1I2-OS-10a-1

制約付き独立潜在情報分析 (CISA) における制約の選択方法

Condition for the selected topics in Constrained Independent Semantic Analysis (CISA)

西垣 貴央*1　小野田 崇*2
Takahiro Nishigaki　Takashi Onoda

*1東京工業大学
Tokyo Institute of Technology

*2電力中央研究所
Central Research Institute of Electric Power Industry

Recently, we have proposed the Constrained Independent Semantic Analysis (CISA). This method re-generated the topics filled with user constraints after generating the topics by using Independent Semantic Analysis (ISA). Generally ISA and CISA are applied to unknown data, so it is difficult that the user selects the adequate constraints. In this paper, we proposed the selection method of adequate user constraints for generating the topics with higher independence. And the proposed method can decide on a lower bound of the number of topics. In our experiments, we applied the CISA with the proposed selection method to Web corpus. The experiments show that the topics generated by using CISA with the proposed selection method have higher independence than the topics generated by using ISA. And the results of the experiments show that the proposed selection method is useful for generating adequately independent topics.

1. はじめに

近年，Web ページや電子ニュース，またブログやソーシャルネットワークサービスなどインターネット利用の一般化に伴い，個人のハードディスクドライブ (HDD) や Web 上には大量の文書データが蓄積されている．整理されずに蓄積されている大量の文書データの中から，必要な文書データを探し出すことは非常に困難である．膨大な量の文書データから必要な文書データを見つける手段の一つとして，多くのニュースサイトでは社会，科学，政治などのトピックに基づいた文書データの整理を行っている．トピックに基づいて文書整理や文書検索，内容要約など文書データに対する処理の多くを容易にするために，文書データを簡潔に表す潜在的なトピックを見つけることが重要である．

文書データが持つ潜在的なトピックを抽出する方法として，bag-of-words 表現された文書の生成過程を確率的にモデル化したトピックモデルと呼ばれる PLSI や LDA などが存在する．トピックモデルでは，ある文書に含まれる各単語は，文書固有のトピック比率に従ってあるトピックを選択し，その選択されたトピックがもつ固有の単語出現確率分布に従って単語は生成される，と仮定して行われる．一方で，トピックモデルのように複雑なパラメータを使用せずに，文書データの単語の類義性や多義性に着目し，文書データにはトピックが正規分布して存在しているという仮定を用いてトピックを抽出する LSA [1] という方法が存在する．無限の量の文書データからトピックを抽出する場合，文書データに潜在するトピックは正規分布しているという仮定は有効であると考えられる．しかし，有限の量の文書データに潜在するトピックは必ずしも正規分布しているとは限らず，HDD や Web 上に存在する電子ニュースやブログのように限られたトピックが潜在する文書データにおいては，多くのトピックが正規分布ではない分布をしていると考えられる．

ここでは，文書データには正規分布ではない独立なトピックが潜在するという仮定を用いた，独立潜在情報分析 (ISA: Independent Semantic Analysis)[2, 3] に注目する．この ISA では，電子ニュースやブログのように限られたトピックが潜在する文書データの中から，任意の数の正規分布していない独立なトピック（潜在情報）を抽出する．さらに，この ISA で抽出した潜在情報にユーザが適切な制約を与えることで，より独立性が高い潜在情報を得ることができる，制約付き独立潜在情報分析 (CISA: Constrained ISA) [4] という方法が提案されている．しかしながら，CISA によってより独立性が高い潜在情報を得るためには，適切な制約をユーザが選択しなくてはならない．そこで本稿では，CISA によってより独立性が高い潜在情報を得るために必要である適切な制約の選択方法を提案する．また，それに伴い ISA や CISA で課題であった，適切な潜在情報の数についても議論を行う．

以下，2 章では関連研究として ISA および CISA について簡単に紹介し，3 章でユーザが選択する適切な潜在情報の選択方法について，および適切な潜在情報の数についてを述べる．4 章では提案した方法を用いて CISA をベンチマークデータに適用した結果について示し，最後に 5 章でまとめと今後の展望について述べる．

連絡先: 連絡先: 西垣貴央, 東京工業大学大学院 総合理工学研究科 知能システム科学専攻, 〒 226-8502 神奈川県横浜市緑区長津田町 4259 J2-53, nishigaki@ntt.dis.titech.ac.jp

