社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－145 (10)
2008－ＦＩ－ 91 (10)
2008／6／20

多重多型トピックモデルを用いたアノテーション付きテキストからのエンティティ検索

江口 浩二† 塩崎 仁博†

†神戸大学大学院工学研究科情報知能学専攻
〒657–8501 神戸市灘区六甲台町 1–1
E-mail: eguchi@port.kobe-u.ac.jp

あらまし 最近，確率的トピックモデルに基づく情報検索手法が提案され，言語モデルの枠組みにおいて潜在的ディリクレ配分法（LDA）またはその変形を用いた実験で良好な結果が報告されている．しかしながら，アノテーション付き文書を検索するタスクに対しては，LDA に基づく手法ではアノテーションによって特定された属性型を直接利用することができない．本稿では，アノテーション付き文書コレクションのための新たなアドホック検索手法を提案する．提案手法は多重多型トピックモデルに基づく．これは，Wikipedia におけるエンティティ，カテゴリラベル，その他の語を典型とする，複数種の単語型を直接扱うことができる．この多重多型トピックモデルをアドホック検索に適用する方法を新たに提案し，Wikipedia を用いたエンティティ検索に関する実験によって提案手法の有効性を示す．

Entity Ranking from Annotated Text Collections using Multitype Topic Models

Koji Eguchi† Hitohiro Shiozaki†

†Kobe University
Department of Computer Science and Systems Engineering
1-1 Rokkoudai, Nada-ku, Kobe, 657-8501, Japan
E-mail: eguchi@port.kobe-u.ac.jp

Abstract Very recently, topic model-based retrieval methods have produced good results using Latent Dirichlet Allocation (LDA) model or its variants in language modeling framework. However, for the task of retrieving annotated documents, LDA-based methods cannot directly make use of multiple attribute types that are specified by the annotations. In this paper, we explore new retrieval methods using a 'multitype topic model' that can directly handle multiple word types, such as annotated entities, category labels and other words that are typically used in Wikipedia. We investigate how to effectively apply the multitype topic model to retrieve documents from an annotated collection, and show the effectiveness of our methods through experiments on entity ranking using a Wikipedia collection.

1 はじめに

近年，確率的トピックモデルのいくつかが情報検索の有効性を改善する目的で応用されている [1, 2]．これには，例えば，確率的潜在意味インデクシング（PLSI）[1] や，潜在的ディリクレ配分法（LDA）[3] に基づく検索モデルがある．これらの手法は新聞記事などの非構造化文書に適用されたが，構造化文書はこれとは異なる性質を持つため，上記のような手法をそのまま適用することはできない．構造化文書の重要な特徴の一つは，複数種の属性型で表現された表現力の高い文書表現であり，Wikipedia の例では，エンティティ名と一般記述，および文書レベルのメタデータで記述されている．このようなアノテーション付き文書に対する場合，前述の PLSI や LDA などのトピックモデルは複数種の単語型を直接扱うことができない．これに対して多重多型トピックモデルは以上に述べたような複数種の単語型を直接扱うことができ，それら属性型の間の依存性を反映したトピックを表現することを可能とする [4]．

本稿では，多重多型トピックモデルに基づく検索モデルを提案し，アノテーション付き文書に対する検索有効性を改善する方法について検討する．さらに，Wikipedia におけるエンティティ検索タスクに対して提案手法の有効性を示す．Wikipedia では，各エンティティはエンティティID が対応付けられた文書として表現されており，各文書はテキスト記述，関連エンティティへのリンク，カテゴリラベルなどで構成される．本稿では，関連エンティティへのリンクは，アンカーテキストに出現するエンティティ名を特定するためだけに用いる．従って，各文書は 3 つの構成要素すなわちエンティティ名と他の一般語，カテゴリラベルからなる．

2 関連研究

確率的トピックモデルでは，文書が複数のトピックの混合分布として，各トピックが単語の分布として表現される [1, 3, 5, 6, 7, 8]．Hofmann は，トピックモデリングの先駆的な研究として確率的潜在意味インデクシング（PLSI: Probabilistic Latent Semantic Indexing）を提案した [1]．Blei らは，PLSI モデルを拡張し，各文書に関するトピックの多項分布にディリクレ事前分布を導入することにより，潜在的ディリクレ配分法（LDA: Latent Dirichlet Allocation）を提案した [3]．これにより，PLSI モデルのもつ過適合問題や新たな文書を生成することができない問題を解消した．LDA に関するグラフィカルモデル表現を図 1 に示し，その生成過程を以下に示す．

