Vol.2010-SLP-82 No.9
2010/7/23

情報検索のための確率的言語モデル

江　口　浩　二[†1]

情報検索のための確率的言語モデルは 1998 年に Ponte と Croft によって提案されてから，情報検索やそれに関連する課題に対する新たなアプローチとして注目を浴びてきた．その特徴の一つに，それまでに研究されてきたベクトル空間モデルや古典的確率型検索モデルで導入された発見的方法を極力用いず，数理的に説明可能な枠組みである点が挙げられる．その表現能力と柔軟性の高さにより，適用範囲は非構造なテキストデータに対する種々のタスクだけでなく，構造化文書検索やクロスメディア検索にも及ぶ．そこで，本稿では，情報検索のための確率的言語モデルついて概要を述べ，その研究動向を紹介する．

Probabilistic Language Models for Information Retrieval — A Survey

Koji Eguchi[†1]

Probabilistic language models were first applied to information retrieval by Ponte and Croft in 1998. Since then, this approach has attracted extensive attentions as a new paradigm of information retrieval and its related technologies. It is particularly notable that this approach avoids having to use heuristics that was heavily employed in vector-space model and classical probabilistic retrieval models and thus this approach is mathematically explicable. Because of its high expressive power and flexibility, the probabilistic language modeling approach has been applied not only to the tasks over unstructured text data, but also to structured document retrieval and cross-media retrieval. This paper gives an overview and survey of probabilistic language models for information retrieval.

†1 神戸大学大学院システム情報学研究科
Graduate School of System Informatics, Kobe University

1. はじめに

情報検索のための確率的言語モデルは 1998 年に Ponte と Croft[1] によって提案されてから，情報検索やそれに関連する課題に対する新たなアプローチとして注目を浴びてきた．それまでに研究されてきたベクトル空間モデル[2),3)] や古典的確率型検索モデル[4)] で導入された発見的方法を極力用いず，数理的に説明可能な枠組みを指向している点が特徴と言える．また，このことから，統計科学や統計的学習理論とも相性がよいことで知られている．その表現能力と柔軟性の高さにより，適用範囲は非構造なテキストデータに対する種々のタスクだけでなく，構造化文書検索やクロスメディア検索にも及ぶ．当該技術は，その国際的な注目の高さに比して，国内では十分に知られているとは言い難い．本稿では，この 10 年余りで大きく発展を遂げ，今なお研究が盛んな，情報検索のための確率的言語モデルについて概要を述べ，その研究動向を紹介する．

2. クエリ尤度モデル

一般に，情報検索は，蓄積された情報から利用者の要求に応じて適合する部分を取り出す問題である[⋆1]．蓄積された情報が文書という単位で表現される状況を考える．このとき，クエリと文書コレクションが与えられた状況で，クエリに適合する度合いに従って文書をランキングする機能が極めて重要となる．本稿においては，特に断りがない限り，情報検索は上に述べたような問題を指すものとする．

文書 D がクエリ Q に適合する確率 $P(D|Q)$ を求めることができれば，これに従って情報検索を実現することができる．古典的確率型検索モデル[4)] では $P(D|Q)$ を計算するにあたって様々な発見的方法を駆使するのに対して，ここでは異なるアプローチを考える．まず，ベイズの定理を用いて次式を得る．

$$P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)} \propto P(Q|D)P(D) \quad (1)$$

上のように $P(Q)$ は文書に依存しないので省略できる．また，$P(D)$ は文書に関して何らかの事前知識がない限り，一様であると見なす（見なさざるを得ない）[⋆2]．このとき，$P(D|Q)$ による文書ランキングにおいては，$P(Q|D)$ が本質的な意味を持つ．すなわち，式 (1) を次

⋆1 広義の情報検索として，情報フィルタリングや情報提示などを含めることもある．
⋆2 文書に関する事前知識を導入する場合について本章で後述する．

のように簡略化する．

$$P(D|Q) \propto P(Q|D) \tag{2}$$

上式の右辺は文書に関する言語モデルがクエリを生成する確率を表しており，上式に基づく情報検索手法は**クエリ尤度モデル**（query likelihood model）と呼ばれる．

