Pure High-order Word Dependence Mining via Information Geometry

Conference or Workshop Item

Pure High-Order Word Dependence Mining via Information Geometry

Yuexian Hou[1], Liang He[1], Xiaozhao Zhao[1], and Dawei Song[2]

[1] School of Computer Sci & Tec, Tianjin University, Tianjin, China
{krete1941,roba269,0.25eye}@gmail.com
[2] School of Computing, The Robert Gordon University, Aberdeen, United Kingdom
d.song@rgu.ac.uk

Abstract. The classical bag-of-word models fail to capture contextual associations between words. We propose to investigate the "high-order pure dependence" among a number of words forming a semantic entity, i.e., the high-order dependence that cannot be reduced to the random coincidence of lower-order dependence. We believe that identifying these high-order pure dependence patterns will lead to a better representation of documents. We first present two formal definitions of pure dependence: Unconditional Pure Dependence (UPD) and Conditional Pure Dependence (CPD). The decision on UPD or CPD, however, is a NP-hard problem. We hence prove a series of sufficient criteria that entail UPD and CPD, within the well-principled Information Geometry (IG) framework, leading to a more feasible UPD/CPD identification procedure. We further develop novel methods to extract word patterns with high-order pure dependence, which can then be used to extend the original unigram document models. Our methods are evaluated in the context of query expansion. Compared with the original unigram model and its extensions with term associations derived from constant n-grams and Apriori association rule mining, our IG-based methods have proved mathematically more rigorous and empirically more effective.

Keywords: Language Model, Word Association, High-order Pure Dependence, Information Geometry, Query Expansion, Log likelihood Ratio Test.

1 Introduction

The classical bag of words models, such as the Vector Space Model (VSM) [18] and unigram language model (LM) [16], represent a document as a weighted vector or probabilistic distribution of words. Although it has been proved useful in practice, there is a major limitation: the contextual information between words, which is the key to form meaningful semantic entities, is missing. In many cases, the semantic entities are not necessarily limited to syntactically valid phrases or named entities. More generally they can be high-order association (also referred as high-order *dependence*) patterns, which are often beyond pair-wise relations, e.g. {"climate", "conference", "Copenhagen"}.

G. Amati and F. Crestani (Eds.): ICTIR 2011, LNCS 6931, pp. 64–76, 2011.

Recently, there have been attempts to extract term relationships, e.g., through the Apriori method in [20], co-occurrence analysis [19], and Word-net relations [13]. In this paper, we propose to consider high-order *pure dependence*, i.e., the high-order dependence that cannot be reduced to the random coincidence of lower-order dependence. Usually these dependence patterns cannot be simply judged by co-occurrence frequencies. For example, the words *a*, *the* and *of* almost co-occur in every English article. However, we cannot say that they form a pattern representing a semantic entity. The high frequency of their co-occurrence can be explained as some kind of "coincidence", because each of them or pairwise combinations has a high frequency independently. On the other hand, the co-occurrence of the words "climate", "conference" and "Copenhagen" implies a un-separable high-level semantic entity, which can not be fully explained as the random coincidence of, e.g., the co-occurrence of "Copenhagen" and "conference" (which can be any other conferences in Copenhagen) and the occurrence of "climate". We consider a high-order dependence among words "pure", if and only if the joint probability distribution of these words is significantly different from the product w.r.t any possible decomposition into lower-order joint distributions or marginal distributions. In the language of graphical model, it requires that the joint distribution can not be factorized.

Formally, given a set of binary random variables $\mathbb{X} = \{X_1, \ldots, X_n\}$, where X_i denotes the occurrence ($X_i = 1$) or absence ($X_i = 0$) of the i-th word. Let $x_i \in \{0, 1\}$ denote the value of X_i. Let $p(\boldsymbol{x})$, $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$, be the joint probability distribution over $\mathbb{X}$. Then the n-order pure dependence over $\mathbb{X}$ can be defined as follows.

Definition 1. *(UPD): $\mathbb{X} = \{X_1, \ldots, X_n\}$ is of n-order Unconditional Pure Dependence (UPD), iff it can NOT be unconditionally factorized, i.e., there does NOT exist a k-partition $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_k\}$ of $\mathbb{X}$, $k > 1$, such that $p(\boldsymbol{x}) = p(\boldsymbol{c}_1) \cdot p(\boldsymbol{c}_2) \cdots p(\boldsymbol{c}_k)$, where $p(\boldsymbol{c}_i)$, $i = 1, \ldots, k$, is the joint distribution over $\mathbb{C}_i$.*

In practice, it is also useful to strengthen our definition of pure dependence in order to eliminate conditional random coincidences. This leads to the following definition of *conditional pure dependence*.