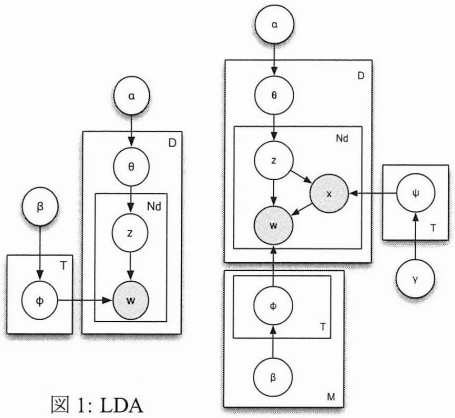

図 1: LDA

図 2: GESwitchLDA

(1)For all d documents sample $\theta_d \sim Dir(\alpha)$

(2)For all t topics sample $\phi_t \sim Dir(\beta)$

(3)For each of the N_d words w_i in document d:

(a)Sample a topic $z_i \sim Mult(\theta_d)$

(b)Sample a word $w_i \sim Mult(\phi_{z_i})$

LDA モデルの推定には，Blei らは変分ベイズ法を用いたが，その後 Teh らは Collapsed 変分ベイズ法を適用することで推定精度を改善した [9]．これら変分ベイズ法やその変形を用いる代わりに，Griffiths らは LDA モデルの推定にギブスサンプリング法を適用した [6]．モデル推定精度という観点からは，十分な繰り返し回数が得られるならば，上述の変分ベイズに基づく手法よりもギブスサンプリングが勝る [9]．

さらに，Newman らは，LDA を拡張し，単語とエンティティを扱うことができる SwitchLDA を提案した [8]．我々はこれを任意個の単語型を扱うことが可能な多重多型トピックモデル GESwitchLDA に一般化し，事前に固有表現認識によってアノテーションが付与された新聞記事を用いて，エンティティ間のリンク予測などのタスクに適用した [4]．GESwitchLDA についての詳細は 3.1 節で述べる．

トピックモデルの重要な応用の一つにアドホック検索[1] が挙げられる．Hofmann は PLSI モデルを提案し，

[1]狭義の情報検索，すなわち，未知のクエリに対して固定的な文書コレクションから検索結果を返すようなタスクを指す [10]．情報フィルタリングに対する概念として定義される．

アドホック検索タスクに適用した [1]．この実験では，推定されたトピックはベクトル空間の基底として用いられ，クエリ・文書間の類似度の計算には，このベクトル空間上のそれぞれのベクトルの内積が用いられた．最近では LDA モデルがアドホック検索に適用されている．Wei と Croft は言語モデルに基づく情報検索の枠組みにおいて，LDA モデルをスムージングの手段として利用した [11]．彼らは，次式に示される通り，従来から用いられてきた文書の言語モデル表現と LDA モデルを用いて，文書に関する線形混合分布を構築した．

$$P(w|d) = \lambda\Big(\frac{N_d}{N_d+\mu}P_{ml}(w|d) + \frac{\mu}{N_d+\mu}P_{ml}(w|coll)\Big) + (1-\lambda)P_{lda}(w|d) \quad (1)$$

ここで，N_d は文書 d における語ののべ総数，μ はスムージング・パラメータを示す．$P_{ml}(w|d)$ と $P_{ml}(w|coll)$ は，それぞれ文書 d と文書コレクション全体における語 w の最尤推定量すなわち相対頻度により求める．$P_{lda}(w|d)$ は次式によって得ることができる．

$$P_{lda}(w|d) = \sum_{t}^{T} P(w|t)P(t|d) \quad (2)$$

T はトピック数を示し，$P(t|d)$ と $P(w|t)$ は次式のギブスサンプリングなどで推定できる [6]．

$$P(t|d) = \frac{C^{TD}_{td,-i}+\alpha}{\sum_t C^{TD}_{td,-i}+T\alpha}$$
$$P(w|t) = \frac{C^{WT}_{wt,-i}+\beta}{\sum_w C^{WT}_{wt,-i}+W\beta} \quad (3)$$

T はトピック数，W は文書コレクションにおける語彙数，α と β はディリクレ事前分布の超パラメータを示す．$C^{TD}_{td,-i}$ は文書 d においてトピック t が割り当てられた頻度（トピック z_i を除く）を示し，$C^{WT}_{wt,-i}$ は語 w にトピック t が割り当てられた頻度（トピック z_i を除く）を示す．

アノテーションなどにより複数種の属性型をともなう表現で記述された文書コレクションに対しては，LDA モデルもそれに基づく検索モデルも直接は適用できない．LDA モデルが異なる型の語を区別しないからである．本稿では，すでに言及した多重多型トピックモデルを用いて，Wikipedia を典型としたアノテーション付き文書に対するアドホック検索モデルを提案する．さらに，単語型を無視して推定した LDA に基づく検索モデルと比較してその有効性を示す．