クエリ尤度モデルは 1998 年に Ponte と Croft[1] によって最初に提案された[⋆1]．その後現在まで，クエリ尤度モデルは言語モデルに基づく情報検索の基本的なアプローチとなっている[5)–7)]．クエリ尤度モデルには，文書に関して仮定する分布の型やスムージングの方法などによって種々のバリエーションがある．本章の以下ではクエリ尤度モデルの基本的な実現方法について概説する．それらのモデルは従来の検索モデルと同等かそれ以上に効果的で，従来の TF-IDF 法による情報検索[3] の代替手段となり得る．

2.1 多項分布に基づくクエリ尤度モデル

クエリ尤度モデルにおいて文書を言語モデルで表現する際に仮定する分布の型として，多変量ベルヌーイ分布[1]，多項分布[5),6)] などが用いられる．なかでも現在最も用いられることが多いのは多項分布である．多項分布による文書モデルは，ユニグラム言語モデルとも呼ばれ，語が独立に生起すると仮定する．このとき，文書 D のユニグラム言語モデル θ_D からクエリ Q が生成される尤度，すなわちクエリ尤度 $P(Q|\theta_D)$ は次式のように定義される．

$$P(Q|\theta_D) = \prod_{\ell=1}^{|Q|} P(q_\ell|\theta_D) \tag{3}$$

上の式では $Q = q_1, \ldots, q_{|Q|}$ すなわち個々のクエリ語が互いに異なるものとして扱った．クエリ Q において語彙 $w_i \in V = \{w_1, \ldots, w_{|V|}\}$ が複数回出現し得ることを考慮すると，次式を得る．

$$P(Q|\theta_D) = \prod_{w_i \in V} P(w_i|\theta_D)^{c(w_i,Q)} \tag{4}$$

ここで $c(w_i, Q)$ は，クエリ Q において語彙 w_i が出現する頻度を指す．上式は正規化定数を省略した多項分布の式に他ならない．このように，当該モデルによる検索の問題は $P(w_i|\theta_D)$ の推定問題に帰着する．

2.2 多変量ベルヌーイ分布に基づくクエリ尤度モデル

文書モデルに多変量ベルヌーイ分布を用いる場合について述べる．それぞれの単語 w_i に対して 2 値確率変数 X_i を定義して，クエリに単語 w_i が出現する場合に $X_i = 1$，出現しない場合に $X_i = 0$ とする．X_i はそれぞれの単語において互いに独立であると仮定する．このとき，多変量ベルヌーイモデル θ_D のパラメータ数は，語彙数と同じになる．そのようなモデルはクエリ中の単語の出現の有無によって形式化でき，制約は $P(X_i = 1|\theta_d) + P(X_i = 0|\theta_d) = 1$ $(i = 1, \ldots, |V|)$ となる．多変量ベルヌーイモデルによるとクエリ尤度は次式で得られる．

$$P(Q|\theta_D) = \prod_{w_i \in Q} P(X_i = 1|\theta_D) \prod_{w_j \notin Q} \big(1 - P(X_j = 1|\theta_D)\big) \tag{5}$$

右辺における一つ目の項は文書がクエリ中に現れる語を生成する確率，二つ目の項はその他の語を生成しない確率を示す．ここでは検索の問題は $P(X_i = 1|\theta_D)$ の推定問題とみなせる．

多変量ベルヌーイモデルにおいては，多項分布モデルに見られるように語の出現頻度を捉えることができないという短所がある．多変量ベルヌーイモデルは Ponte と Croft が最初に提案したクエリ尤度モデルにおいて用いられた[1] が，現在では，多項分布モデルの方が一般的に広く用いられている．実際に，多項分布モデルの方が多変量ベルヌーイモデルよりも有効であるとする実験結果が報告されている[6]．一方で，語の出現頻度が比較的重要でない状況において多変量ベルヌーイモデルの有効性が示されている[8]．

2.3 文書モデルの推定

現在，情報検索のための言語モデルにおいて最も広く用いられているのは多項分布モデルであるので，これを想定して，文書モデルのパラメータ θ_D の推定方法について述べる．