Definition 2. *(CPD): $\mathbb{X} = \{X_1, \ldots, X_n\}$ has n-order Conditional Pure Dependence (CPD), iff it can NOT be conditionally factorized, i.e., there does NOT exist $\mathbb{C}_0 \subset \mathbb{X}$ and a k-partition $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_k\}$ of $\mathbb{V} = \mathbb{X} - \mathbb{C}_0$, $k > 1$, such that $p(\boldsymbol{v}|\boldsymbol{c}_0) = p(\boldsymbol{c}_1|\boldsymbol{c}_0) \cdot p(\boldsymbol{c}_2|\boldsymbol{c}_0) \cdots p(\boldsymbol{c}_k|\boldsymbol{c}_0)$, where $p(\boldsymbol{v}|\boldsymbol{c}_0)$ is the conditional joint distribution over $\mathbb{V}$ given $\mathbb{C}_0$, and $p(\boldsymbol{c}_i|\boldsymbol{c}_0)$, $i = 1, 2, \ldots, k$, is the conditional joint distribution over $\mathbb{C}_i$ given $\mathbb{C}_0$.*

Remark 1. *Definition 2 permits an empty $\mathbb{C}_0$. Hence CPD entails UPD.*

To our best knowledge, there has not been any efficient method to characterize the above high-order pure dependence in both sufficient and necessary senses. For a given partition $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_k\}$ of $\mathbb{X}$, the method in [21] and [3] can efficiently decide whether $p(\boldsymbol{x}) = p(\boldsymbol{c}_1) \cdot p(\boldsymbol{c}_2) \cdots p(\boldsymbol{c}_k)$. However, it is an exponential task if we directly test all possible partitions of $\mathbb{X}$ and identify the n-order UPD. In

a configuration of graphical model, it can be shown that the decision problem of UPD or CPD is NP-hard [4].

Regarding the issue of efficiency, one may develop heuristics based on pair-wise dependence measures, e.g., covariance and correlation coefficient. Nonetheless, they usually suffer from the ad-hoc nature in tuning the threshold to decide significant pure dependence. Chi-square statistic can avoid the ad-hoc threshold, but it is indirect in the high-order case. Association rule mining can also be used to find highly frequent word associations. However, it does not guarantee the resulting associations are pure dependence. On the other hand, the complete n-gram method is straightforward, but it often leads to a large amount of redundant and noisy information.

In this paper, we propose to use Information Geometry (IG) [2], which provides relevant theoretical insights and useful tools, to tackle these difficulties in a consistent framework. IG studies joint distribution by way of differential geometry. A space of probability distributions is considered as a differentiable manifold, each distribution as a point on the manifold with the parameters of the model as coordinates. There are different kinds of coordinate systems to fit the manifold (detailed in Section 3), and it turns out that the so called mixed coordinate systems with orthogonality are especially useful for our purpose. Based on the coordinate orthogonality, we can derive a set of statistics and methods for analyzing word dependence patterns by decomposing the dependence into various orders. As a result, the 2nd-order, 3rd-order and higher-order pure dependence can be singled out and identified by the log likelihood ratio test.

The main theoretical contributions of this paper are that we propose a series of theoretically proven sufficient criteria for identifying UPD or CPD, respectively, and the corresponding efficient implementations that use the log likelihood test to the θ-coordinate of IG. The proposed IG-based methods can control confidence level theoretically. Then we apply the extracted high-order pure dependence (UPD or CPD) patterns in query expansion by incorporating then into the unigram document representation in the Relevance Model [9].

2 Related Work

This paper focuses on effective extraction and utilization of high-order pure word dependence patterns in the context of information retrieval (IR). There have been studies on incorporating dependence in language models. For example, Niesler et al. [15] presented a variable-length category-based n-gram language model, and Zhang et al. [23] proposed a framework for combining n-grams in different orders. Gao et al. presented a dependence language model to incorporate grammatical linkages [5]. The Markov Random Field (MRF) model captures short and long range term dependencies [11][12]. Song et al. [20] presented methods generating word associations based on association rule mining. Many enhancements to the classical bag-of-word representation of documents have been introduced, e.g., via the use of second-order co-occurrence information to build context vectors for word sense discrimination [19] and the combination of text

data with external knowledge (Wordnet) [13]. However, none of them explicitly considered high-order pure dependence.

The IG is systematically introduced by Amari [2] and has been successfully applied in the fields such as the study of neural spikes [14]. Based on IG, Hofmann [6] defined a Fisher kernel for learning document similarities by Support Vector Machines (SVM). However, the issue of high-order pure dependence was not considered in his work. In general, the application of IG in text processing tasks is not yet widely studied.

3 Preliminaries of Information Geometry

To illustrate our theoretical results and the corresponding algorithmic framework, it is necessary to explain the relevant background of IG [1][2][17][8].

