

# 多型トピックモデルを用いた Wikipedia 検索

## Wikipedia Retrieval using Multitype Topic Models

江口 浩二<sup>1\*</sup> 塩崎 仁博<sup>1</sup>  
Koji Eguchi<sup>1</sup> Hitohiro Shiozaki<sup>1</sup>

<sup>1</sup> 神戸大学大学院工学研究科情報知能学専攻

<sup>1</sup> Department of Computer Science and Systems Engineering, Kobe University

**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 a new retrieval method 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 a type-annotated collection, and then show that our proposed method significantly outperforms several state-of-the-art methods through experiments in the task of 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。これには、確率的潜在意味インデクシング (PLSI) [Hofmann 99] や、潜在的ディリクレ配分法 (LDA) [Blei 03] に基づく検索モデル [Wei 06] などがある。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、構造化文書はそ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、典型的な例に Wikipedia が挙げられる。Wikipedia では、各文書は項目解説と関連する他の事典項目（以下、エンティティ）へのリンク、および文書レベルのメタデータなどで記述されている。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捉えることを可能とする [Shiozaki 08b]。

本論文では、多型トピックモデルに基づく検索モデ

ルを新たに提案し、アノテーション付き文書に対して有効な検索を実現する方法について検討する。さらに、検索質問に対して Wikipedia から関連する項目を検索するタスク（以下、エンティティ検索）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（事典項目）はエンティティ ID が対応付けられた文書（項目解説）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。本論文では、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と考える。

## 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 [Hofmann 99, Blei 03, Ueda 03, Griffiths 04, Newman 06, Shiozaki 08b]。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 [Hofmann 99]。Blei らは、PLSI モデルを拡張し、各文書 ( $d$ ) に関するトピック空間の多項分布 ( $Mult(\theta_d)$ ) にディリクレ事前分布 ( $Dir(\alpha)$ ) を

\*連絡先：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

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 [Blei 03] . 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 のグラフィカルモデル表現を図 1(a) に示す . ここでは、各トピック ( $t$ ) に関する単語空間の多項分布 ( $Mult(\phi_t)$ ) にもディリクレ事前分布 ( $Dir(\beta)$ ) が導入されている . なお、グラフィカルモデルとは、確率変数またはパラメータを頂点とし、それらの依存関係関係を有向グラフで表現したものである . 網掛けの頂点は観測変数、それ以外の頂点は潜在変数または未知パラメータを示す . 矩形部分は、その隅に示された回数だけサンプリングが繰り返されることを表す . ただし、 $N_d$  は文書  $d$  の延べ語数、 $T$  はトピック数、 $D$  は文書数を示す . また、図 1(a) に示された LDA のグラフィカルモデル表現に対応する文書生成過程を書き下すと、以下ようになる .

1. すべての文書  $d$  に対してディリクレ事前分布  $Dir(\alpha)$  から  $\theta_d$  をサンプリングする .
2. すべてのトピック  $t$  に対してディリクレ事前分布  $Dir(\beta)$  から  $\phi_t$  をサンプリングする .
3. 文書  $d$  における  $N_d$  語の単語  $w_i$  それぞれに対して、
  - (3-a) 多項分布  $Mult(\theta_d)$  からトピック  $z_i$  をサンプリングする .
  - (3-b) 多項分布  $Mult(\phi_{z_i})$  から単語  $w_i$  をサンプリングする .

この文書生成過程では、潜在変数や未知パラメータを仮定し、グラフィカルモデルで示されたような依存関係に従って、最終的に単語が生成された結果が観測される文書であると考えられる [Steyvers 04] . 実用的には、文書の単語分布から未知パラメータの推定 (モデル推定) を行う必要がある . LDA モデルの推定に、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 [Teh 07] . 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 [Griffiths 04] . 推定されたモデルの精度という観点からは、十分な繰り返し回数を得られるならば、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法が勝る [Teh 07] .

ところで、トピックモデルの重要な応用の一つにアドホック検索<sup>1</sup>が挙げられる . すでに言及した Hofmann の PLSI モデルはアドホック検索タスクに適用された [Hofmann 99] . ここではいくつかの試みがなされたが、その一つとして、推定されたトピックがベクトル空間

<sup>1</sup> 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 [Baeza-Yates 99] . 情報フィルタリングに対する概念として定義される .