3 多型文書のための検索モデル

3.1 多重多型トピックモデル GESwitchLDA

Newman らは LDA モデルを拡張していくつかのエンティティ・トピックモデルを提案した [8]．これらはテキストにおいて言及されたエンティティとトピックの間の依存性を表現する試みである．SwitchLDA はここで提案されたモデルの一つである．これらは 2 つの単語型を扱うことができるが，我々は任意個の単語型の間の依存性を表現すべく一般化した GESwitchLDA を開発し，イベントを表現するために who 型エンティティと where 型エンティティとその他の一般語の間の依存関係をモデル化した [4]．このモデルは多重多型トピックモデルとも呼ぶ．

GESwitchLDA のグラフィカルモデル表現を図 2 に示す．個々のトピック t に対して単語型の寄与率を表す多項分布 ψ とそれに対応する超パラメータ γ で決定されるディリクレ事前分布が導入されている．また，トピック t に関する多項分布 ϕ_t とそれに対応する超パラメータ β で決定されるディリクレ事前分布は，種類数 M の個々の単語型ごとに区別される．GESwitchLDA モデルの生成過程は以下の通りである．

(1)For all d documents sample $\theta_d \sim Dir(\alpha)$

(2)For all t topics:

(a)Sample $\psi_t \sim Dir(\gamma)$

(b)For each word type $y \in \{0, \cdots, M-1\}$:

Sample $\phi_t^{(y)} \sim Dir(\beta^{(y)})$

(3)For each of the N_d words w_i in document d:

(a)Sample a topic $z_i \sim Mult(\theta_d)$

(b)Sample a flag $x_i \sim Mult(\psi_{z_i})$

(c)For each word type $y \in \{0, \cdots, M-1\}$:

If $(x_i = y)$ sample a type-y word $w_i \sim Mult(\phi_{z_i}^{(y)})$

GESwitchLDA モデルを推定するにはギブスサンプリングなどを用いることができる [4]．表 1 は，GESwitchLDA を用いて Wikipedia から抽出したトピックの例を示す．

3.2 多型クエリ尤度モデル

アドホック検索タスクの基本的なアプローチの一つに確率的言語モデルに基づくクエリ尤度モデルがある [12, 13, 14]．このモデルにおいては，次式に示される

表 1: Wikipedia から GESwitchLDA を用いて推定した多重多型トピックの例．各列はトピックに対応し，それぞれについて最も尤度の高い語をその尤度とともに上段に，エンティティを中段，カテゴリラベルを下段に示す．

software	0.0266	beer	0.0298	game	0.0551
windows	0.0191	wine	0.0278	player	0.0471
system	0.0171	tea	0.0268	card	0.0401
file	0.0161	drink	0.0218	cards	0.0379
version	0.0122	sugar	0.017	players	0.0288
support	0.0115	coffee	0.0169	play	0.0206
microsoft	0.0114	alcohol	0.0164	games	0.0147
code	0.0106	made	0.0116	played	0.0137
files	0.0095	drinking	0.0107	hand	0.0132
source	0.0093	bottle	0.0106	points	0.0123
Microsoft_Windows	0.0254	Wine	0.0303	Poker	0.0335
Linux	0.0213	Beer	0.0219	Board_game	0.0204
Microsoft	0.0154	Grape	0.017	Card_game	0.0186
Open_source	0.0153	Soft_drink	0.0112	Playing_card	0.0172
Operating_system	0.0143	Coca-Cola	0.011	Betting_(poker)	0.0122
Unix	0.0132	Coffee	0.0106	The_Price_Is_Right	0.0121
Mac_OS_X	0.012	Lager	0.0095	Game	0.0116
GNU_General_Public_License	0.0113	Brewery	0.0095	World_Series_of_Poker	0.0107
Computer_software	0.0102	Alcoholic_beverage	0.0089	Contract_bridge	0.0104
Free_software	0.0079	Vodka	0.0088	Gambling	0.0103
software	0.1259	beverages	0.0857	games	0.1738
computing	0.0483	alcoholic_beverages	0.056	mental-skill_games	0.1171
free_software	0.0324	food_and_drink	0.0403	tabletop_games	0.0993
application_software	0.0296	beer	0.0381	card_games	0.0515
operating_systems	0.0259	alcohol	0.0366	board_games	0.0414
cultural_movements	0.0255	non-alcoholic_drink	0.0359	playing_cards	0.0389
system_software	0.0228	wine	0.0339	entertainment	0.0368
systems	0.0219	distilled_beverages	0.0254	personal_life	0.034
engineering	0.0183	beverage_companies	0.0215	poker	0.0334
software_by_operating_system	0.0179	brewers_and_breweries	0.0207	gambling	0.0275