文書モデルの最も単純な推定方法は次式のように相対頻度を用いることである．

$$P_{ml}(w_i|\theta_D) = \frac{c(w_i, D)}{|D|} \tag{6}$$

ここで，$c(w_i, D)$ は文書 D において語 w_i が出現する頻度，$|D|$ は文書 D の文書長すなわち延べ総語数を示す．これは文書における単語の出現が観測された状況での，多項分布のパラメータに関する最尤推定に他ならない．ところが，式 (4)，式 (6) より，文書中に出現しない，すなわち文書モデルにおいて非零の確率値が割り振られていない語 w_i がクエリに一つでも存在すると，他にどのようなクエリ語が存在していようとも尤度 $P(Q|\theta_D)$ が 0 になり，クエリ語が一つも存在しない文書との比較ができなくなることに注目されたい．このような状況は情報検索タスクにおいて好ましくない．これを**零確率問題**（zero-probability problem）と呼ぶ．また，文書長が小さい場合に，最尤推定では正確なモデル推定を行うことができないことが多い．これらの問題に対処するために，文書モデルの**スムージング**

⋆1 ほぼ同時期に Hiemstra らの研究もある[5)]．

(smoothing) が用いられる.

スムージングでは,何らかの方法で文書に出現しない語に対しても微小な確率値を割り振る.それにより,より多くのクエリ語が出現しない文書ほど,式 (4) に従って,それらの微小な確率値が掛け合わさることになり,結果としてのクエリ尤度の値が小さくなる.このようにして前述の零確率問題に対処することができる.スムージングの実現方法としては,音声認識や機械翻訳のための言語モデルで研究されてきた手法を適用することが考えられるが,情報検索タスクの特徴を捉えたスムージング手法も独自に発展してきた.以下では代表的なスムージング手法を紹介する.

2.3.1 線形補間法

線形補間法 (linear interpolation または Jelinek-Mercer smoothing)[10] では,文書コレクションに関するモデルを用いてスムージングを行う.このとき,次式のように,スムージングの度合いを定数 $\lambda \in [0,1]$ で表現するのが特徴である.

$$P(w_i|\hat{\theta}_D;\lambda) = \lambda \frac{c(w_i,D)}{|D|} + (1-\lambda)P(w_i|\theta_C) \tag{7}$$

ここで,C は文書コレクションを示し,$P(w|\theta_C)$ はコレクションモデル (collection language model, background language model または reference language model) と呼ばれるものである.これは,例えば,以下のようにして推定できる.

$$P(w_i|\theta_C) = \frac{\sum_{D\in C} c(w_i,D)}{\sum_{D\in C}|D|} \tag{8}$$

式 (7) からわかる通り,$\lambda \approx 1$ のときに式 (6) で示した文書の最尤推定モデルが強調され,逆に $\lambda \approx 0$ のときは軽視される.

2.3.2 ディリクレ・スムージング

線形補間法ではスムージングの度合いを固定したのに対して,これを文書長に応じて可変とした拡張手法に**ディリクレ・スムージング** (Dirichlet smoothing または Dirichlet prior smoothing)[9] がある.これはディリクレ事前分布を用いて文書多項分布に事前知識を導入するものであり,次式によってスムージングを行う.

$$P(w_i|\hat{\theta}_D;\mu) = \frac{c(w_i,D)+\mu P(w_i|\theta_C)}{|D|+\mu} \tag{9}$$

これをコレクションモデルによる補間として表現すると,次式のように変形できる.

$$P(w_i|\hat{\theta}_D;\mu) = \frac{|D|}{|D|+\mu}\frac{c(w_i,D)}{|D|} + \frac{\mu}{|D|+\mu}P(w_i|\theta_C) \tag{10}$$

ただし,μ は正の値を持つパラメータである.補間の度合いが文書長によって決まる係数で表現されている点に注意されたい.これは,長い文書は観測サンプル数が十分に大きいのでスムージングの必要性が低く,逆に短い文書では必要性が高いことを反映している.実際に,式 (10) において $|D| \to \infty$ のとき,式 (6) で示した文書の最尤推定モデルと等価となる.式 (7) と比較すると,$\lambda = \frac{|D|}{|D|+\mu}$ となっている.λ を固定することは文書長に多様性がある場合にスムージングをうまくコントロールできないことを意味する.

Zhai ら[9] は,以上に述べた補間に基づくスムージングと,音声認識の分野で広く用いられるバックオフ・スムージング (back-off smoothing) を,情報検索の実験において比較している.そこでは,補間に基づくスムージングの方がバックオフ・スムージングよりも有効であり,補間に基づくスムージングのなかでもディリクレ・スムージングが線形補間法よりも有効であることが経験的に示されている.