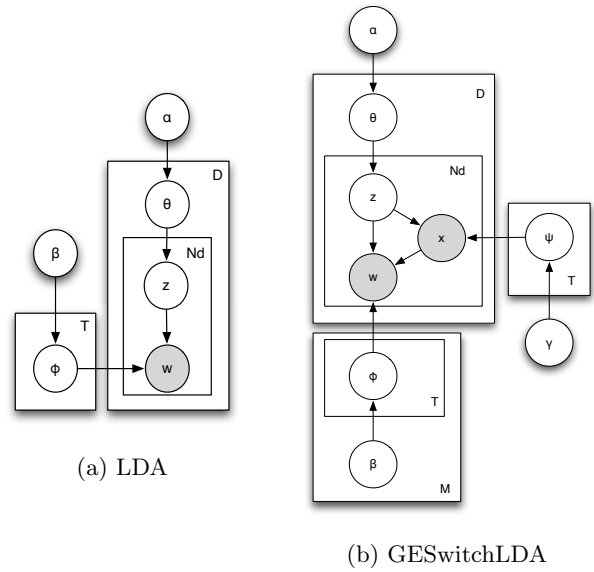


図 1: LDA と GESwitchLDA のグラフィカルモデル表現

の基底として用いられ、クエリ・文書間の類似度の計算に当該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 . さらに最近では 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 [Wei 06] . 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 .

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

ここで、 $N_d$  は文書  $d$  における延べ語数、 $\mu$  はスムージング・パラメータ [Zhai 01] 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 .

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

$P(t|d)$  と  $P(w|t)$  はギブスサンプリング法などで推定できる [Griffiths 04] .

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 . 本論文では、次節で述べる多型トピックモデル GESwitchLDA 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新たに提案する。

### 3 多型文書のための検索モデル

#### 3.1 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 [Newman 06]。これらはテキストにおいて言及されたエンティティ名すなわち固有表現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間の依存性を表現すべく一般化した GESwitchLDA を開発し、ニュースイベントを表現するために文書中に出現する人物名などの who 型エンティティ名と地名などの where 型エンティティ名、さらにその他の一般語の間の依存関係をモデル化した [Shiozaki 08b]。このモデルを多型トピックモデルとも呼ぶ。本節では GESwitchLDA の概要を述べる。

GESwitchLDA のグラフィカルモデル表現を図 1(b) に示す。文書  $d$  に関するトピック空間の多項分布  $Mult(\theta_d)$  とそれに対応するディリクレ事前分布  $Dir(\alpha)$  については LDA の場合と同様であるが、トピック  $t$  に関する単語空間の多項分布  $Mult(\phi_t^{(x)})$  とそれに対応するディリクレ事前分布  $Dir(\beta^{(x)})$  は、個々の単語型  $x$  ごとに区別された。また、個々のトピック  $t$  に対して  $M$  種類の単語型それぞれの寄与率を表す単語型空間の多項分布  $Mult(\psi_t)$  とそれに対応するディリクレ事前分布  $Dir(\gamma)$  が導入された。2 で示した LDA モデルの文書生成過程と同様に、GESwitchLDA モデルの文書生成過程を書き下すと以下ようになる。

1. すべての文書  $d$  に対してディリクレ事前分布  $Dir(\alpha)$  から  $\theta_d$  をサンプリングする。
2. すべてのトピック  $t$  に対して、
  - (2-a) ディリクレ事前分布  $Dir(\gamma)$  から  $\psi_t$  をサンプリングする。
  - (2-b) それぞれの単語型  $y \in \{0, \dots, M-1\}$  に対して、ディリクレ事前分布  $Dir(\beta^{(y)})$  から  $\phi_t^{(y)}$  をサンプリングする。
3. 文書  $d$  における  $N_d$  語の単語  $w_i$  それぞれに対して、
  - (3-a) 多項分布  $Mult(\theta_d)$  からトピック  $z_i$  をサンプリングする。
  - (3-b) 多項分布  $Mult(\psi_{z_i})$  から単語型  $x_i$  をサンプリングする。
  - (3-c) 単語型  $y \in \{0, \dots, M-1\}$  の単語それぞれに対して、 $x_i = y$  のとき単語型  $y$  の単語  $w_i$  を多項分布  $Mult(\phi_{z_i}^{(y)})$  からサンプリングする。