3.1 Coordinates of Probability Distributions

In IG, a family of probability distributions is considered as a differentiable manifold with certain coordinate system. In the case of binary random variables, we use three basic coordinate systems, namely *p-coordinates*, *η-coordinates*, and *θ-coordinates* [14]. To be specific, if we define an assignment over $\mathbb{X}$, denoted by $a_{\mathbb{X}} =< a_1, a_2, \ldots, a_n >$ (or $a_{\mathbb{X}} = a_1 a_2 \ldots a_n$ in short), which determines a certain value of $\boldsymbol{x}$ by assigning $a_i \in \{0, 1\}$ to X_i, $1 \leq i \leq n$, then the coordinate systems of IG can be defined as follows:

1. *p*-coordinates:

$$p_{a_{\mathbb{X}}} = p_{a_1 a_2 \ldots a_n} = Pr\{X_1 = a_1, \ldots, X_n = a_n\} > 0 \tag{1}$$

   where $p_{a_{\mathbb{X}}}$ is the joint probability and $a_i \in \{0, 1\}, 1 \leq i \leq n$. Note that it is sufficient to determine a n-variable joint distribution using $2^n - 1$ probabilities, due to the constraint $\sum_{a_1, a_2, \ldots, a_n} p_{a_1 a_2 \ldots a_n} = 1$. Also note that IG requires that any probability term is not zero. This requirement can be met by using any common smoothing method.
2. η-coordinates:

$$\begin{aligned} \eta_i &= E[x_i], \quad 1 \leq i \leq n \\ \eta_{ij} &= E[x_i x_j], \quad 1 \leq i < j \leq n \\ &\vdots \\ \eta_{12\ldots n} &= E[x_1 x_2 \ldots x_n] \end{aligned} \tag{2}$$

   Note we define the order of a η-coordinate by the number of its subscripts. For example, η_1 is 1-order, and η_{23} is 2-order. In the information retrieval context, a η-coordinate is effectively equivalent to the document frequency of a single term or a term combination, up to a normalization factor.
3. θ-coordinates: The coordinate system specially relevant to our goal is the θ-coordinates, which can be derived from the log-linear expansion of $p(\boldsymbol{x})$:

$$\log p(\boldsymbol{x}) = \sum_i \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j + \cdots + \theta_{12\ldots n} x_1 x_2 \ldots x_n - \Psi \tag{3}$$

where Ψ is the normalization term corresponding to $\Psi = -\log p(\mathbf{0})$. It is easy to check that Formula (3) is an exact expansion since all x_i's are binary [14]. Note that we can also define the order of a θ-coordinate the same as in the η-coordinates.

As an example, we consider the case of $n = 3$. For the p-coordinate system, tuple-word joint distribution can be determined by arbitrary 7 out of 8 probabilities, e.g. $\{p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}\}$. The transform between p-coordinates and η-coordinates is trivial, say, $p_{111} = \eta_{123}, p_{011} = \eta_{23} - \eta_{123}, p_{100} = \eta_1 - \eta_{12} - \eta_{13} + \eta_{123}$. Based on formula (3), θ-coordinates can be given by the following equation if we have known p-coordinates:

$$\theta_{12\ldots n} = \log \prod_{k=0}^{n} \prod_{a_{\mathbb{X}} \in A_{\mathbb{X}}^{(k)}} p_{a_{\mathbb{X}}}^{(-1)^{n-k}} \tag{4}$$

where $A_{\mathbb{X}}^{(k)}$ denotes the set of all assignments, which assign 1 to k out of n variables, exactly. And based on formula (4), $\mathbb{X} = \{X_1, X_2, X_3\}$, $A_{\mathbb{X}}^{(0)} = \{000\}$, $A_{\mathbb{X}}^{(1)} = \{100, 010, 001\}$, $A_{\mathbb{X}}^{(2)} = \{101, 011, 110\}$, $A_{\mathbb{X}}^{(3)} = \{111\}$. Then we have

$$\theta_{123} = \log \frac{p_{111} p_{100} p_{010} p_{001}}{p_{110} p_{101} p_{011} p_{000}}.$$

Using the coordinate systems defined by the above, the set of all n-order joint probability distributions forms a d-dimensional manifold S_n, where $d = 2^n - 1$.

3.2 Coordinate Orthogonality

The Fisher information of two coordinate parameters ξ_i and ξ_j is defined as

$$g_{ij}(\boldsymbol{\xi}) = E\left[\frac{\partial \log p(\boldsymbol{x}, \boldsymbol{\xi})}{\partial \xi_i} \frac{\partial \log p(\boldsymbol{x}, \boldsymbol{\xi})}{\partial \xi_j}\right]$$

Here $E[\cdot]$ means the expectation with respect to $p(\boldsymbol{x}, \boldsymbol{\xi})$. In IG, the coordinate parameters ξ_i and ξ_j are called *orthogonal* when $g_{ij}(\boldsymbol{\xi}) = 0$ at any $\boldsymbol{\xi}$ [14].