2. 関連研究

本章では，文書データには正規分布ではない独立なトピックが潜在するという仮定を用いた独立潜在情報分析 ISA (Independent Semantic Analysis) と，その手法にユーザ制約を加えた制約付き独立潜在情報分析 CISA (Constrained ISA) について簡単に述べる．以下，記号の小文字はスカラー，小文字太字はベクトル，大文字太字は行列を表す．

2.1 ISA: Independent Semantic Analysis

文書データ $\mathbf{x}_{1,\ldots,N}$ は，独立な潜在情報 $\mathbf{s}_{1,\ldots,K}$ と，“文書での潜在情報の強度” を表す混合行列 $\mathbf{A}(\mathbf{x},\mathbf{s})$ を用いて，独立な潜在情報の線形和として次のように表すことができる．

$$\mathbf{x}_i = a(\mathbf{x}_i,\mathbf{s}_1)\cdot\mathbf{S}_1 + a(\mathbf{x}_i,\mathbf{s}_2)\cdot\mathbf{S}_2 + \cdots + a(\mathbf{x}_i,\mathbf{s}_k)\cdot\mathbf{S}_k$$

ここで，$a(\mathbf{x}_i,\mathbf{s}_1)$ は文書データ $\mathbf{x}_i$ における独立な潜在情報 $\mathbf{s}_1$ の強度を示す値である．

また，文書データ $\mathbf{x}_{1,\ldots,N}$ と独立な潜在情報 $\mathbf{s}_{1,\ldots,K}$ は単語 $\mathbf{c}_{1,\ldots,M}$ の値によって表現される．文書データを各単語 $\mathbf{c}$ が文書データ $\mathbf{x}$ の中での強さを “文書データでの単語の強度” と呼ぶ行列 $\mathbf{R}(\mathbf{x},\mathbf{c})$ による表現ができる．同様に独立な潜在情報を各単語 $\mathbf{c}$ が独立な潜在情報 $\mathbf{s}$ を特定する力を “潜在情報での単語の重要度” と呼ぶ行列 $\mathbf{V}(\mathbf{s},\mathbf{c})$ によって表す．さらに，各文書データ $\mathbf{x}$ が独立な潜在情報 $\mathbf{s}$ を特定する力を “潜在情報での文書データの重要度” と呼ぶ行列 $\mathbf{U}(\mathbf{s},\mathbf{x})$ によって表す．

ISA では文書データ $\mathbf{x}$ から独立な潜在情報 $\mathbf{s}$ を推定し，“文書での潜在情報の強度”$\mathbf{A}(\mathbf{x},\mathbf{s})$ に基づいて，各文書データがどの独立な潜在情報からどの程度影響を受けているのかを分析する．ISA のアルゴリズムを以下に示す．このときユーザは，真の潜在情報の数 k はわからないものとする．

1. 文書データ集合 $\mathbf{X}$ を，文書データを行に，単語を列にとった行列 $\mathbf{R}(\mathbf{x},\mathbf{c})$ として整理する．
2. $\mathbf{R}(\mathbf{x},\mathbf{c})$ を正規化し，$\hat{\mathbf{R}}(\mathbf{x},\mathbf{c})$ を求める．
3. ステップ 2. で求めた $\hat{\mathbf{R}}(\mathbf{x},\mathbf{c})$ を次のように分解する．$\mathbf{U}^{\mathrm{T}}\cdot\hat{\mathbf{R}}\cdot\mathbf{V}=\mathbf{D}\Longleftrightarrow\hat{\mathbf{R}}=\mathbf{U}\cdot\mathbf{D}\cdot\mathbf{V}^{\mathrm{T}}$．$\mathbf{U}$ と $\mathbf{V}$ は独立な潜在情報での文書データと単語の重要度を示す行列である．また $\mathbf{D}$ は特異値の対角行列であり，その大きさの順に k 個の成分を抜き出し，$\mathbf{U}_k$，$\mathbf{D}_k$，$\mathbf{V}_k$ を作成する．
4. ステップ 3. で得られた $\mathbf{U}_k$，$\mathbf{D}_k$ を用いて，各潜在情報間の独立性が最大となるときの，“文書データにおける潜在情報の強度”$\mathbf{A}(\mathbf{x},\mathbf{s})$ を，FPICA [5] に基づいたアルゴリズムによって求める．
5. ステップ 4. で求めた “文書データにおける潜在情報の強度”$\mathbf{A}(\mathbf{x},\mathbf{s})$ の値によって，各文書データがどの潜在情報から派生しているのかを決定する．

$$\mathcal{C}_j=\{\mathbf{x}_i\mid\arg\max_{\mathbf{s}}a(\mathbf{x}_i,\mathbf{s}_j)\},$$
$$i\in\{1,\cdots,N\},\ j\in\{1,\cdots,K\}$$