以上に述べた GESwitchLDA モデルを推定するにはギブスサンプリング法などを用いることができる [Shiozaki 08b]。表 1 は、GESwitchLDA を用いて Wikipedia から抽出したトピックの例を示す。3.3 節で提案するアドホック検索モデルは GESwitchLDA に基づくものである。

#### 3.2 多型クエリ尤度モデル

本節では 3.3 節で提案する検索モデルのもう一つの基礎をなすクエリ尤度モデルとその拡張である多型クエリ尤度モデルについて概説する。クエリ尤度モデルは、アドホック検索タスクの基本的なアプローチの一つであり、確率的言語モデルに基づく [Ponte 98, Hiemstra 98, Song 99]。これは文書モデル・クエリモデル間の (負の) クロスエントロピーによる文書ランキングとして解釈される。

$$\sum_{w \in q} P(w|q) \log P(w|d) \quad (3)$$

ただし、 $d$  は文書、 $q$  はクエリ、 $w$  は語を示す。クエリモデルは  $q$  における語  $w$  の最尤推定量、すなわち  $P(w|q) = P_{ml}(w|q) = c(w, q)/|q|$  によって得られる。ここで、 $c(w, q)$  は  $q$  における語  $w$  の頻度を示す。文書モデル  $P(w|d)$  を推定するには次式のディリクレ・スムージングを用いることができる [Zhai 01]。

$$P(w|d) = \frac{N_d}{N_d + \mu} P_{ml}(w|d) + \frac{\mu}{N_d + \mu} P_{ml}(w|coll) \quad (4)$$

ここで、 $P_{ml}(w|d)$  と  $P_{ml}(w|coll)$  はそれぞれ文書  $d$  と文書コレクション  $coll$  における語  $w$  の最尤推定量によって得られる。また、 $N_d$  は文書  $d$  における延べ語数、 $\mu$  はディリクレ・スムージング [Zhai 01] のパラメータを示す。

さて、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、(3) 式を修正する必要がある。次式のように修正する。

$$\sum_{x \in \mathbf{x}} \nu_x \sum_{w \in q_x} P(w|x, q) \log P(w|x, d) \quad (5)$$

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  $\nu_x$  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。 $\sum \nu_x = 1$  であるとする。文書における特定の単語型  $x$  のみに着目したモデル  $P(w|x, d)$  には、(4) 式のディリクレ・スムージングを修正した次

表 1: Wikipedia から GESwitchLDA を用いて推定した多型トピックの例．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一般語をその尤度とともに上段に，エンティティ名を中段，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
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
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

式を用いることができる．

$$P(w|x, d) = \frac{N_{xd}}{N_{xd} + \mu_x} P_{ml}(w|x, d) + \frac{\mu_x}{N_{xd} + \mu_x} P_{ml}(w|x, coll) \quad (6)$$

ここで， $N_{xd}$  は文書  $d$  における単語型  $x$  を伴う語の延べ総数を示す．また， $P_{ml}(w|x, \cdot)$  は単語型  $x$  を伴う語  $w$  の最尤推定量を示し， $\mu_x$  は単語型  $x$  ごとに定めるディリクレ・スムージング・パラメータを示す．

(5) 式および (6) 式は Ogilvie らの手法 [Ogilvie 03] と類似する．ただし，彼らは HTML 文書集合に対する既知事項検索タスクを想定し，いくつかの HTML タグで特定されたテキストの断片に対して適用している．

### 3.3 GESwitchLDA に基づく検索モデル

本節では，アノテーション付き文書コレクションのための，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，関連エンティティ名，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 [Wei 06] 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と考えられる．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もしくはクエリ尤度モデルにおける文書モデルを用いて，以下の要領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．

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．

$$P(w|x = 0, d) = \lambda \left( \frac{N_{wd}}{N_{wd} + \mu_w} P_{ml}(w|x = 0, d) \right.$$

$$\left. + \frac{\mu_w}{N_{wd} + \mu_w} P_{ml}(w|x = 0, coll) \right) + (1 - \lambda) P_{tm}(w|x = 0, d) \quad (7)$$

$$P(w|x = 1, d) = \lambda \left( \frac{N_{ed}}{N_{ed} + \mu_e} P_{ml}(w|x = 1, d) \right. + \frac{\mu_e}{N_{ed} + \mu_e} P_{ml}(w|x = 1, coll) \left. \right) + (1 - \lambda) P_{tm}(w|x = 1, d) \quad (8)$$

$$P(w|x = 2, d) = \lambda \left( \frac{N_{ld}}{N_{ld} + \mu_\ell} P_{ml}(w|x = 2, d) \right. + \frac{\mu_\ell}{N_{ld} + \mu_\ell} P_{ml}(w|x = 2, coll) \left. \right) + (1 - \lambda) P_{tm}(w|x = 2, d) \quad (9)$$