From the definition of Fisher information, a direct observation is that, if ξ_i is orthogonal to ξ_j, the log-likelihood increment induced by $\Delta\xi_i$ is uncorrelated to the log-likelihood increment induced by $\Delta\xi_j$. Based on this observation, it can show that the maximum likelihood estimations of orthogonal parameters are independent to each other, and hence it entails a simple procedure of hypothesis test [14]. Note that such a simplification does not hold for other non-orthogonal parameterizations, e.g., correlation coefficients.

In Section 4, we will explicitly prove the theoretical connection between the n-order θ-coordinate and CPD (or UPD), which justifies that the θ-coordinate is

a relevant metric of high-order pure dependence. We thus aim to find a mixed coordinate system, denoted by ζ-coordinates, in which the high-order θ-coordinate parameter is orthogonal to all lower-order η-coordinates. This mixed coordinate system does exist: Generally, it can be shown that $\theta_{12\ldots n}$ is orthogonal to any η-coordinate less than n-order [14], and hence the (2^n-1)-dimensional ζ-coordinates can be given by $[\boldsymbol{\eta_1}, \ldots, \boldsymbol{\eta_{n-1}}, \theta_{12\ldots n}]^T$, where $\boldsymbol{\eta_1} = [\eta_1, \ldots, \eta_n]^T$, $\boldsymbol{\eta_2} = [\eta_{12}, \eta_{13}, \ldots, \eta_{(n-1)n}]^T$ and etc.

3.3 Coordinate Parameter Estimation

The θ-coordinates plays a central role in the identification of high-order pure dependence. However, a direct computation for high-order θ-coordinates can be numerically unstable. In addition, we desire a quantitative statistical significance level of the investigated θ-coordinate. Owing to the orthogonality between η-coordinates and θ-coordinates, Nakahara and Amari [14] develop a very efficient framework of Log Likelihood Ratio Test (LLRT) for θ-coordinates. However, Nakahara and Amari left the computation of high-order g_{dd} (the bottom-right element of the Fisher information matrix of ζ-coordinates) as an open problem, which is a necessary step for implementing the LLRT framework. To facilitate the LLRT framework, in the following Proposition 1, we develop a closed-form formula for computing g_{dd} in general[1].

Proposition 1

$$g_{dd} = \frac{1}{\sum_{\boldsymbol{x}} 1/p(\boldsymbol{x})} \tag{5}$$

The proof of Proposition 1 can be found in [7].

In the mixed ζ-coordinates, because of the orthogonality, the maximum likelihood estimation of the η's and the $\theta_{12\ldots n}$ can be performed independently [14]. Usually we can first estimate the η's from the corpus, and then calculate the $\hat{\theta}_{12\ldots n}$. In general, a larger absolute value of $\hat{\theta}_{12\ldots n}$ indicates a greater possibility that the word pattern is of pure dependence.

To guarantee a theoretic confidence level of the estimation for θ, the hypothesis test is needed. Here the null hypothesis $H_0 : \theta = \theta_0$, against $H_1 : \theta \neq \theta_0$. And we consider their log likelihood:

$$l_0 = \log p(\boldsymbol{x}; \hat{\boldsymbol{\eta}}, \theta_0), \quad l_1 = \log p(\boldsymbol{x}; \hat{\boldsymbol{\eta}}, \hat{\theta}).$$

We adopt the statistic of likelihood ratio test used in [14]

$$\begin{aligned}
\lambda &= 2\log\frac{l_1}{l_0} = 2\sum_{i=1}^{N}\log\frac{p(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}, \hat{\theta})}{p(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}, \theta_0)} \\
&\approx 2N \cdot E\big[\log\frac{p(\boldsymbol{x}; \hat{\boldsymbol{\eta}}, \hat{\theta})}{p(\boldsymbol{x}; \hat{\boldsymbol{\eta}}, \theta_0)}\big] = 2N \cdot D\big[p(\boldsymbol{x}; \hat{\boldsymbol{\eta}}, \hat{\theta}) : p(\boldsymbol{x}; \hat{\boldsymbol{\eta}}, \theta_0)\big] \\
&\approx N g_{dd}(\hat{\theta} - \theta_0)^2
\end{aligned} \tag{6}$$

[1] Recently, Nakahara independently gets a theoretical result similar to Proposition 1 (according to our personal communication with Nakahara).

Here N is the number of documents, $D\big[\cdot : \cdot\big]$ denotes the Kullback-Leibler divergence, $\hat{\theta}$ can be estimated by (4), g_{dd} is the Fisher information of the mixed coordinates ζ in the θ-direction at point $(\hat{\boldsymbol{\eta}}; \hat{\theta})$ and can be given by Proposition 1. Also note that the last approximation equation is entailed by the well-known approximate relation between Kullback-Leibler divergence and Riemannian distance [14]. In this paper, we are interested in identifying significant pure dependence w.r.t the θ-parameter (the relation between pure dependence and the θ-parameter is discussed in Section 4). Hence we let $\theta_0 = 2$ and only apply the LLRT to those $|\hat{\theta}|$'s that are greater than θ_0. On the other hand, if $|\hat{\theta}| \leq \theta_0$, we simply consider that the pure dependence is absent.