この手法の課題として，潜在情報数が未知のためユーザが適当に指定した潜在情報数の結果が，ユーザの望む結果と異なる場合がある．そのような場合において，ISA で求めた潜在情報に潜在情報の数を減らすようなユーザ制約を加えることで，少ない潜在情報数でより独立性が高い潜在情報を求める CISA が提案されている．

2.2 CISA: Constrained ISA

CISA では，ISA で求めた k 個の潜在情報のうち 2 個にユーザ制約を加えて，$k-1$ 個のより独立な潜在情報を求めることができる．その際，次の仮定のもとで行われる．

I. ISA で求めた k 個の潜在情報は，ISA で求めた $k-1$ 個の潜在情報を含んでいる．

$$ISA(k-1)\in ISA(k) \tag{1}$$

ここで $ISA(k)$ は潜在情報数 k で ISA を行って得た潜在情報を示す．

II. $ISA(k)$ のある 1 個の潜在情報は $ISA(k-1)$ の潜在情報のうちの 1 個から分裂して生成される．

この仮定のもとで，k 個の潜在情報から $k-1$ 個の潜在情報を求める CISA についてのアルゴリズムを次に示す．

a. 制約を加える潜在情報 $\mathbf{s}_i$ と $\mathbf{s}_\ell$ を選択する．
   - 制約を加えた新しい潜在情報 $\mathbf{ns}$ の初期値を $\mathbf{ns}=(\mathbf{s}_i+\mathbf{s}_\ell)/2$ とする．

b. ユーザが選択した潜在情報以外の $k-2$ 個の潜在情報を初期値として，ステップ 1. で求めた $\mathbf{ns}$ とのコサイン相関が低いものから順番に各潜在情報の独立性が最大となるように更新する．更新方法は FPICA [5] を用いて行う．

c. $k-2$ 個の新たな潜在情報が得られたら，最後にユーザがステップ 1. で求めた $\mathbf{ns}$ の独立性が最大となるように更新する．更新方法は FPICA [5] を用いて行う．

これらのステップ a. b. c. を複数回繰り返すことで，潜在情報を 1 個ずつ減らしていくことが可能である．

CISA の課題として，ステップ a. での 1) 独立性がより高い潜在情報の推定を行うことができる，適切な 2 個の潜在情報をユーザが選択するのは難しい点と，CISA のアルゴリズムを複数回繰り返すことで 2) 潜在情報の数は最低 2 個まで減らせるが，適切である考えられる潜在情報数の数が分からない点，ということがある．

3. 提案方法

本章では，CISA における課題である，1) 制約として選択する潜在情報の選び方ついて，および 2) 適切である考えられる潜在情報の数についての課題を解決する方法について述べる．

3.1 選択する潜在情報の選び方について

CISA は，未知のデータに適用する場合を想定しているため，独立性がより高い潜在情報の推定のために，適切な潜在情報をユーザが選択するのは非常に困難である．そこでここでは，より独立性が高い潜在情報の推定のために，ユーザが選択する適切な潜在情報の選択方法について提案する．

CISA を適用する際，前章で述べた 2 つの仮定を用いている．この仮定のもとでユーザは 2 個の潜在情報を選択する．ユーザが選択する適切な潜在情報の条件は，次の 2 つである．

- $ISA(k-1)$ のときには存在しなかった新たな潜在情報
- その新たな潜在情報が生成される際に，分裂したと考えられる潜在情報

以上を満たす潜在情報を選択するために，我々は $ISA(k-1)$ と $ISA(k)$ とのコサイン相関を計算する．式 (1) という仮定があるので，$ISA(k-1)$ と $ISA(k)$ とのコサイン相関を求めると，$ISA(k)$ においてある 2 個の潜在情報は，$ISA(k-1)$ においてある 1 個の潜在情報とのコサイン相関の値が大きいものとなる．これは，$ISA(k-1)$ でのその潜在情報が，$ISA(k)$ を求める際に 2 個に分裂したと考えられるためである．この 2 個の潜在情報をユーザが選択するのに適切な潜在情報として提示する．