2.4 文書事前分布

本章の冒頭の式 (1) において,$P(D)$ は文書に関して何らかの事前知識がない限り,一様であると見なすと述べた.例えば,リンク解析に基づいて Web ページの重要度を与える PageRank[11] を典型とした,文書に関するクエリに依存しない事前知識を利用することを考える.式 (1) における $P(D)$ は**文書事前分布** (document prior) と呼ばれ[⋆1],これによって文書のクエリ非依存な優先度を,クエリ尤度 $P(Q|D)$ と自然な形で結合することができる.このようなベイズ統計学的な発想は,情報検索タスクにおいて複数の特徴を組み合わせるのに役立つことがある.

典型的な例として,Kraaij ら[12] は,Web を対象にした既知事項検索である**指定ページ発見** (named page finding) の実現のため,サイトトップページは URL 文字列が短い傾向にあることに着目し,訓練データを用いて文書事前分布 $P(D)$ を推定し,式 (1) を用いることで検索性能を大幅に改善した.

3. クエリ尤度モデルの拡張

2 では,コレクションモデルによる,比較的単純なスムージング手法を用いたクエリ尤度モデルを取り上げた.本章では,これらの拡張に関していくつかの重要な研究動向を示す.

3.1 翻訳モデル

同義語と多義語の問題に対処するため,Berger と Lafferty[13] は,語 s が語 t に言い換え

⋆1 ベイズの定理において $P(D)$ に対応する項は一般に事前確率または事前分布と呼ばれる.

られる（翻訳される）確率 $P(t|s)$ を用いて，クエリ尤度を計算する翻訳モデル（translation model）を提案し，クエリ尤度モデルを拡張した．ここで「翻訳」という表現を用いているが，ここでは単言語における言い換えを意味していることに注意されたい．このモデルでは，クエリ $Q = Q = q_1, \ldots, q_{|Q|}$ に対するクエリ尤度が以下の式によって計算される．

$$P(Q|\theta_D) = \prod_{\ell=1}^{|Q|} \sum_{w_i \in V} P(q_\ell|w_i) P(w_i|\theta_D) \tag{11}$$

$P(q_\ell|w_i)$ は 語 w_i が q_ℓ に翻訳される確率を示す．

元々の翻訳モデルは，統計的機械翻訳の目的で IBM 社の研究グループによって提案された[14]．上で用いられたものはその中で最も基本的なモデルである．このモデルを推定するには，十分な量の適合判定データすなわちクエリ・適合文書対が訓練データとして必要となる．一般的にこのような訓練データは容易に得られないため，Berger と Lafferty の研究[13]ではこのような訓練データを発見的手法によって機械的に作成している．

翻訳モデルは，その自然な拡張として言語横断検索に応用できる[15]．また，興味深い応用例として，FAQ アーカイブ，コミュニティ指向質問応答サービスなどにおける質問・回答対に対する検索がある[16),17)]．「夕食をとるのにどこか良いところをご存じですか？」，「レストラン○○でおいしいメキシコ料理が食べれますよ．」という例に共通の語彙は一つとしてないことからわかる通り，質問と回答には語彙的なギャップが存在する．これを翻訳モデルで対処しようとするのが上記の研究のねらいである．

3.2 クラスタ型スムージング

2.3 節で述べた，線形補間法やディリクレ・スムージングなどの補間に基づくスムージングでは，同一のコレクションモデルですべての文書のスムージングが行われる．文書モデルをより適切に推定するという目的のもとでは，文書に出現する語の同義語や類義語などを反映したモデル推定ができないか，と考えるのは自然な発想である．このような考えのもと，事前に文書コレクションを解析して得た文書クラスタを，文書モデルのスムージングの手段として利用する研究がある[18)–20)]．このような目的で用いる文書クラスタリングは，ハードクラスタリングすなわち各文書をただ一つのクラスタに割り当てる手法[19)] と，ソフトクラスタリングすなわち各文書の複数クラスタへの帰属の度合を出力する手法[18),20)] に大別できる．次式は，ソフトクラスタリングの典型的な手法である潜在的ディリクレ配分法（latent Dirichlet allocation: LDA）[21)] を用いた Wei らの研究[20)] による．

$$P(w_i|\hat{\theta}_D; \lambda, \mu) = \lambda\Big(\frac{|D|}{|D|+\mu} P_{ml}(w_i|\theta_D) + \frac{\mu}{|D|+\mu} P(w_i|\theta_C)\Big) + (1-\lambda) P_{tm}(w_i|D) \tag{12}$$