Asymptotically, according to Wilks' theorem, we have $\pm\sqrt{N g_{dd}(\hat{\theta} - \theta_0)^2} \sim N(0,1)$. Here N(0,1) denotes the standard normal distribution. Hence $\lambda \sim \chi^2(1)$, that is, the χ^2 distribution with degree of freedom 1. Then we can control the probability of error theoretically.

4 The Spectrum of High-Order Pure Dependence

In this Section, we first introduce two extra definitions on high-order pure dependence, namely Pair-wise Pure Dependence (PPD) and Theta Pure Dependence (TPD), which are the sufficient criteria of UPD and CPD, respectively. Note that, from an algorithmic perspective, PPD or TPD are far more feasible than directly deciding UPD or CPD. Finally, we clarify the spectrum of all kinds of high-order pure dependence defined by this paper.

Definition 3. *(PPD):* $\mathbb{X} = \{X_1, \ldots, X_n\}$ *has n-order Pair-wise Pure Dependence (PPD), iff every 2-order* θ*-coordinate* θ_{ij}*,* $1 \leq i < j \leq n$*, is significantly different from zero.*

Definition 4. *(TPD):* $\mathbb{X} = \{X_1, \ldots, X_n\}$ *has n-order Theta Pure Dependence (TPD), iff the n-order* θ *coordinate* $\theta_{12\ldots n}$ *is significantly different from zero.*

In Definitions 3 and 4, the significance level can be decided w.r.t an appropriate confidence interval of the LLRT described in Section 3.3. The following two propositions show the spectrum relation between PPD, TPD, UPD, and CPD.

Proposition 2. *PPD* $\Rightarrow$ *UPD.*

Proof. We will prove ¬UPD $\Rightarrow$ ¬PPD. Assume $\mathbb{X} = \{X_1, \ldots, X_n\}$ does NOT have the n-order UPD, i.e., there exists a nontrivial partition $\{\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_k\}$ of $\mathbb{X}$, such that $p(\boldsymbol{x}) = p(\boldsymbol{c}_1) \cdot p(\boldsymbol{c}_2) \cdots p(\boldsymbol{c}_k)$. Without loss of generality, we assume that X_1 and X_2 belong to $\mathbb{C}_1$ and $\mathbb{C}_2$, respectively. Summarize all variables of $p(\boldsymbol{x})$, except for X_1 and X_2. We have $\sum_{x_3 \ldots x_n} p(\boldsymbol{x}) = p(x_1)p(x_2)$. Hence, X_1 is independent to X_2, and θ_{12} vanishes by the definition of θ-coordinates (Formula 4). The proposition follows. □

Table 1. 2-order and 3-order pure dependence patterns (TREC AP8889)

Orders	2-order PD		3-order PD		
1	soviet	union	bush	jackson	vote
2	bush	democrat	bush	democrat	dole
3	bush	dole	republican	elect	presidenti
4	israel	palestinian	israel	palestinian	peac
5	attornei	judg	attornei	judg	trial
6	govern	rebel	militari	troop	rebel
7	militari	soldier	militari	troop	soldier

Index by Lemur toolkits v4.1 with Porter Stemmer

Proposition 3. *TPD* $\Rightarrow$ *UPD; TPD* $\Rightarrow$ *CPD*

Proof. We will first prove $\neg$UPD $\Rightarrow$ $\neg$TPD. First, we give several definitions and notations. Let $\mathbb{C} \subset \mathbb{X}$, $a_{\mathbb{C}}$ is a sub-assignment of $a_{\mathbb{X}}$ iff $a_{\mathbb{C}}$ assigns the same value to $\mathbb{C}$ as $a_{\mathbb{X}}$. We call an assignment (or sub-assignment) odd iff it assigns odd number of 1's to variables. Otherwise, it is an even assignment.

Let us consider the term inside the logarithmic function of $\theta_{12\ldots n}$, i.e., $\prod_{k=0}^{n} \prod_{a_{\mathbb{X}} \in A_{\mathbb{X}}^{(k)}} p_{a_{\mathbb{X}}}^{(-1)^{n-k}}$. According to Formula 4, if n is odd, the numerator and denominator of this term can be rewritten as $\prod_{a_{\mathbb{X}} \text{ is odd}} p_{a_{\mathbb{X}}}$ and $\prod_{a_{\mathbb{X}} \text{ is even}} p_{a_{\mathbb{X}}}$, respectively. On the other hand, if n is even, the numerator and denominator will be interchanged.