イメージ図を図 1 に示す．小さい丸が各データ点を示しており，直線が $ISA(3)$ によって得られた 3 個の潜在情報で，破線が $ISA(2)$ によって得られた潜在情報である．直線と破線のコサイン相関を調べると $ISA(3-1)$ と $ISA(2-1)$，$ISA(3-2)$ と $ISA(2-2)$ のコサイン相関が高く，$ISA(3-3)$ は $ISA(2-2)$ の方がコサイン相関は高くなるが，他のものと比べると低い．この場合，ユーザが選択する適切な潜在情報は $ISA(3-2)$ と $ISA(3-3)$ となる．

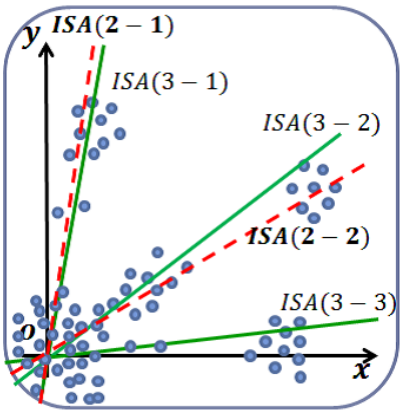

図 1: $ISA(3)$ で選択する適切な潜在情報のイメージ図

表 1 の上段にベンチマークデータである CLUTO[6] の LA Times [7] での，潜在情報数 7 の時と 6 とのコサイン相関を示す．$ISA(7)$ でのコサイン相関の高いものを太字としている．このとき，CISA でユーザが選択する適切な潜在情報は $\mathbf{s}_{7-5}$ と $\mathbf{s}_{7-7}$ である．この 2 個の潜在情報は，$ISA(6)$ の $\mathbf{s}_{6-5}$ が 2 個に分裂したと考えられる．このように，コサイン相関を用いて CISA でユーザが選択する適切な潜在情報を決定する．

3.2 適切と考えられる潜在情報の数について

CISA を用いることによって，潜在情報の数は最低 2 個にまで減らすことが可能である．しかし，新聞などの大規模な文書データに 2 個の潜在情報しか存在しないということは考えにくい．この節では，適切と考えられる潜在情報の数を提示する方法について述べる．

CISA を適用する際，ISA で求めた k 個の潜在情報は，ISA で求めた $k-1$ 個の潜在情報を含んでおり，残りの 1 つは $ISA(k-1)$ のうちの 1 つから分裂して生成されるという仮定で行われる．しかし，潜在情報の数が減ると分裂したと考えられる潜在情報以外の潜在情報のコサイン相関の値が低くなる場合がある．その場合，ISA で求めた k 個の潜在情報は，ISA で求めた $k-1$ 個の潜在情報を含んでいるという仮定が成り立たなくなる．その仮定が成り立たなくなる数を適切と考えられる潜在情報の数としてユーザに提示する．例えばイメージ図 1 で，直線 $ISA(3-1)$ と点線 $ISA(2-1)$ とのコサイン相関の値がある値よりも低くなる場合，$ISA(3)$ の潜在情報は $ISA(2)$ の潜在情報を含んでいるという仮定が成り立たなくなる．その場合，適切であると考えられる潜在情報数は 3 としてユーザに提示する．仮定が成り立たないとするコサイン相関の値は，45 度以上となるときのコサイン相関の値とする．

表 1 の下段に，$ISA(7)$ の潜在情報と $ISA(6)$ の潜在情報との角度が最も小さい時の値を示す．これを見ると $ISA(\mathbf{s}_{7-5})$ と $ISA(\mathbf{s}_{7-7})$ 以外の制約に関係のない潜在情報 $\mathbf{s}_{7-6}$ と最も近い潜在情報との角度が 60.4093 度となっており，45 度を超えていることが分かる．これはつまり，LA Times において $ISA(7)$ と $ISA(6)$ の間には，$ISA(7)$ は $ISA(6)$ を含んでいるという仮定が成り立っていないことを示している．以上より，LA Times の潜在情報数は 7 が適切であると考えられる．

4. 実験

この章では，前章で紹介した方法を用いてユーザ制約を選択し CISA を適用した結果を紹介する．比較方法としては，推定した潜在情報間の相互情報量 [10] を使用する．相互情報量の値が 0 の場合，その潜在情報間は完全に独立している．推定した潜在情報の相互情報量は次の式を用いて計算することができる．

$$MI(\mathbf{S}) = \sum_{i=1}^{k} H(\mathbf{s}_i) - H(\mathbf{S})$$

$H(\mathbf{S})$ は潜在情報の情報量で次のように求められ，$P_{\mathbf{s}}(i)$ は i 番目の潜在情報の確率を示す．

$$H(\mathbf{s}_i) = -\sum_{i=1}^{k} P_{\mathbf{s}}(i) \log \frac{1}{P_{\mathbf{s}}(i)},\ P_{\mathbf{s}}(i) = \frac{\sum_{l=1}^{n} 1(\mathbf{s}(\mathbf{x}_l) - i)}{n}$$