ここでは，まず最尤推定文書モデルに対してディリクレ・スムージングが適用され，次に $P_{tm}(w_i|D)$ を用いた補間が行われている．$P_{tm}(w_i|D)$ のみを用いた場合は，トピックの粒度が粗すぎて検索に有効でないことが指摘されている[20)]．トピック数を K とするとき，$P_{tm}(w_i|D)$ は次式によって得られる．

$$P_{tm}(w_i|D) = \sum_{k=1}^{K} P(w_i|t_k) P(t_k|D) \tag{13}$$

文書に関するトピック分布 $P(t_k|D)$，および，トピックに関する単語分布 $P(w_i|t_k)$ は，ベイズ推定によって求める[21)–23)]．

以上に述べたクラスタ型スムージングの効果については，とくに LDA を用いた方法[20)] が有望であることが実験的に示されている．近年，トピックモデルの研究は非常に活発に行われており，それらの情報検索への適用は重要な課題であると言える．例えば，筆者らは LDA の変種である多型トピックモデル[24),25)] を構造化文書検索に適用し，その有効性を示している[26)]．

3.3 語間依存性モデル

文書多項分布すなわちユニグラム言語モデルに基づくクエリ尤度モデルの拡張を考えるとき，直感的には N グラムモデル（N-gram model）への展開が考えられる．例として，バイグラムに基づくクエリ尤度は次式で与えられる[6)]．

$$P(Q|\theta_D) = P(q_1|\theta_D) \prod_{i=2}^{|Q|} P(q_i|q_{i-1}, \theta_D) \tag{14}$$

$P(q_i|q_{i-1}, \theta_D)$ は文書 D のもとで q_{i-1} の直後にクエリ語 q_i が生成される確率である．N グラムモデルが単語の位置に関する依存性を捉えているに対し，文法構造に関する依存性に着目する研究もある．Gao らの研究[27)] を例にとると，ある文書に対するクエリの対数尤度は次式のようになる．

$$\log P(Q|\theta_D) = \log P(L|D) + \sum_{i=1}^{|Q|} \log P(q_i|\theta_D) + \sum_{(i,j)\in L} MI(q_i, q_j|L, D) \tag{15}$$

ここで，L は構文解析によって得られたクエリ Q における語間依存関係集合で，MI は相互情報量である．L が空であるとき上式は多項分布に基づくクエリ尤度の対数と等価であり，そうでないとき右辺の最終項はクエリ語間の依存性を捉える．

以上に述べた研究では，語間依存性を考慮することによってある程度の効果が得られるが，次に述べる手法と比較するとその効果は限定的であると言える．

一方，**近接特徴量**（proximity feature）を用いた手法では有望な結果が報告されている．なかでも，Metzler と Croft[28] は，**マルコフ確率場モデル**（Markov random field model）に基づいてクエリにおける語間依存性を反映した文書ランキングを実現している．これをさらに拡張した筆者らの研究[29] では，クエリを構成する複合語間の緩やかな依存関係と，個々の複合語における構成語間の緊密な依存関係に着目した文書ランキングを実現し，日本語 Web 検索の実験において有効性を実証している．これは日本語をはじめとするいくつかの東アジア言語に対する情報検索の実現において，しばしば問題となる複合語の扱いに対して有効な解決策を与える．語間依存性のモデリングの展開として，より複雑な語間依存性を呈する自然言語クエリ（natural language query または sentence query）への対処が挙げられる．

4. 分布間距離に基づく検索モデル

2.1 で述べたクエリ尤度モデルにおいて，式 (4) の降順による文書のランキングは，次に示すクエリモデル θ_Q と文書モデル θ_D 間の負の相対エントロピー（KL ダイバージェンス）の降順による文書ランキングと等価である．

$$-KL(\theta_Q||\theta_D) = \sum_{w_i \in Q} P(w_i|\theta_Q) \log P(w_i|\theta_D) - \sum_{w_i \in Q} P(w_i|\theta_Q) \log P(w_i|\theta_Q) \quad (16)$$

$$\propto \sum_{w_i \in Q} P(w_i|\theta_Q) \log P(w_i|\theta_D) \quad (17)$$

なお，θ_Q は最尤推定によって得られる．式 (16) の右辺第 2 項は文書 D に依存しないため，文書ランキングに影響しない．そこで，式 (17) ではこの項を省略している．つまり，式 (4) の降順による文書のランキングは，式 (17) の右辺に示された，クエリモデル θ_Q と文書モデル θ_D 間の負の交差エントロピー（クロスエントロピー）の降順と等価であるとも言える．このような分布間距離に基づく文書ランキングはさらに発展させることができる．本章では，この拡張形として分布間距離に基づく検索モデルについて述べる．このようなアプローチとして典型的なものに，Lavrenko と Croft による**適合モデル**（relevance model）[31] が挙げられる．以下に，その概要を述べる．