If the joint distribution $p(\boldsymbol{x})$ can be factorized, without loss of generality, assume that there exists a partition $\{\mathbb{C}_1, \mathbb{C}_2\}$ of $\mathbb{X}$, such that $p(\boldsymbol{x}) = p(\boldsymbol{c}_1) \cdot p(\boldsymbol{c}_2)$. Then, for an arbitrary given assignment $a_{\mathbb{X}}$, we have $p_{a_{\mathbb{X}}} = p_{a_{\mathbb{C}_1}} p_{a_{\mathbb{C}_2}}$. Let's count the occurring number of $p_{a_{\mathbb{C}_1}}$ in the numerator and denominator, respectively. We can see that the occurring number of $p_{a_{\mathbb{C}_1}}$ in the numerator is the same as the occurring number of $p_{a_{\mathbb{C}_1}}$ in the denominator, since the number of odd assignments is exactly the same as the number of even assignments. It turns out that every occurrence of $p_{a_{\mathbb{C}_1}}$ or $p_{a_{\mathbb{C}_2}}$ in the numerator can be eliminated by the corresponding occurrence in the denominator. Hence, we have $\prod_{k=0}^{n} \prod_{a_{\mathbb{X}} \in A_n^{(k)}} p_{a_{\mathbb{X}}}^{(-1)^{n-k}} = 1$, which entails a vanishing $\theta_{12\ldots n}$. Up to now, we indeed prove that TPD $\Rightarrow$ UPD.

If $p(\boldsymbol{x})$ can be conditionally factorized, we could show that $\theta_{12\ldots n}$ also vanishes by a similar approach. Hence, TPD $\Rightarrow$ CPD follows. $\square$

5 Implementation and Complexity Analysis

PPD requires that every pair of variables is significantly dependent. In order to decide whether n variables form a PPD pattern, we need perform C_n^2 times of LLRT on the involved 2-order θ parameters. In each 2-order LLRT procedure, we need sum all samplings to obtain the corresponding 4 p-coordinates and compute the corresponding g_{33}. These steps takes $O(N)$ time, where N is the number of samplings. Hence the identifying procedure of n-order PPD takes $O(n^2 N)$ time

in total. In practice, we are often interested in finding all maximal PPD patterns up to a given order $n_0 < n$. Here the maximal PPD pattern refers to the PPD pattern that cannot be enlarged. This problem is the maximal clique problem of the graph generated by the following rule: 1 A variable is denoted by a vertex; 2 An edge connects two vertices iff the corresponding two variables form a 2-order PPD pattern. As Tsukiyama et al. showed [22], it is possible to list all maximal cliques in a graph in an amount of time that is polynomial per generated clique. Hence our problem can be efficiently solved if the number of all maximal PPD patterns, up to n_0-order, is a polynomial function of n_0. The number of PPD patterns can be controlled by an appropriate significance level of LLRT.

In order to decide whether n variables form a TPD pattern, we need only to perform a single LLRT on the involved n-order θ parameter. The estimate of a n-order θ takes $O(N)$ time. Hence, the identifying procedure of a n-order TPD only takes $O(N)$ time in total.

Mining all TPD patterns, up to n_0-order, are much time-consuming since high-order TPD patterns can not be directly derived from the lower-order TPD patterns. Hence we adopt two pre-selection sets as the candidates of TPD patterns: 1 all PPD patterns up to n_0-order; 2 all frequent co-occurrence patterns, up to n_0-order, w.r.t certain frequency threshold. We then test whether the corresponding θ-coordinates of the candidate patterns are significantly different from zero. The TPD generated from the above two pre-selection sets are called TPD1 and TPD2, respectively.

As an illustration, here we show some interesting dependence patterns extracted from TREC AP8889 by PPD methods in Table 1.

6 Application

6.1 An Extended Relevance Model

In the framework of Relevance Model (RM), we estimate the probability distribution $P(w|R)$, where w is an arbitrary word and R is the unknown underlying relevance model, which is usually approximated by the topmost documents (e.g. n=50) of the initial retrieval. Then we pick up the words w with high probability $P(w|R)$, forming an expanded query.

The mining of $P(w|R)$ can be extended to incorporate the word patterns with high-order pure dependence. In this section, we provide an extended relevance model, which employs the high-order pure dependence as a complement of the classic relevance model. We pick the top n returned documents of the initial retrieval, and extract the high-order dependence patterns using various different methods. For each dependence pattern c in the dependence set C, we calculate

$$P(c|R) = \frac{\text{Number of chunks containing } c}{\text{Total number of chunks}}.$$

Intuitively, we belief that a word in some high-order pure dependence patterns should carry more semantic importance. Hence we interpolate the weight due to high-order pure dependence with the weight estimated using the interpolated relevance model RM3 [9][10].

$$D_{combine}(w|R) = \lambda D(w|R) + (1-\lambda)P(w|R). \quad (7)$$

where $D(w|R) = \sum_{c:w\in c} P(c|R)$.