ように，各文書がクエリ q を生成する尤度の順に文書をランキングする．

$$P(q|d) = \prod_{w \in q} P(w|d)^{c(w,q)} \tag{4}$$

d は文書，q はクエリ，そして w は q における語を示す．$c(w,q)$ は q における語 w の頻度を与える．したがって，$P(q|d)$ はいわゆる bag-of-words すなわち文書において各語が独立であるとする仮定に基づいて，多項分布で表現された文書モデルがクエリ語を生成する尤度を示す．(4) 式による文書ランキングは，次式で表される，文書モデルとクエリモデル間の（負の）クロスエントロピーと等価である．

$$\sum_{w \in q} P(w|q) \log P(w|d) \tag{5}$$

文書モデル $P(w|d)$ を推定するのに，次式のディリクレ・スムージングを用いることができる [15]．

$$P(w|d) = \frac{N_d}{N_d + \mu} P_{ml}(w|d) + \frac{\mu}{N_d + \mu} P_{ml}(w|coll) \tag{6}$$

ここで，$P_{ml}(w|d)$ は文書 d における語 w の最尤推定量に基づく分布を表し，$P_{ml}(w|coll)$ は文書コレクション全体における語 w の最尤推定量に基づく分布を表す．N_d は文書 d における語の総数，μ はディリクレ・スムージングのパラメータを示す．複数の単語型で表現された文書（以下，多型文書）に適用するには，(5) 式を修正する必要がある．

多型文書を仮定し，(5) 式で示されたクエリ尤度モデルを次式のように修正する．

$$\sum_{x \in \mathbf{x}} \nu_x \sum_{w \in q_x} P(w|x,q) \log P(w|x,d) \quad \text{where} \quad \sum \nu_x = 1 \tag{7}$$

ここで，$\mathbf{x}$ は単語型の集合，x は特定の単語型を示す．重み付けパラメータ ν_x によって，ランキングにおける単語型のバランスを調整することができるが，この値は経験的に定める．

文書における特定の単語型 x のみに着目したモデル $P(w|x,d)$ を，(6) 式のディリクレ・スムージングを修正した次式を用いて推定することができる．

$$P(w|x,d) = \frac{N_{xd}}{N_{xd} + \mu_x} P_{ml}(w|x,d) + \frac{\mu_x}{N_{xd} + \mu_x} P_{ml}(w|x,coll) \tag{8}$$

ここで，N_{xd} は文書 d における単語型 x を伴う語ののべ総数を示す．また，$P_{ml}(w|x,\cdot)$ は単語型 x を伴う語 w の最尤推定量に基づく分布を示す．

3.3 GESwitchLDA に基づく検索モデル

本稿では，アノテーション付き文書コレクションのための，多重多型トピックモデル（3.1 節）に基づく検索モデルについて，Wikipedia を用いた例で説明する．$x=0$，$x=1$ および $x=2$ はそれぞれ対応する語 w が一般語，エンティティ，カテゴリラベルであることを示す．LDA を用いたアドホック検索 [11] の場合と同様に，GESwitchLDA のみを用いるのは情報検索のための文書表現としては粗すぎる．従って，我々は GESwitchLDA と 3.2 節で述べた多型クエリ尤度モデルにおける文書モデルを用いて，次の 2 通りの方法でアドホック検索のための新たな文書モデルを構築する．

手法 (a)：$P(w|x,d)$ を用いる方法

手法 (b)：$P(w,x|d)$ を用いる方法

まず，手法 (a) の詳細を述べる．我々は GESwitchLDA モデルと多型クエリ尤度モデルにおける文書モデルを用いて，次式のようにして線形混合モデルを構築する．

$$\begin{aligned}P(w|x=0,d)=&\\ &\lambda\Big(\frac{N_{wd}}{N_{wd}+\mu_w}P_{ml}(w|x=0,d)\\ &+\frac{\mu_w}{N_{wd}+\mu_w}P_{ml}(w|x=0,coll)\Big)\\ &+(1-\lambda)P_{tm}(w|x=0,d)\end{aligned}\tag{9}$$

$$\begin{aligned}P(w|x=1,d)=&\\ &\lambda\Big(\frac{N_{ed}}{N_{ed}+\mu_e}P_{ml}(w|x=1,d)\\ &+\frac{\mu_e}{N_{ed}+\mu_e}P_{ml}(w|x=1,coll)\Big)\\ &+(1-\lambda)P_{tm}(w|x=1,d)\end{aligned}\tag{10}$$

$$\begin{aligned}P(w|x=2,d)=&\\ &\lambda\Big(\frac{N_{\ell d}}{N_{\ell d}+\mu_\ell}P_{ml}(w|x=2,d)\\ &+\frac{\mu_\ell}{N_{\ell d}+\mu_\ell}P_{ml}(w|x=2,coll)\Big)\\ &+(1-\lambda)P_{tm}(w|x=2,d)\end{aligned}\tag{11}$$