クエリを Q，適合性を表す 2 値確率変数を $R \in \{r, \bar{r}\}$ とする．Q は観測されるものの R は観測されない現実的な状況を考える．このとき，ある語 w が R のもとで生成される確率を次のように近似する．

$$P(w|R=r) \approx P(w|Q, R=r) \approx \sum_{D_i \in C} P(w|\theta_{D_i}) P(\theta_{D_i}|Q, R=r) \quad (18)$$

上式はユーザから与えられたクエリの背後に潜在する情報要求に関する言語モデルを表しているといえ，これを適合モデル[31] と呼ぶ．

次に，適合モデルを用いて検索対象の文書集合のランキングを行う方法について述べる．適合モデルを推定することは，与えられたクエリ Q の背後に潜在する情報要求に関する言語モデル θ_Q を推定することに他ならない．これと文書 D_i に関する言語モデル θ_{D_i} との負の交差エントロピー，すなわち，

$$-H(\theta_Q||\theta_{D_i}) = \sum_{w \in V} P(w|\theta_Q) \log P(w|\theta_{D_i}) \quad (19)$$

の降順に文をランキングする．ただし，V は文書集合に含まれる語彙集合を示す．

以上に述べた適合モデルは情報検索でしばしば用いられる疑似適合フィードバック（pseudo-relevance feedback）の言語モデルによる実現と見なすことができ，同義語と多義語の問題を軽減する効果がある．数々の実証実験において，適合モデルはクエリ尤度モデルと比較して検索有効性が高いことが報告されている[31]．適合モデルは通常の情報検索だけでなく，言語横断検索[32]，クロスメディア検索[33] など，その応用範囲の広さと性能の高さで知られている．また，筆者らは，クエリ尤度モデルと適合モデルを拡張して，意見検索すなわちクエリで示された事物に関する意見を検索するタスクを実現している[34]．

5. おわりに

情報検索のための確率的言語モデルは様々なタスクに適用されてきた．すでに本文で紹介した Web 検索[12] や言語横断検索[15),32]，質問・回答対検索[16),17]，構造化文書検索[26]，日本語情報検索[29]，意見検索[34]，クロスメディア検索[8),33] の他にも，本文では紹介しきれなかった研究事例が数多く存在する．筆者のサーベイ論文[35] でもいくつかの研究事例を紹介しているので参照されたい．言語モデルの表現力と柔軟性の高さは，現在までに適用されている問題以外に対しても適用できる可能性を秘めており，今後の発展を注視したい．