We consider $D_{combine}(w|R)$ as the new weight for word w in our extended relevance model. The following experimental results shows that this extended model outperforms the classical model significantly in most cases.

6.2 Experimental Setup

We evaluate our model using four TREC collections: AP8889 with topic 101-150 (the *title* field), AP8889 with topic 151-200 (the *title* field), AP8889 with topic 201-250 (the *desc* field), and WSJ9092 with topic 201-250 (the *desc* field). Lemur 4.12 is used for indexing and retrieval.The first-round retrieval is carried out by a baseline language modeling (LM) approach with $\mu = 1000$. The Relevance Model (RM) is selected as the second baseline method with 50 feedback documents.

6.3 Results and Analysis

Figure 1 shows the 11-point interpolated average precision on TREC AP8889 and WSJ9092 datasets. We can see that all the query expansion method outperform the baseline language model, while the combined extended model is the best.

To further examine the merit of our IG-based high-order pure dependence model, we furthermore compare it with several other high-order dependence models, as shown in Table 2 (To keep it clean, we do not draw the curves of all methods on Figure 1). In Table 2, "Apr" indicates the Apriori method, which has many successful applications for finding the interesting item patterns. "CO" ("ConstOrder") indicates considering all the possible k-order word patterns. Due to the time and space limitations, we only examined the $k \leq 3$ case. "PPD", "TPD1" and "TPD2" indicate the methods described in Section 5. The combined methods are described in Section 6.1.

We can see that all high-order models outperform the baseline uni-gram RM. This verifies our intuition that the uni-gram RM and the high-order model are complementary to each other. Note that the best result can be achieved when the coefficient λ in (7) is set to about 0.1.

Table 2. MAP Performance comparison

QE Methods	AP8889 101-150	AP8889 151-200	AP8889 201-250	WSJ9092 201-250
LM	0.2331	0.3138	0.0862	0.1948
RM	0.3086	0.4042	0.0879	0.2060
PPD	0.2963 (-4.99%)	0.3859 (-4.53%)	0.0865 (-1.59%)	0.2402 (+16.60%)*
RM+CO	0.3109 (+0.75%)*	0.4101 (+1.46%)	0.0949 (+7.96%)	0.2121 (+2.96%)
RM+Apr	0.3093 (+0.23%)*	0.4168 (+3.12%)*	0.0900 (+2.39%)	0.2176 (+5.63%)*
RM+PPD	**0.3173** (+2.82%)*	0.4218 (+4.35%)*	0.0999 (+13.65%)*	**0.2488** (+20.78%)*
RM+TPD1	0.3153 (+2.17%)*	**0.4232** (+4.70%)*	**0.1003** (+14.11%)*	0.2441 (+18.50%)*
RM+TPD2	0.3166 (+2.58%)*	0.4191 (+3.69%)*	0.0972 (+10.58%)*	0.2211 (+7.33%)*

Significant improvements (at level 0.05) over RM are marked with "".

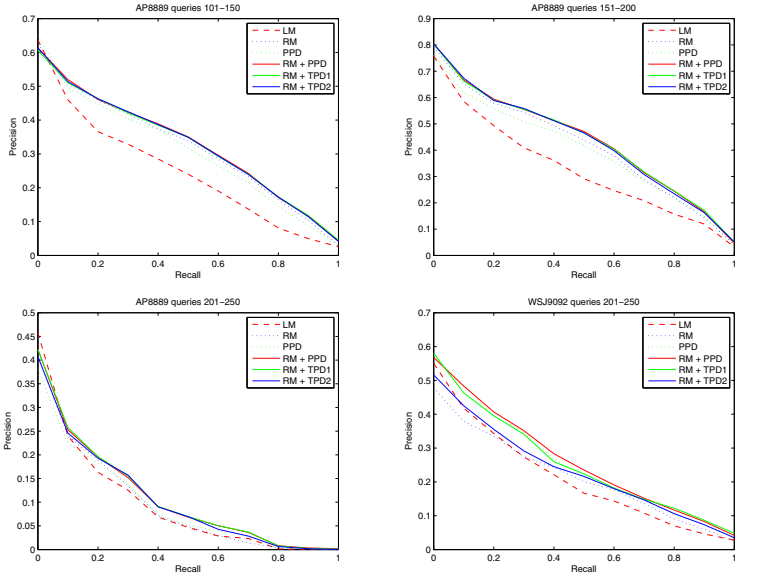

Fig. 1. P-R curve on TREC AP and WSJ

We can also note the PPD/TPD method outperform ConstOrder method and Apriori method significantly, especially on the WSJ9092 dataset. We believe one of the reasons is that the query we selected for WSJ9092 dataset (the *desc* field of topic 201-250) are long and complicated, in which case our IG-based high-order pure model have more advantages.