表 2 に LA Times，KOS blog と NIPS [8]，20newsgroups [9] のそれぞれの潜在情報数の時の相互情報量の値と，その時の，$ISA(k)$ と $ISA(k-1)$ の制約に関係のない潜在情報間の角度で最も大きい値を示す．これは，LA Times では表 1 の下段と同じで，制約に関係のない $ISA(\mathbf{s}_{7-5})$ と $ISA(\mathbf{s}_{7-7})$ 以外の潜在情報の中で最も大きい値を示している．他のベンチマークデータでの結果も同様である．また，それぞれのデータは，LA Times は文書数 6279，単語数 31472，同様に KOS blog は文書数 3430，単語数 6906，NIPS は文書数 1500，単語数 12419 で，20newsgroups は文書数 11267，単語数 37577 の文書データである．

表 2 の全てのデータセットの結果を見ると，ISA によって求めた $ISA(k-1)$ の潜在情報の相互情報量よりも，CISA によって求めた $CISA(k \to k-1)$ の潜在情報の相互情報量の方が小さい値となっている事がわかる．これは ISA によって求めた潜在情報よりも CISA によって得られた潜在情報のほうが独立性が高いことを示している．また，表 2 での KOS blog の結果では $CISA(11 \to 10)$ のときに，制約に関係の無い潜在情報の角度が 46.4279 度となっており 45 度を超えている．これは潜在情報数 11 が適切だと考えられる潜在情報数である．同様に，LA Times では適切だと考えられる潜在情報の数は 6 で，20newsgroups での適切な潜在情報の数は 9 である．

一方で NIPS の結果では潜在情報の数を 2 個にまで減らしても，角度が 45 度を超えることは無かった．これは潜在情報の数は 2 個が適切であることを示している．NIPS は Neural Information Processing Systems の論文の文書データであり，少ない潜在情報で構成されていると考えられる．得られた潜在情報での単語の重要度の大きいものは，潜在情報が 2 個のとき，“network” “unit” “input” “neural” の潜在情報と，“learning” “function” “algorithm” “model” の潜在情報となっている．また，KOS blog は政治的なブログデータの集まりで推定した潜在情報を分析してみると，予算に関する単語が集まった潜在情報や，外交に関する潜在情報，国内での政策についての潜在情報，投票に関する潜在情報，人種に関する潜在情報などに分かれており，NIPS と比較すると様々な単語で構成されていることがわかった．

以上の結果より，CISA でユーザが経験的に選択する潜在情報を提案手法により自動的に決めた CISA の結果は，ユーザが経験的に選択する潜在情報を選択した場合と同様に，ISA の結果よりも独立性が高いことが示せた．また適切であると考えられる潜在情報の数も示すことができた．

5. おわりに

本論文では，CISA においての課題であった，1) ユーザが選択する適切な潜在情報の選択方法についておよび，2) 適切であ

表 1: LA Times で $ISA(7)$ と $ISA(6)$ とのコサイン相関

$ISA(6)\backslash ISA(7)$	$\mathbf{s}_{7-1}$	$\mathbf{s}_{7-2}$	$\mathbf{s}_{7-3}$	$\mathbf{s}_{7-4}$	$\mathbf{s}_{7-5}$	$\mathbf{s}_{7-6}$	$\mathbf{s}_{7-7}$
$\mathbf{s}_{6-1}$	**0.9858**	0.0157	0.0138	-0.0220	0.1161	0.0217	0.1094
$\mathbf{s}_{6-2}$	0.0025	**0.9952**	-0.01439	0.0263	-0.0982	0.0060	0.0001
$\mathbf{s}_{6-3}$	-0.0026	-0.0025	**0.9953**	0.0145	-0.0846	-0.0163	-0.0258
$\mathbf{s}_{6-4}$	-0.0041	-0.0239	-0.0123	**0.7349**	0.0087	0.3793	0.2628
$\mathbf{s}_{6-5}$	0.0027	0.0523	0.03871	0.1258	**0.6382**	0.1708	**-0.7331**
$\mathbf{s}_{6-6}$	-0.0075	-0.0115	0.0061	-0.6448	-0.0433	**0.4938**	0.0042
最も小さい角度	9.6520	5.6185	5.5485	42.7077	50.3386	**60.4093**	42.8542