そして，次の値の大きさの順に文書をランキングする．

$$\sum_{x \in \{0,1,2\}} \nu_x \sum_{w \in q_x} P(w|x, q) \log P(w|x, d) \quad (10)$$

ここで， $\sum \nu_x = 1$  とする．

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

$$P_{tm}(w|x, d) = \sum_t P(w|x, t) P(t|d) \quad (11)$$

$P(t|d)$  と  $P(w|x, t)$  は次のようにギブスサンプリングを用いて推定する [Shiozaki 08b, Newman 06, Griffiths 04] ．

$$P(t|d) = \frac{C_{td,-i}^{TD} + \alpha}{\sum_t C_{td,-i}^{TD} + T\alpha} \quad (12)$$

$$P(w|x = 0, t) = \frac{C_{wt,-i}^{WT} + \beta^{(0)}}{\sum_w C_{wt,-i}^{WT} + W\beta^{(0)}} \quad (13)$$

$$P(w_e|x = 1, t) = \frac{C_{et,-i}^{ET} + \beta^{(1)}}{\sum_e C_{et,-i}^{ET} + E\beta^{(1)}} \quad (14)$$

$$P(w_\ell|x = 2, t) = \frac{C_{\ell t,-i}^{LT} + \beta^{(2)}}{\sum_\ell C_{\ell t,-i}^{LT} + L\beta^{(2)}} \quad (15)$$

$$P(x|t) = \frac{n_{t,-i}^x + \gamma}{n_{t,-i}^{all} + 3\gamma} \quad (16)$$

```

<title>Compilers that can compile both C and
C++</title>
...
<categories>
<category id="43172">integrated development environ-
ments </category>
<category id="14030">compilers</category>
</categories>
<description>I want some compilers that can translate
C and C++ source-codes to object-codes. These com-
pilers may also compile some other languages.
</description>
...

```

(a) トピックデータ (抜粋)

- GNU Compiler Collection
- CodeWarrior
- Visual C Plus Plus

(b) 適合エンティティの例

図 2: トピックデータ (抜粋) と適合エンティティの例

ここで,  $n_t^x = \sum_{w_x} C_{w_x t}^{W_x T}$ ,  $n_t^{all} = \sum_x n_t^x$  である. 単語型は一般語, エンティティ名, カテゴリラベルを想定してそれぞれ  $w, e, \ell$  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. また,  $T$  はトピック数を示し, ディリクレ事前分布の超パラメータ  $\alpha, \beta, \gamma$  は 3.1 節で言及した定義に従う.  $C_{td}^{TD}$  は文書  $d$  においてトピック  $t$  が割り当てられた頻度を示し,  $C_t^T$  は「 $\cdot$ 」(一般語, エンティティ名またはカテゴリラベル) にトピック  $t$  が割り当てられた頻度を示す (式中の添字  $-i$  はトピック  $z_i$  を除くことを示す). なお, 以上に説明した手法では (16) 式における  $P(x|t)$  を積極的に用いないが, 別途に提案する手法 [Shiozaki 08a] ではこれを用いている.

## 4 実験

### 4.1 タスク定義と評価尺度

Wikipedia では, ある事典項目に関する項目解説 (文書) において他の関連する事典項目 (エンティティ名) が一般語とともに用いられる. そして, 個々の関連エンティティ名は, その定義やプロフィールなどを解説する別の文書にハイパーリンクによって対応づけられている. つまり, 各エンティティ名は特定の文書に対応するので, 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する定義等を与えてい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. 適合判定方法の詳細は後述する.