そして，次の値の大きさの順に文書をランキングする．

$$\sum_{x\in\{0,1,2\}}\nu_x\sum_{w\in q_x}P(w|x,q)\log P(w|x,d)\qquad\text{where}\ \sum\nu_x=1\tag{12}$$

次に手法 (b) の詳細を述べる．我々は GESwitchLDA モデルと従来型のクエリ尤度モデルにおける文書モデルを用いて，次式の要領で線形混合モデルを構築する．

$$\begin{aligned}P(w,x=0|d)=&\\ &\lambda\Big(\frac{N_d}{N_d+\mu_w}P_{ml}(w,x=0|d)\\ &+\frac{\mu_w}{N_d+\mu_w}P_{ml}(w,x=0|coll)\Big)\\ &+(1-\lambda)P_{tm}(w,x=0|d)\end{aligned}\tag{13}$$

$$\begin{aligned}P(w,x=1|d)=&\\ &\lambda\Big(\frac{N_d}{N_d+\mu_e}P_{ml}(w,x=1|d)\\ &+\frac{\mu_e}{N_d+\mu_e}P_{ml}(w,x=1|coll)\Big)\\ &+(1-\lambda)P_{tm}(w,x=1|d)\end{aligned}\tag{14}$$

$$\begin{aligned}P(w,x=2|d)=&\\ &\lambda\Big(\frac{N_d}{N_d+\mu_\ell}P_{ml}(w,x=2|d)\\ &+\frac{\mu_\ell}{N_d+\mu_\ell}P_{ml}(w,x=2|coll)\Big)\\ &+(1-\lambda)P_{tm}(w,x=2|d)\end{aligned}\tag{15}$$

以上に述べた方法 (b) の場合，次の値の大きさの順に文書をランキングする．

$$\sum_{(w,x)\in q}P(w,x|q)\log P(w,x|d)\tag{16}$$

ここで，q は語 w と単語型 x の対からなる集合として表される．

以上に述べた GESwitchLDA に基づく検索モデルの手法 (a) と手法 (b) において，$P_{tm}(w|x,d)$ と $P_{tm}(w,x|d)$ は次のようにして計算する．

$$P_{tm}(w|x,d)=\sum_t^T P(w|x,t)P(t|d)\tag{17}$$

$$\begin{aligned}P_{tm}(w,x|d)&=\sum_t^T P(w,x|t)P(t|d)\\&=\sum_t^T P(w|x,t)P(x|t)P(t|d)\end{aligned}\tag{18}$$

$P(t|d)$，$P(w|x,t)$ および $P(x|t)$ は次のようにギブスサンプリングを用いて推定する [8, 4].

$$\begin{aligned}P(t|d)&=\frac{C^{TD}_{td,-i}+\alpha}{\sum_t C^{TD}_{td,-i}+T\alpha}\\ P(w|x=0,t)&=\frac{C^{WT}_{wt,-i}+\beta^{(0)}}{\sum_w C^{WT}_{wt,-i}+W\beta^{(0)}}\\ P(w_e|x=1,t)&=\frac{C^{ET}_{et,-i}+\beta^{(1)}}{\sum_e C^{ET}_{et,-i}+E\beta^{(1)}}\end{aligned}$$

$$P(w_\ell|x=2,t) = \frac{C^{LT}_{\ell t,-i}+\beta^{(2)}}{\sum_\ell C^{LT}_{\ell t,-i}+L\beta^{(2)}}$$

$$P(x|t) = \frac{n^x_{t,-i}+\gamma}{n^{all}_{t,-i}+3\gamma}$$

ここで，$n^x_t=\sum_{w_x} C^{W_xT}_{w_xt}$，$n^{all}_t=\sum_x n^x_t$ であり，$C^{\cdot\cdot}_{\cdot\cdot}$ の表記は (3) 式と同様である．T と α についても (3) 式と同様であるが，β と γ は 3.1 節で言及した定義に従う．単語型は一般語，エンティティ，カテゴリラベルを想定してそれぞれ w, e, ℓ あるいは w_x ($x \in \{0,1,2\}$) で示し，これらの文書コレクションにおける異なり語数をそれぞれ W，E，L で示す．