参 考 文 献

1) J.M. Ponte and W.B. Croft, “A language modeling approach to information retrieval,” Proc.SIGIR 1998, pp.275–281, 1998.
2) G. Salton, A. Wang, and C.S. Yang, “A vector space model for automatic indexing,” Communications of the ACM, vol.18, no.11, pp.613–620, 1975.
3) G. Salton, Automatic text processing: The transformation, analysis, and retrieval of information by computer, Addison-Wesley Longman Publishing, 1989.
4) S. Robertson and K.S. Jones, “Relevance weighting of search terms,” Journal of the American Society for Information Science, vol.27, no.3, pp.129–146, 1976.
5) D. Hiemstra, “A linguistically motivated probabilistic model of information retrieval,” Research and Advanced Technology for Digital Libraries, vol.1513, pp.569–584, Lecture Notes in Computer Science, Springer-Verlag, 1998.
6) F. Song and W.B. Croft, “A general language model for information retrieval,” Proc.CIKM 1999, pp.316–321, 1999.
7) J. Lafferty and C. Zhai, “Document language models, query models, and risk minimization for information retrieval,” Proc.SIGIR 2001, pp.111–119, 2001.
8) S.L. Feng, R. Manmatha, and V. Lavrenko, “Multiple Bernoulli relevance models for image and video annotation,” Proc.CVPR 2004, vol.II, pp.1002–1009, 2004.
9) C. Zhai and J. Lafferty, “A study of smoothing methods for language models applied to information retrieval,” ACM TOIS, vol.22, no.2, pp.179–214, 2004.
10) F. Jelinek and R.L. Mercer, “Interpolated estimation of markov source parameters from sparse data,” Pattern Recognition in Practice, pp.381–397, 1980.
11) S. Brin and L. Page, “The anatomy of a large-scale hypertexual web search engine,” Computer Networks and ISDN Systems, vol.30, no.1-7, pp.107–117, 1998.
12) W. Kraaij, T. Westerveld, and D. Hiemstra, “The importance of prior probabilities for entry page search,” Proc.SIGIR 2002, pp.27–34, 2002.
13) A. Berger and J. Lafferty, “Information retrieval as statistical translation,” Proc.SIGIR 1999, pp.222–229, 1999.
14) P.F. Brown, *et al.*, “A statistical approach to machine translation,” Computational Linguistics, vol.16, no.2, pp.79–85, 1990.
15) J. Xu, R. Weischedel, and C. Nguyen, “Evaluating a probabilistic model for cross-lingual information retrieval,” Proc.SIGIR 2001, pp.105–110, 2001.
16) A. Berger, *et al.*, “Bridging the lexical chasm: Statistical approaches to answer-finding,” Proc.SIGIR 2000, pp.192–199, 2000.
17) J. Jeon, W.B. Croft, and J.H. Lee, “Finding similar questions in large question and answer archives,” Proc.CIKM 2005, pp.84–90, 2005.
18) T. Hofmann, “Probabilistic latent semantic indexing,” Proc.SIGIR 1999, pp.50–57, 1999.
19) X. Liu and W.B. Croft, “Cluster-based retrieval using language models,” Proc.SIGIR 2004, pp.186–193, 2004.
20) X. Wei and W.B. Croft, “LDA-based document models for ad-hoc retrieval,” Proc.SIGIR 2006, pp.178–185, 2006.
21) D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet allocation,” Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.
22) T.L. Griffiths and M. Steyvers, “Finding scientific topics,” Proc.National Academy of Sciences of the United States of America, vol.101, pp.5228–5235, 2004.
23) Y.W. Teh, D. Newman, and M. Welling, “A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation,” Advances in Neural Information Processing Systems, vol.19, pp.1353–1360, MIT Press, 2007.
24) D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers, “Statistical entity-topic models,” Proc.SIGKDD 2006, pp.680–686, 2006.
25) H. Shiozaki, K. Eguchi, and T. Ohkawa, “Entity network prediction using multi-type topic models,” IEICE Transactions on Information and Systems, vol.E91-D, no.11, pp.2589–2598, 2008.
26) 江口浩二，塩崎仁博，“多型トピックモデルを用いたアノテーション付き文書に対する検索手法，” 信学論，vol.J92-D，no.3，pp.311–320，2009.
27) J. Gao, J.-Y. Nie, G. Wu, and G. Cao, “Dependence language model for information retrieval,” Proc.SIGIR 2004, pp.170–177, 2004.
28) D. Metzler and W.B. Croft, “A markov random field model for term dependencies,” Proc.SIGIR 2005, pp.472–479, 2005.
29) K. Eguchi and W.B. Croft, “Query structuring and expansion with two-stage term dependence for Japanese web retrieval,” Information Retrieval, vol.12, no.3, pp.251–274, 2009.
30) D. Metzler, T. Strohman, Y. Zhou, and W.B. Croft, “Indri at TREC 2005: Terabyte Track,” Proc.TREC 2005, pp.1–7, NIST Special Publication 500-266, 2005.
31) V. Lavrenko and W.B. Croft, “Relevance based language models,” Proc.SIGIR 2001, pp.120–127, 2001.
32) V. Lavrenko, M. Choquette, and W.B. Croft, “Cross-lingual relevance models,” Proc.SIGIR 2002, pp.175–182, 2002.
33) J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” Proc.SIGIR 2003, pp.119–126, 2003.
34) K. Eguchi and V. Lavrenko, “Sentiment retrieval using generative models,” Proc.EMNLP 2006, pp.345–354, 2006.
35) 江口浩二，“情報検索のための確率的言語モデルに関する動向と課題，” 信学論，vol.J93-D，no.3，pp.157–169，2010.