我々は評価ワークショップ INEX-2007 Entity Ranking Track<sup>2</sup> [Vries 08] 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, 情報要求が記述されたトピックデータ<sup>3</sup>とそれに対応する適合判定データからなる. トピックデータの例を図 2(a) に示す. title フィールドは利用者の情報ニーズを数語からなるフレーズで表現したもので, description フィールドはそれを 1-2 文程度で表現したものである. また, categories フィールドは, 利用者が求めるエンティティに該当するカテゴリとその ID を示したものである. 本実験ではトピックデータの title フィールドと categories フィールドをクエリとして用いた. なお, title フィールドから抽出した単語集合を分割し, 文書コレクションにおいてエンティティ名として出現するものがあればそれをエンティティ型, それ以外を一般語型とした. 適合判定データは, 行ごとにトピックの ID, エンティティの ID, 2 値の適合判定結果を表し, これをリストアップしたデータである. 各エンティティ ID は文書に対応付けられている. 適合判定データの構築方法としては, 評価ワークショップに参加したグループが提出したエンティティ検索の結果を用いて, プーリングすなわち上位一定件数の和集合をとる過程を経て, プーリング結果に含まれる個々のエンティティ ID に対応する文書の適合性を人手で判定する. このとき, 判定者 (原則としてトピックデータ作成者) は, トピックデータ全体に示された情報ニーズに基づいて個々のエンティティの適合性を判定する. このとき, 多くの場合は, エンティティ名を参照するので十分であるが, 必要に応じてエンティティ名に対応づけられた文書内容を閲覧する. 図 2(a) に示したトピックに対応する適合エンティティ名の例を同図 (b) に示す. 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 [Denoyer 06], 本論文でもこれを用いた. 表 2 に当該コレクションの統計データを示す.

評価尺度としては, 平均精度 (MAP: mean average precision — non-interpolated) [Baeza-Yates 99], 幾何平均精度 (GMAP: geometric mean average precision) [Robertson 06] および MRR (mean reciprocal rank) [Voorhees 99] を用いた. MAP は情報検索の評価指標

<sup>2</sup>(<http://inex.is.informatik.uni-duisburg.de/2007/xmlSearch.html>).

<sup>3</sup>ここでいうトピックデータは, 情報要求を所定の書式で書き下したものである. 3.1 節などで述べたトピックモデルとは異なるものであることに注意せよ.

表 2: Wikipedia コレクションの統計データ

# of documents	659,338
# of vocabulary of words	232,148
# of vocabulary of entities	668,059
# of vocabulary of categories	79,776
# of total frequencies of words	117,329,210
# of total frequencies of entities	17,678,000
# of total frequencies of categories	9,772,936

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おり、次式で求められる。

$$\frac{1}{|q|} \sum_{q \in q} \left[ \frac{1}{|r_q|} \sum_{r \in s_q} prec(r) \right]$$

ここで、 $q$  はクエリ集合、 $s_q$  はクエリ  $q$  に対して検索された適合文書のランクからなる集合、 $r_q$  は適合文書集合、 $prec(r)$  はランク  $r$  における精度 (precision, 適合率とも訳される) を示す。MAP では上式のように各クエリごとの平均精度 (ここでは上式の  $[\cdot]$  の部分を指す) をすべてのクエリにわたって算術平均するのに対して、GMAP は幾何平均を用いることによって得られる。それによりこの尺度では頑健な (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をすべてのクエリにわたって平均するものであり、質問応答タスクなどの評価でしばしば用いられる。

## 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 [Callan 92] 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、エンティティ名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 [Griffiths 04, Wei 06]。  $P(t|d)$  についても同様に 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概要について述べる。従来型のクエリ尤度モデル (以下「QL」) で用いた (4) 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

表 3: 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
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
GESI+MQL ( $T=800, 2:2:3$ )	0.2749	0.1464	0.5580

表 4: トピック数のみを変化させたときの評価結果

	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
GESI ( $T=400, 1:1:1$ )	0.0789	0.0157	0.1724
GESI ( $T=800, 1:1:1$ )	0.1281	0.0243	0.2657
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

た。また、多型クエリ尤度モデル (以下「MQL」) で用いた (6)-(9) 式におけるディリクレ・スムージングのパラメータは、簡単のため、 $\mu_w = \mu_e = \mu_\ell = 50$  に設定した。(7)-(9) 式による GESwitchLDA に基づく検索モデル (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。(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。以上の  $\lambda$  の値は、訓練データを用いて、前節で述べた MAP を最大化するように、経験的に定めた。

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、GESwitchLDA のみに基づく文書モデルを、それぞれ「GESI」で示す。