4 実験

4.1 タスク定義と評価尺度

Wikipedia では，エンティティはその定義等を説明する文書として表現される．言い換えれば，各文書は特定のエンティティに対応するので，Wikipedia におけるエンティティ検索タスクは適合性に基づく文書検索に，ある程度類似する．ここで，エンティティ検索と文書検索の主な相違点は，前者の適合性では特定のエンティティに関して定義し説明することが要求されるのに対して，後者では必ずしもそうでないことである．例えば，あるエンティティに係る一般的な情報について説明するものの，そのエンティティの定義を述べていない文書は，文書検索タスクでは適合とされるであろうが，エンティティ検索タスクでは不適合とされる．

我々は評価ワークショップ「INEX-2007 Entity Ranking Track」[2] で構築された 28 の訓練用評価データと 46 のテスト用評価データを用いた．それぞれの評価データは，トピックと対応する適合判定からなる．クエリはトピックの title フィールドから抽出して用いた．この評価ワークショップで用いられた Wikipedia コレクションは英語で記述された 659,388 件からなり [16]，本稿でもこれを用いた．

評価尺度としては，平均精度（MAP: mean average precision —non-interpolated）[10]，幾何平均精度（GMAP: geometric mean average precision）[17] および MRR（mean reciprocal rank）[18] を用いた．MAP は情報検索の評価指標として広く受け入れられており，安定的かつ理解しやすいことで知られている．MAP が算術平均を用いるのに対して，GMAP は幾何平均を用いることによって得られ，それにより頑健な（検索の難易度が高いクエリに対しても比較的有効な）検索システムが重視される．MRR は，最も上位にランキングされた適合エンティティのランクの逆数に基づく指標であり，質問応答タスクなどの評価でしばしば用いられる．

[2] ⟨http://inex.is.informatik.uni-duisburg.de/2007/xmlSearch.html⟩.

4.2 実験設定

Wikipedia コンテンツを構成する 3 つの要素すなわちエンティティ名と他の一般語，カテゴリラベルは，それぞれ異なる名前空間で排他的に扱った．また，418 のストップワード [19] を除去し，10 文書未満にしか出現しない一般語も除去した．ただし，エンティティとカテゴリラベルについてはその出現頻度を問わず除去しなかった．トピック数は $T=400$ または $T=800$ とした．GESwitchLDA を推定するため，2 つの独立したマルコフ連鎖に対してギブスサンプリングを実行し，それぞれで推定されたトピックを貪欲法によって対応付け，$P(w,x|t)$ を平均した．$P(t|d)$ についても同じ要領で 2 つのマルコフ連鎖の平均によって得た．また，比較のために用いた LDA モデルについても基本的に同じ要領で推定したが，単語型 x は無視した．

以下では，訓練データを用いて経験的に定めた各種パラメータの値について述べる．前節で述べた MAP を最大化するように決定した．従来型のクエリ尤度モデル（以下，「QL」）で用いた (6) 式におけるディリクレ・スムージングのパラメータは，$\mu=250$ に設定した．また，多型クエリ尤度モデル（以下，「MQL」）で用いた (8)-(11) 式および (13)-(15) 式におけるディリクレ・スムージングのパラメータは，$\mu_w=\mu_e=\mu_\ell=50$ とした．(9)–(11) 式による GESwitchLDA に基づく検索モデルの手法 (a)（以下，「GESI+MQL」）においては，$T=400$，$T=800$ に対してそれぞれ $\lambda=0.6$，$\lambda=0.5$ とした．また，(13)–(15) 式による手法 (b)（以下，「GESD+QL」）においては，$T=400$，$T=800$ に対してそれぞれ $\lambda=0.6$，$\lambda=0.5$ とした．(1) 式の LDA に基づく検索モデル（以下，「LDA+QL」）に関しては，$T=400$，$T=800$ に対してそれぞれ $\lambda=0.7$，$\lambda=0.5$ とした．

なお，$\lambda=0$ のとき，上で述べた LDA+QL において，クエリ尤度モデルに関する部分を無視し，LDA に基づく文書モデルのみを用いることになるが，次節の実験結果ではこれを「LDA」で示すことにする．同様に，$\lambda=0$ のときの，手法 (a) の GESwitchLDA のみに基づく文書モデルを「GESI」，手法 (b) の GESwitchLDA のみに基づく文書モデルを「GESD」で示す．

表 2: 最適パラメータによる訓練データとテストデータを用いた評価結果

	MAP	GMAP	MRR
training			
QL	0.2267	0.0644	0.4892
MQL (1:1:2)	0.2406	0.0645	0.5140
LDA+QL (T=800)	0.2636	0.1004	0.5229
GESD+QL (T=800)	0.2644	0.0946	0.5458
GESI+MQL (T=800, 2:2:3)	0.2866	0.1198	0.5654
testing			
QL	0.2193	0.1056	0.5115
MQL (1:1:2)	0.2298	0.1143	0.5448
LDA+QL (T=800)	0.2633	0.1366	0.5045
GESD+QL (T=800)	0.2623	0.1313	0.5155
GESI+MQL (T=800, 2:2:3)	0.2749	0.1464	0.5580