表 2: ベンチマークデータでの ISA と CISA 潜在情報の相互情報量

LA Times	$ISA(8)$	$CISA(9 \to 8)$	$ISA(7)$	$CISA(8 \to 7)$	$ISA(6)$	$CISA(7 \to 6)$
相互情報量	1.8040	**1.7545**	1.4076	**1.3044**	1.1270	**1.0331**
最大の角度	\	16.3649	\	23.8776	\	**60.4093**
KOS blog	$ISA(12)$	$CISA(13 \to 12)$	$ISA(11)$	$CISA(12 \to 11)$	$ISA(10)$	$CISA(11 \to 10)$
相互情報量	3.9449	**3.6023**	3.2479	**2.9372**	2.4470	**2.4421**
最大の角度	\	27.2652	\	37.1302	\	**46.4179**
NIPS	$ISA(4)$	$CISA(5 \to 4)$	$ISA(3)$	$CISA(4 \to 3)$	$ISA(2)$	$CISA(3 \to 2)$
相互情報量	0.3665	**0.3604**	0.2369	**0.1826**	0.1066	**0.0737**
最大の角度	\	27.9038	\	27.6578	\	15.7894
20newsgroups	$ISA(10)$	$CISA(11 \to 10)$	$ISA(9)$	$CISA(10 \to 9)$	$ISA(8)$	$CISA(9 \to 8)$
相互情報量	2.5213	**2.2828**	2.0995	**1.9479**	1.5993	**1.5804**
最大の角度	\	22.5665	\	14.0346	\	**67.3468**

ると考えられる潜在情報の数についての 2 つを解決する方法について述べた．CISA を用いる際には，$ISA(k-1)$ と $ISA(k)$ の間には包含関係があるという仮定を用いていることを利用して，1) 2) の課題を克服する方法を提案した．提案した方法を用いて，LA Times，KOS blog，NIPS，20newsgroups に ISA および CISA を適用した．その結果，$ISA(k-1)$ で得た潜在情報の相互情報量よりも $CISA(k \to k-1)$ で得られた潜在情報の相互情報量のほうが小さく，独立性が高いことが示せた．また，包含関係にあるという仮定が成り立たなくなる時の潜在情報の数を適切である潜在情報の数としてユーザに提示することで，潜在情報の数の下限を見つけることができた．

今後の課題には，適切な潜在情報数として提示した数の妥当性について調べる必要がある．また，潜在情報の数を減らすだけではなく，増やす方法についても検討したい．

参考文献

[1] Scott Deerwester et al., “Indexing by latent semantic analysis”, Journal of the American Society for Information Science, Vol. 41, No. 6, pp.391-407, 1990.

[2] Takahiro Nishigaki and Takashi Onoda, “Indepnedence based Clusteirng”, 2012 Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), pp. 389-390, 2012.

[3] Takahiro Nishigaki and Takashi Onoda, “Clustering based on independent component”, 2012 International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 74-78, 2012.

[4] Takahiro Nishigaki and Takashi Onoda, “Constrained Clustering Based on Semantic Information”, Conference on Technologies and Applications of Artificial Intelligence (TAAI), pp. 268-273, 2013.

[5] A. Hyvarinen, E.Oja, “A Fast Fixed-Point Algorithm for Independent Component Analysis”, Neural Computation, Vol.9, No.7, pp. 1483-1492, 1997.

[6] George Karypis, “CLUTO - A Clustering Toolkit”, http://glaros.dtc.umn.edu/gkhome/views/cluto/ , Department of Computer Science and Engineering, University of Minnesota, 2002.

[7] S. Zhong, J. Ghosh, “A comparative study of generative models for document clustering”, Data Mining Workshop on Clustering High Dimensional Data and Its Applications, 2003.

[8] M. Lichman, “UCI Machine Learning Repository”, http://archive.ics.uci.edu/ml , 2013.

[9] Jason Rennie, “The 20 Newsgroups data set”, http://qwone.com/ jason/20Newsgroups/ , 2007.

[10] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján ”Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature ”, Journal of Machine Learning Research (JMLR). Vol 13, pp.27-66, 2012.