To show the different performance between TPD and PPD, we compare the results from different parameter λ's. It is shown that the averaged performance is almost the same, but the TPD method is more stable on sub-optimal parameter setting, suggesting that, if we cannot afford the time to train the parameters of the model, TPD method is "safer". In addition, the set of TPD patterns is often much reduced, which can offer a more economic high-order model.

7 Conclusions and Future Work

We analytically clarified a spectrum of high-order pure dependence, and proposed a novel framework based on Information Geometry to extract high-order pure word dependence patterns from documents. In this IG-based framework, we developed a set of rigorously-established justifications and feasible algorithms to single out high-order pure dependence by a well-founded statistical procedure (i.e. the log likelihood ratio test). We also integrate the automatically derived high-order pure dependence patterns into the Relevance Model. Evaluation results demonstrated the usefulness of the high-order pure dependence, and the effectiveness and robustness of our IG-based approach.

Our future work will be focused on addressing the following issues. First, we will perform a systematic analysis to clarify the semantic distinctions between PPD and TPD. Second, we will compare our approach with stronger baselines that utilize term dependence in IR, e.g., the dependence language model [5] and the MRF model [11]. Finally, we exploit the integration of a suitable level of syntactical dependence information into our framework.

Acknowledgements. The authors would like to thank anonymous reviewers for their constructive comments. This work is supported in part by the Natural Science Foundation of China (NSFC, grant 61070044); NSF of Tianjin (grant 09JCYBJC00200); the NSFC-RSE (Royal Society of Edinburgh) International Joint Project Scheme; the EU FP7 through its Marie Curie IRSES (grant 247590); and the UK's Engineering and Physical Sciences Research Council (grant EP/F014708/2).

References

1. Amari, S.: Information geometry on hierarchy of probability distributions. IEEE Transactions in Information Theory 47(5), 1701–1711
2. Amari, S., Nagaoka, H.: Methods of Information Geometry. American Mathematical Society, Providence (2001)
3. Bakirov, N.K., Rizzo, M.L., Székely, G.J.: A multivariate nonparametric test of independence. Journal of Multivariate Analysis 79(8), 1742–1756
4. Chickering, D., et al.: Large-sample learning of bayesian networks is np-hard. The Journal of Machine Learning Research 5, 1287–1330
5. Gao, J., Nie, J.Y., et al.: Dependence language model for information retrieval. In: Proceedings of SIGIR 2004, pp. 170–177 (2004)
6. Hofmann, T.: Learning the similarity of documents: An information-geometric approach to document retrieval and categorization
7. Hou, Y., et al.: Efficient factorization test and high-order pure dependence mining. Submitted to NIPS 2011 (2011)
8. Jeffreys, H.: An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A 186 (1946)
9. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: Proceedings of SIGIR 2001, pp. 120–127 (2001)
10. Lv, Y., Zhai, C.: A comparative study of methods for estimating query language models with pseudo feedback. In: Proceedings of CIKM 2009, pp. 1895–1898 (2009)
11. Metzler, D., Croft, W.B.: A markov random field model for term dependencies. In: Proceedings of SIGIR 2005, pp. 472–479 (2005)
12. Metzler, D., Croft, W.B.: Latent concept expansion using markov random fields. In: Proceedings of SIGIR 2007, pp. 311–318 (2007)
13. Mihalcea, R., Corley, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of AAAI 2006, pp. 775–780 (2006)
14. Nakahara, H., Amari, S.: Information geometric measure for neural spikes. Neural Computation 14(10), 2269–2316
15. Niesler, T.R., Woodland, P.C.: A variable-length category-based n-gram language model. In: Proceedings of IEEE ICASSP 1996, pp. 164–167 (1996)

16. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of SIGIR 1998, pp. 275–281 (1998)
17. Rao, C.R.: Information and Accuracy Attainable in the Estimation of Statistical Parameters. Bull. Calcutta. Math. Soc. 37 (1945)
18. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11)
19. Schütze, H.: Automatic word sense discrimination. Computational Linguistics 24(1), 97–123
20. Song, D., Huang, Q., Rueger, S., Bruza, P.: Facilitating query decomposition in query language modeling by association rule mining using multiple sliding windows. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 334–345. Springer, Heidelberg (2008)
21. Taskinen, S., Oja, H., Randles, R.H.: Multivariate nonparametric tests of independence. Journal of the American Statistical Association 100(471), 916–925
22. Tsukiyama, S., Ide, M., Ariyoshi, H., Shirakawa, I.: A new algorithm for generating all the maximal independent sets. SIAM Journal on Computing 6(3), 505–517
23. Zhang, S., Dong, N.: An effective combination of different order n-grams. In: Proceedings of O-COCOSDA 2003, pp. 251–256 (2003)