은 최적화 문제로 표현할 수 있다.

$$\text{argmin}_{U,V}\frac{1}{2}\|X-UV^T\|^2$$
$$subject\ to\ V^TV=I, U\geq 0, v\geq 0 \tag{13}$$

일반적으로, 위의 (13)과 같이 조건이 주어진 최적화 문제는 라그랑지안 승수를 사용하여 조건에서 벗어나는 경우에 제약을 가하는 방식으로 풀게 된다. 비음수 행렬의 직교 분해에서는 라그랑지안 승수로 구성된 행렬 $\Lambda$ $\Lambda$를 사용한 항 $\text{tr}\{\Lambda(V^TV-I)\}$을 더하여 구할 수 있다. [3] 에서 이러한 접근을 기초로 적절한 근사를 사용하여 곱셈의 업데이트 방법을 제안한 바 있다.

여기서는 위에 소개된 방법과는 달리, $V^TV=I$를 만족시키는 제한된 공간 위에서의 최대 감소 방향을 계산하는 방식으로 비음수 행렬의 직교 분해를 위한 새로운 방법을 유도하였다. 행렬 $U$가 고정되어 있다고 가정하면, (13)의 식은 $V$에 대한 식으로 볼 수 있다. 식 (13)을 $V$가 $V^TV=I$를 만족하는 $n\times K$ 행렬이라는 조건을 만족하는 공간 위에서 최적화 하는 방법은 다양하게 연구되어 왔다[10,11]. 하지만 기존의 방법들은 비음수 조건이 있는 경우에 적합하지 않으므로, 여기서는 $V$에 대한 곱셈의 업데이트를 사용하여 비음수 조건을 만족시킴과 동시에 직교 제한 $V^TV=I$을 만족시키는 방법을 소개한다. $n\times K$ 차원을 가지며 $V^TV=I$조건을 만족시키는 직교 행렬들로 이루어진 공간은 Stiefel 다양체라는 이름으로 불린다[12].

Stiefel 다양체 위에서의 한 점 $V$로부터의 접선 방향을 정의하는 식은 조건식 $V^TV=I$를 미분함으로써 다음과 같이 얻을 수 있다.

$$V^T\Delta+\Delta^TV=0, \tag{14}$$

즉, $V^T\Delta$가 반대칭 행렬이 됨을 알 수 있다. Stiefel 다양체에서의 표준 계량[11]은 다음과 같다.

$$g_c(\Delta,\Delta)=\text{tr}\,\Delta^T(I-\frac{1}{2}VV)\Delta \tag{15}$$

참고로 유클리디안 공간에서의 계량은 다음과 같이 표현된다.

$$g_e(\Delta,\Delta)tr\Delta^T\Delta \tag{16}$$

$V$에 대한 어떤 함수 $\varepsilon$이 주어져 있을 때, $V$의 각각의 원소에 대한 편도함수를 다음과 같이 정의할 수 있다.

$$[\nabla_V\varepsilon]_{ij}=\frac{\partial\varepsilon}{\partial V_{ij}} \tag{17}$$

$U$가 고정되어 있다고 가정하면, 함수 $\varepsilon$(13)는 Stiefel 다양체 위에서 정의되어 있다고 볼 수 있으며, 이 때 함수 $\varepsilon$의 최대 감소 방향은 $V$의 모든 접벡터 $\Delta$에 대하여 다음 식을 만족하는 접벡터 $\widetilde{\nabla}_V\varepsilon$ 가 된다.

$$\begin{aligned} g_e(\nabla_V\varepsilon,\Delta) &= tr\{(\nabla_V\varepsilon)^T\Delta\} \\ &= g_c((\widetilde{\nabla}_V\varepsilon,\Delta)) \\ &= tr\left\{(\widetilde{\nabla}_V\varepsilon)^T(I-\frac{1}{2}VV^T)\Delta\right\} \end{aligned} \tag{18}$$

위의 식을 $V^T\widetilde{\nabla}_V\varepsilon$가 반대칭이라는 조건을 이용하여 풀면 $\widetilde{\nabla}_V\varepsilon$ 를 다음과 같이 구할 수 있다.

$$\widetilde{\nabla}_V\varepsilon=\nabla_V\varepsilon-V(\nabla_V\varepsilon)^TV. \tag{19}$$

따라서, (7)에서 계산한 편도함수를 이용하여 Stiefel 다양체 위에서의 최대 감소 방향을 다음과 같이 계산할 수 있다.