表 3: トピック数のみを変化させたときの訓練データを用いた評価結果

	MAP	GMAP	MRR
LDA (T=400)	0.0933	0.0154	0.2549
LDA (T=800)	0.1309	0.0256	0.2574
LDA+QL (T=400)	0.2617	0.1025	0.5607
LDA+QL (T=800)	0.2636	0.1004	0.5229
GESD (T=400)	0.0723	0.0124	0.1340
GESD (T=800)	0.1254	0.0213	0.2511
GESI (T=400, 1:1:1)	0.0789	0.0157	0.1724
GESI (T=800, 1:1:1)	0.1281	0.0243	0.2657
GESD+QL (T=400)	0.2497	0.0965	0.5275
GESD+QL (T=800)	0.2644	0.0946	0.5458
GESI+MQL (T=400, 1:1:1)	0.2649	0.1163	0.5305
GESI+MQL (T=800, 1:1:1)	0.2751	0.1146	0.5578

4.3 実験結果

4.2 節で述べたように，提案する MQL，GESI+MQL および GESD+QL のパラメータと，比較対象の QL と LDA+QL のパラメータは，訓練データを用いて MAP を最大化するよう，経験的に定めた．これらの最適パラメータを用いてテストデータを対して実験を行い，4.1 節で述べた MAP，GMAP および MRR で評価を行った．訓練データとテストデータを用いた評価結果を表 2 に示す．この表から，提案する GESI+MQL（手法 (a)）は QL と比較して MAP で 25.3%，GMAP で 38.6% の改善を得たことがわかる．最近の技術である LDA+QL と比較した場合，我々の GESI+MQL は MAP で 4.4%，GMAP で 7.1% の改善を得た．さらに，Wilcoxon の符号付き順位検定（両側）によると，GESI+MQL は 0.05 の有意水準で，QL と比較した場合，LDA+QL と比較した場合の両方で統計的に有意であった．GESD+QL（手法 (b)）については，QL と比較した場合は 0.05 の水準で有意であったが，LDA+QL と比較した場合は有意差は認められなかった．

表 4: 単語型重みのみを変化させたときの訓練データを用いた評価結果（T=800）

	MAP	GMAP	MRR
MQL (1:1:1)	0.2202	0.0630	0.4889
MQL (1:1:2)	0.2406	0.0645	0.5140
MQL (1:2:1)	0.2007	0.0598	0.4762
MQL (2:1:1)	0.1768	0.0479	0.4566
MQL (1:1:3)	0.2397	0.0601	0.5098
MQL (2:2:3)	0.2374	0.0648	0.4925
GESI (1:1:1)	0.1281	0.0243	0.2657
GESI (1:1:2)	0.1139	0.0188	0.2176
GESI (1:2:1)	0.1273	0.0204	0.2578
GESI (2:1:1)	0.1303	0.0248	0.3045
GESI (1:1:3)	0.0987	0.0152	0.1839
GESI (2:2:3)	0.1207	0.0217	0.2334
GESI+MQL (1:1:1)	0.2751	0.1146	0.5578
GESI+MQL (1:1:2)	0.2864	0.1168	0.5694
GESI+MQL (1:2:1)	0.2615	0.1025	0.5342
GESI+MQL (2:1:1)	0.2316	0.0874	0.4280
GESI+MQL (1:1:3)	0.2830	0.1135	0.5992
GESI+MQL (2:2:3)	0.2866	0.1198	0.5654

表 5: カテゴリ型重みを 0 に設定したときの訓練データを用いた評価結果（T=800）

	MAP	GMAP	MRR
MQL (1:1:0)	0.1046	0.0247	0.2855
GESI (1:1:0)	0.0933	0.0135	0.2328
GESI+MQL (1:1:0)	0.1530	0.0438	0.3429

表 3 に，単語型の重みを固定してトピック数を 400 から 800 に変化させた場合の実験結果を示す．すべてのトピックモデルに基づく手法の有効性は MAP と MRR に関して概ね改善した．ただし，これは計算コストを代償とする．

表 4 は，単語型を変化させたときの MQL，GESI および GESI+MQL の実験結果である．表 5 は，カテゴリラベルに対する単語型重みを 0 に設定したときの，MQL，GESI および GESI+MQL の実験結果である．なお，表中の「ν_0:ν_1:ν_2」は一般語，エンティティ，カテゴリラベルについてのそれぞれの単語型重みの比を示す．これら 2 つの表から，MQL はカテゴリラベルの単語型重みを増やしたときに改善したが，GESI は一般語の単語型重みを増やしたときに改善したことがわかる．最終的な GESI+MQL については，$\nu_0 : \nu_1 : \nu_2 = 2 : 2 : 3$ のときの有効性が最大であった．表 5 においてカテゴリラベルの単語型重みが 0 のときに有効性が著しく低下したことから，Wikipedia のエンティティ検索タスクにおいてはカテゴリデータが重要な役割を担うことが