## 4.3 実験結果

4.2 節で述べたように、MQL および GESI+MQL 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 を最大化するように、経験的に定めた。このようにして決定したパラメータ値を用いてテストデータに対して実験を行い、4.1 節で述べ

表 5: 単語型重みのみを変化させたときの評価結果 ( $T=800$ )。下線は MAP に関して最も有効性が高かったことを示す。

	MAP	GMAP	MRR
(QL)	0.2267	0.0644	0.4892
MQL (1:1:1)	0.2202	0.0630	0.4889
MQL (1:1:2)	<u>0.2406</u>	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
(LDA)	0.1309	0.0256	0.2574
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)	<u>0.1303</u>	0.0248	0.3045
GESI (1:1:3)	0.0987	0.0152	0.1839
GESI (2:2:3)	0.1207	0.0217	0.2334
(LDA+QL)	0.2636	0.1004	0.5229
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)	<u>0.2866</u>	0.1198	0.5654

表 6: カテゴリ型重みを 0 に設定したときの評価結果 ( $T=800$ )

	MAP	GMAP	MRR
(QL)	0.2267	0.0644	0.4892
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

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 3 に示す。なお、表中の「 $\nu_0:\nu_1:\nu_2$ 」の表記は、(5) 式と (10) 式において用いた一般語、エンティティ名、カテゴリラベルの単語型重みの比を示し、訓練データを用いて決定した値が記されている。この表から、提案する GESI+MQL は QL と比較して MAP で 25.3%, GMAP で 38.6% の改善率を得たことがわかる<sup>4</sup>。最近の技術である LDA+QL と比較した場合は MAP で 4.4%, GMAP で 7.1% の改善率を得た。

さて、提案手法と比較対象の評価値の平均を比較するだけでは、その改善がごく少数のトピックについてのみ得られたものであることを否定できないため、有意な改善が実証されたと言えない。そこで、トピックごとの 2 対の評価値 (MAP) に関して、Wilcoxon の符号付き順位検定 (両側) による仮説検定を行い、その差が統計的に有意なものであることを示す。その結

<sup>4</sup>これらの改善率は、提案手法の評価値と比較対象の評価値の差を、比較対象の評価値で除して得た値の百分率として計算したものである。これ以降の改善率についても同様である。

果、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。

次に、訓練データを用いた、より詳細な実験の結果を分析する。ここで訓練データを用いたのは、評価の妥当性を確保するため、テストデータでパラメータの調整を一切行わなかったからである。表 4 は、GESwitchLDA に基づく文書モデル (GESI) を最尤推定量に基づく文書モデルと混合 (+MQL, +QL) することで、LDA と LDA+QL の場合と同様、大きく改善することを示している。また、この表から、単語型の重みを固定してトピック数を 400 から 800 に変化させたときに、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ことがわかる。ただし、これは計算コストを代償とする。

表 5 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。単語型重みが適切であれば、MAP と MRR によれば MQL は QL に勝るが、GMAP によれば両者は同等であった。GESI に関しては、MRR によれば LDA に勝るが、他の 2 つの尺度では両者は同等であった。ところが、GESI と MQL の混合である GESI+MQL では、表 4 に示されている通り、MQL と比較しても LDA+QL と比較しても 3 つの評価尺度で安定した改善が認められた。このことから、GESI と MQL は両者の効果を組み合わせることによってより大きな性能が得ていると考えられる。

表 6 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。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。

## 5 むすび

確率的トピックモデルに基づいて、アノテーションにより異なる属性が付与された語からなる文書のための新たなアドホック検索モデルを提案した。これは多型トピックモデルを用いて、単語型同士の依存性を捉えつつ、文書と語をトピックに応じて非排他的に分割する。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。この多型トピックモデルと多型クエリ尤度モデルを組み合わせることで有効な検索モデルを実現した。これは言い換えると、最尤推定量に基づく多型文書モデルを多型トピックモデルを用いて平滑化する手法であり、これまでに検討されてこなかった新たな観点によるものである。