$$\begin{aligned} \widetilde{\nabla}_V\varepsilon &= (-X^TU+VU^TU)-V(-X^TU+VU^TU)^TV \\ &= VU^TXV-X^TU \\ &= [\widetilde{\nabla}_V\varepsilon]^+-[\widetilde{\nabla}_V\varepsilon]^- \end{aligned} \tag{20}$$

(5)의 관계식에 위의 식을 대입하면 다음과 같은 곱셈의 업데이트 식을 유도할 수 있다.

$$V\leftarrow V\odot\frac{X^TU}{VU^TXV}, \tag{21}$$

위의 식은 직교 조건이 없는 비음수 행렬 분해의 업데이트식 (22)에서 분모의 $U$가 직교 조건 $V^TV=I$을 만족할 경우의 관계식 $U=XV$에 의하여 치환된 것으로 볼 수 있다. 분자의 $X^T$과 분모의 $VU^T$에 의하여 행렬 분해의 오차(2)를 줄이게 되는데, 이는 재구성된 값 $VU^T$가 데이터 $X^T$보다 작을 경우 $V$가 커지는 방향으로 업데이트가 됨으로써 오차를 줄이기 때문이다. 이와 마찬가지로, 분자의 $U$와 분모의 $XV$에 의하여 직교 조건을 만족시키는 방향으로 업데이트가 진행될 것임을 알 수 있다.

(24)식이 여기에서 제안하고자 하는 비음수 행렬의 직교 분해 알고리즘의 핵심적인 부분이며, 행렬 $U$의 업데이트는 별다른 제한이 없으므로 비음수 행렬 분해에서와 동일한 방법으로 수행이 가능하다.

## 4. 실험 결과

제안된 비음수 행렬의 직교 분해 방법을 사용하여 여섯 개의 일반적인 문서 데이터 모음(CSTR, k1a, k1b, re0, re1)에 대한 집단화를 시도하고, 성능을 비음수 행렬 분해 방법과, 이전에 제안된 비음수 행렬의 직교 분해 방법(DTPP)[3]과 비교하였다. 모든 문서 데이터에 대하여 어근 추출(stemming)과 일반적인 단어 제거(stop word removal)를 수행하였으며, 집단과의 상호 정보(mutual information)를 바탕으로 1,000 개의 단어를 선택하여 사용하였다. 또한, 데이터 행렬에 정규화-절단 가중치(normalized-cut weighting)[4]를 부여하였다.

집단화의 성능을 측정하고 비교하기 위해서 집단화 정확도를 기준으로 사용하였다. 집단화 정확도를 계산하기 위하여, Kuhn-Munkres 최대 정합 방법[13]을 사용하여 집단화 결과와 실제 집단과의 관계를 찾아내었다. $n$번째 문서의 실제 집단이 $c_n$이고, 집단화 알고리즘의 결과로 인해 찾아진 집단이 $\widetilde{c_n}$이라고 할 때, 집단화 정확도 AC는 다음과 같이 계산할 수 있다.

$$AC=\frac{\sum_{n=1}^{N}\delta(c_n,\widetilde{c_n})}{N}, \tag{22}$$

위의 식에서 $x=y$인 경우에 $\delta(x, y)=1$이 되고, $x \neq y$인 경우 $\delta(x, y)=0$이 되는 델타 함수가 사용되었다. 실험에 사용된 비음수 행렬 분해 및 비음수 행렬의 직교 분해 알고리즘들은 모두 초기 조건에 의해서 조금씩 다른 결과를 내는 알고리즘들이므로, 정확한 측정을 위해 100번의 실험을 서로 다른 초기 조건을 갖고 수행한 후 결과로 얻은 집단화 정확도의 평균을 구하였다. 초기 조건은 매 측정마다 행렬의 각 값을 0 에서 1 사이에서 균등하게 임의로 추출(uniform random sampling)된 양수값으로 채우는 방식으로 결정되었다. 실험 결과 제안된 비음수 행렬의 직교 분해 방법이 비음수 행렬 분해 방법이나 이전에 제안된 비음수 행렬의 직교 분해 방법(DTPP)[3] 보다 나은 성능을 보였다(표 1).

행렬 $V$의 직교 정도의 유지 여부 정도도 $\|V^TV-I\|$를 통해 계산되었다. 직교 정도가 알고리즘이 수행됨에 따라 감소하는 정도를 100회 측정하여 평균을 계산하였다. 실험 결과, 제안된 비음수 행렬의 직교 분해 방법이 비음수 행렬 분해나 이전에 제안된 비음수 행렬의 직교 분해보다 직교 제한을 더 잘 만족시키는 것으로 나타났다. 한 예로 CSTR 문서 데이터 모음에 대한 직교 정도의 변화 양상을 그림 1에 표시하였다.