わかる．

上記のパラメータを調整した実験のすべて（表3–5）で訓練データを用いたことに注意せよ．これは評価の妥当性を確保するため，訓練データのみを用いて最適パラメータを経験的に決定し，テストデータではそれらのパラメータをそのまま用いたからである．

5 おわりに

確率的トピックモデルに基づいて，アノテーション付き文書のための新たな検索モデルを提案した．提案する検索モデルは，拡張したクエリ尤度モデルと多重多型トピックモデルを組み合わせたものであり，この多重多型トピックモデルは異なる単語型を直接扱うことができるものである．多重多型トピックモデルの推定にはギブスサンプリング法を用いた．Wikipediaのエンティティ検索タスクに関する実験によって，クエリ尤度モデルならびにWeiとCroftによって最近開発されたLDAに基づく検索モデルを比較対象として，我々の検索モデルを評価した結果，統計的に有意な改善を得た．

今後の課題としては，Wikipediaにおけるリンク構造の情報を利用することなどが挙げられる．

謝辞

本研究の一部は，科学研究費補助金特定領域研究「情報爆発IT基盤」（19024055），基盤研究(B)（20300038），萌芽研究（18650057）の援助による．

参考文献

[1]Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proceedings of the 22nd Anuual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, California, USA, pp. 50–57 (1999).

[2]Liu, X. and Croft, W. B.: Cluster-based retrieval using language models, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, pp. 186–193 (2004).

[3]Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).

[4]Shiozaki, H., Koji, E. and Ohkawa, T.: Entity network prediction using multitype topic models, *Proceedings of the 12th Pacific-Asia Conference on Knowlede Discovery and Data Mining (PAKDD 2008)*, Osaka, Japan (2008).

[5]Ueda, N. and Saito, K.: Parametric Mixture Models for Multi-labeled Text, *Advances in Neural Information Processing Systems*, Vol. 15, pp. 721–728 (2003).

[6]Griffiths, T. L. and Steyvers, M.: Finding scientific topics, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, pp. 5228–5235 (2004).

[7]Steyvers, M. and Griffiths, T.: *Handbook of Latent Semantic Analysis*, Lawrence Erbaum Associates, chapter 21: Probabilistic Topic Models (2007).

[8]Newman, D., Chemudugunta, C., Smyth, P. and Steyvers, M.: Statistical Entity-Topic Models, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, Pennsylvania, USA, pp. 680–686 (2006).

[9]Teh, Y. W., Newman, D. and Welling, M.: A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation, *Advances in Neural Information Processing Systems*, Vol. 19, pp. 1353–1360 (2007).

[10]Baeza-Yates, R.(ed.): *Modern Information Retrieval*, Addison-Wesley (1999).

[11]Wei, X. and Croft, W. B.: LDA-based document models for ad-hoc retrieval, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, ACM, pp. 178–185 (2006).

[12]Ponte, J. M. and Croft, W. B.: A Language Modeling Approach to Information Retrieval, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 275–281 (1998).

[13]Hiemstra, D.: A Linguistically Motivated Probabilistic Model of Information Retrieval, *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, Vol. 1513, Springer-Verlag, pp. 569–584 (1998).

[14]Song, F. and Croft, W. B.: A general language model for information retrieval, *Proceedings of the 8th International Conference on Information and Knowledge Management*, Kansas City, Missouri, USA, pp. 316–321 (1999).

[15]Zhai, C. and Lafferty, J.: A study of smoothing methods for language models applied to Ad Hoc information retrieval, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, USA, pp. 334–342 (2001).

[16]Denoyer, L. and Gallinari, P.: The Wikipedia XML Corpus, *ACM SIGIR Forum*, Vol. 40, pp. 64–68 (2006).

[17]Robertson, S.: On GMAP: and other transformations, *Proceedings of the 15th ACM Conference on Information and Knowledge Management*, Arlington, Virginia, USA, pp. 78–83 (2006).

[18]Voorhees, E.: The TREC-8 Question Answering Track Report, *Proceedings of the 8th Text REtrieval Conference (TREC-8)* NIST Special Publication 500-246, pp. 77–82 (1999).

[19]Callan, J. P., Croft, W. B. and Harding, S. M.: The INQUERY Retrieval System, *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, Valencia, Spain, pp. 78–83 (1992).