Wikipedia のエンティティ検索タスクに関する実験によって、広く受け入れられているクエリ尤度モデル

ならびに最近開発された 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。また、多型トピックモデルと多型クエリ尤度モデルを組み合わせることで、それぞれを単独で用いるよりも有効であることを確認した。さらに、当該タスクにおいてカテゴリメタデータが重要であることを示した。より詳細な評価は今後の課題としたい。他の課題としては、Wikipedia におけるリンク構造をモデルに組み入れること、エンティティ検索以外の Wikipedia を用いたタスクに焦点を当てた評価を行う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

- [Baeza-Yates 99] Baeza-Yates, R. and Ribeiro-Neto, B. eds.: *Modern Information Retrieval*, Addison Wesley (1999)
- [Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)
- [Callan 92] Callan, J. P., Croft, W. B., and Harding, S. M.: The INQUERY Retrieval System, in *Proceedings of the 3rd International Conference on Database and Expert Systems Applications*, pp. 78–83, Valencia, Spain (1992)
- [Denoyer 06] Denoyer, L. and Gallinari, P.: The Wikipedia XML Corpus, *ACM SIGIR Forum*, Vol. 40, pp. 64–68 (2006)
- [Griffiths 04] 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)
- [Hiemstra 98] Hiemstra, D.: A Linguistically Motivated Probabilistic Model of Information Retrieval, in *Research and Advanced Technology for Digital Libraries*, Vol. LNCS-1513, pp. 569–584, Springer-Verlag (1998)
- [Hofmann 99] Hofmann, T.: Probabilistic Latent Semantic Indexing, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57, Berkeley, California, USA (1999)
- [Newman 06] Newman, D., Chemudugunta, C., Smyth, P., and Steyvers, M.: Statistical Entity-Topic Models, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 680–686, Philadelphia, Pennsylvania, USA (2006)
- [Ogilvie 03] Ogilvie, P. and Callan, J.: Combining Document Representations for Known-Item Search, in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 143–150, Toronto, Canada (2003)
- [Ponte 98] Ponte, J. M. and Croft, W. B.: A Language Modeling Approach to Information Retrieval, in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–281, Melbourne, Australia (1998)
- [Robertson 06] Robertson, S.: On GMAP: and Other Transformations, in *Proceedings of the 15th ACM Conference on Information and Knowledge Management*, pp. 78–83, Arlington, Virginia, USA (2006)
- [Shiozaki 08a] Shiozaki, H. and Eguchi, K.: Entity Ranking from Annotated Text Collections using Multitype Topic Models, in *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, Dagstuhl Castle, Germany*, Vol. LNCS-4862, pp. 279–292, Springer (2008)
- [Shiozaki 08b] Shiozaki, H., Eguchi, K., and Ohkawa, T.: Entity Network Prediction using Multitype Topic Models, *IEICE Transactions on Information and Systems*, Vol. E91-D, No. 11, pp. 2589–2598 (2008)
- [Song 99] Song, F. and Croft, W. B.: A General Language Model for Information Retrieval, in *Proceedings of the 8th ACM Conference on Information and Knowledge Management*, pp. 316–321, Kansas City, Missouri, USA (1999)
- [Steyvers 04] Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T.: Probabilistic Author-Topic Models for Information Discovery, in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 306–315, Seattle, Washington, USA (2004)
- [Teh 07] Teh, Y. W., Newman, D., and Welling, M.: A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation, in *Advances in Neural Information Processing Systems*, Vol. 19, pp. 1353–1360, MIT Press (2007)
- [Ueda 03] Ueda, N. and Saito, K.: Parametric Mixture Models for Multi-labeled Text, in *Advances in Neural Information Processing Systems*, Vol. 15, pp. 721–728, Cambridge, Massachusetts, USA (2003), MIT Press
- [Voorhees 99] Voorhees, E.: The TREC-8 Question Answering Track Report, in Voorhees, E. and Harman, D. K. eds., *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pp. 77–82, NIST Special Publication 500-246 (1999)
- [Vries 08] Vries, de A. P., Vercoustre, A.-M., Thom, J. A., Craswel, N., and Lalmas, M.: Overview of the INEX 2007 Entity Ranking Track, in *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, Dagstuhl Castle, Germany*, Vol. LNCS-4862, pp. 245–251, Springer (2008)
- [Wei 06] Wei, X. and Croft, W. B.: LDA-Based Document Models for Ad-Hoc Retrieval, in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 178–185, Seattle, Washington, USA (2006)
- [Zhai 01] Zhai, C. and Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval, in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 334–342, New Orleans, Louisiana, USA (2001)