표 1 여섯 개의 문서 데이터 모음에 대한 비음수 행렬 분해(NMF), Ding의 비음수 행렬의 직교 분해 방법(DTPP) 및 제안된 비음수 행렬의 직교 분해 방법(ONMF)의 평균 집단화 정확도(100번 수행의 평균치)

| | NMF | DTPP | ONMF |
|---|---|---|---|
| cstr | 0.7568 | **0.7844** | 0.7268 |
| wap | 0.4744 | 0.4281 | **0.4917** |
| k1a | 0.4773 | 0.4311 | **0.4907** |
| k1b | 0.7896 | 0.6087 | **0.8109** |
| re0 | 0.3624 | 0.3384 | **0.3691** |
| re1 | 0.4822 | 0.4452 | **0.5090** |

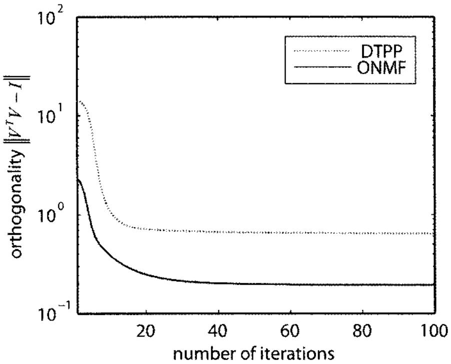


그림 1 CSTR 데이터에 대해 측정된 Ding의 비음수 행렬의 직교 분해 방법(DTPP)과 제안된 비음수 행렬의 직교 분해 방법(ONMF)의 직교 정도의 변화 양상

## 5. 결 론

여기서는 Stiefel 다양체 위에서의 곱셈의 업데이트를 사용한 비음수 행렬의 직교 분해 알고리즘을 제안하였다. 제안된 알고리즘은 문서의 집단화 문제에 성공적으로 적용되어, 실험적으로 기존의 비음수 행렬 분해와 이전에 제안된 비음수 행렬의 직교 분해보다 우수한 성능을 보임이 입증되었다.

## 참 고 문 헌

[1] Lee, D.D., Seung, H.S., "Learning the parts of objects by non-negative matrix factorization," Nature 401, pp. 788-791, 1999.
[2] Li, S.Z., Hou, X.W., Zhang, H.J., Cheng, Q.S. "Learning spatially localized parts-based repre-

sentation," Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, Hawaii, pp. 207-212, 2001.

[3] Ding, C., Li, T., Peng, W., Park, H., "Orthogonal nonnegative matrix tri-factorizations for clustering," Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Philadelphia, PA, 2006.

[4] Xu, W., Liu, X., Gong, Y., "Document clustering based on non-negative matrix factorization," Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR), Toronto, Canada, 2003.

[5] Shahnaz, F., Berry, M., Pauca, P., Plemmons, R., "Document clustering using nonnegative matrix factorization," Information Processing and Management 42, pp. 363-385, 2006.

[6] Lee, D.D., Seung, H.S., "Algorithms for non-negative matrix factorization," Advances in Neural Information Processing Systems (NIPS). Volume 13, MIT Press, 2001.

[7] Gaussier, E., Goutte, C., "Relation between PLSA and NMF and implications," Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR), Salvador, Brazil, 2005.

[8] Choi, S., "Algorithms for orthogonal nonnegative matrix factorization," Proceedings of the International Joint Conference on Neural Networks (IJCNN), Hong Kong, 2008.

[9] Ding, C., He, X., Simon, H.D., "On the equivalence of nonnegative matrix factorization and spectral clustering," Proceedings of the SIAM International Conference on Data Mining (SDM), Newport Beach, CA, pp. 606-610, 2005.

[10] Smith, S.T., "Geometric Optimization Methods for Adaptive Filtering," PhD thesis, Harvard University, May 1993.

[11] Edelman, A., Arias, T., Smith, S.T., "The geometry of algorithms with orthogonality constraints," SIAM J. Matrix Anal. Appl. 20(2), pp. 303-353, 1998.

[12] Stiefel, E., "Richtungsfelder und fernparallelismus in n-dimensionalem mannig faltigkeiten," Commentarii Math. Helvetici 8, pp. 305-353, 1935-1936.

[13] Lovasz, L., Plummer, M., "Matching Theory," Akademiai Kiado, 1986.

유 지 호

정보과학회논문지 : 소프트웨어 및 응용 제 36 권 제 3 호 참조

최 승 진

정보과학회논문지 : 소프트웨어 및 응용 제 36 권 제 3 호 참조