E-016

Kullback Leibler 情報量を用いた情報選別
Information Filtering Using Kullback Leibler Divergence.

柳本 豪一 †
Hidekazu Yanagimoto
大松 繁 †
Sigeru Omatu

1. はじめに

コンピュータやインターネットの普及にともない，膨大な電子化文書が利用できる環境となってきた。このため，膨大な情報の中から必要な情報を取得するため，情報検索技術の研究開発が盛んになっている。特に，利用者が必要とする情報のみを自動的に選別する情報フィルタリングシステムが注目を浴びている。情報フィルタリングシステムでは，利用者の興味をユーザプロファイルとして保存し，その興味に応じて自動的に情報を選別する。したがって，情報フィルタリングの性能は正確に興味を抽出できているかどうかにかかっている。

従来，ユーザプロファイルを作成する手法としては，関連フィードバック [1] がある。関連フィードバックは，ベクトル空間 [2] 上で利用者の検索要求を明確化する手法であり，評価付けされた文書から線形識別を行うためのユーザプロファイルを作成するものである。

一方，従来のベクトル空間上で文書を解析するのではなく，文書を確率モデルで表現することにより，複雑な文書の特徴を捉えようと言う研究が盛んになってきている [3]。確率モデルを用いる際，確率空間上で文書間の類似度を計算する必要があるため，Kullback Leibler Divergence(KL 情報量) が良く用いられる。KL 情報量を用いた研究として，文書をクラスタリングする研究 [4][5] がある。しかし，KL 情報量は非線形関数であるため最適化することが困難であり，利用者の興味を抽出するためにあまり利用されていなかった。

本論文では，遺伝的アルゴリズムを用いることで，KL 情報量を距離関数として，ユーザプロファイルを作成することを提案する。本手法では，遺伝的アルゴリズムで最適化をするために適合度関数の設計を行う。また，本手法の有効性を確認するため，テストコレクション NTCIR2 を用いた評価実験を行う。評価実験より，文書に偏りがある場合に，提案手法が優れていることが確認できた。

2. 提案手法

ここでは，提案手法である KL 情報量を用いた情報選別について説明する。特に，提案手法の主要部分である (1) 文書の確率表現，(2) 単峰性正規分布交叉 (UNDX)，(3) 適合度関数，(4) 遺伝的アルゴリズムの処理，について説明する。

2.1 文書の確率表現

本手法では，すべての文書を離散型確率分布として表現する。そのために，文書を表現する確率分布モデルを決定する必要がある。

ベクトル空間法において，文書は，単語を要素とし，tf*idf 法により求められた重みをその要素の値とする文書ベクトルとして表現される。この場合，単語の出現頻度のみに注目しており，単語の出現順序は考慮されていない。つまり，単語は独立に生成されると仮定している。この仮定から，文書は複数の単語が入っている箱から無作為に単語を抽出して生成されたものとみなすことができる。このように独立試行を繰り返すことによって生成される文書は，多項分布モデルによって記述できる。

文書を多項分布で表現する場合，多項分布における確率変数は単語の出現回数，事象の確率は単語の出現確率に対応する。つまり，多項分布は以下の式で表現され，確率変数 X_i が i 番目の単語の出現頻度，p_i が単語の生成確率を表す。

$$Pr(X_1 = c_1, \cdots, X_n = c_n) = \frac{c!}{\Pi_{i=1}^{n} c_i!} \Pi_{i=1}^{n} p_i^{c_i} \quad (1)$$

ここで，$c = \sum_{i=1}^{n} c_i$ とする。

式 (1) を文書モデルとして利用する場合，実際の文書からパラメータ $(p_1, \cdots, p_n)$ を求める必要がある。ここでは最尤推定によりパラメータを決定する。したがって，各パラメータは以下の式により求められる。

$$p_i = \frac{c_i}{c} \quad (2)$$

以上より，すべての文書を式 (2) より離散型確率分布として表現できる。

2.2 単峰性正規分布交叉 (UNDX)

本手法では，個体をユーザプロファイルそのものとし，遺伝子を実数値ベクトルで表現する。このため，交叉では実数値を扱う必要があるため，UNDX[6] を用いる。UNDX を式 (3)，式 (4) に示す。

$$\mathbf{C}_1 = \mathbf{m} + z_1 \mathbf{e}_1 + \sum_{k=2}^{n} z_k \mathbf{e}_k \quad (3)$$

$$\mathbf{C}_2 = \mathbf{m} - z_1 \mathbf{e}_1 - \sum_{k=2}^{n} z_k \mathbf{e}_k \quad (4)$$

$$\mathbf{m} = \frac{(\mathbf{P}_1 + \mathbf{P}_2)}{2}$$
$$z_1 \sim N(0, \sigma_1^2)$$
$$z_k \sim N(0, \sigma_2^2)(k = 2, \cdots, n)$$
$$\sigma_1 = \alpha d_1$$
$$\sigma_2 = \frac{\beta d_2}{\sqrt{n}}$$
$$\mathbf{e}_1 = \frac{(\mathbf{P}_1 - \mathbf{P}_2)}{|\mathbf{P}_1 - \mathbf{P}_2|}$$
$$\mathbf{e}_i \perp \mathbf{e}_j, (i \neq j), (i, j = 1, \cdots, n)$$

†大阪府立大学大学院 工学研究科

ここで，n は遺伝子が張る空間の次元，$\mathbf{P}_1$，$\mathbf{P}_2$ は両親の遺伝子，$\mathbf{C}_1$，$\mathbf{C}_2$ は交叉によって作成された遺伝子，d_1 は両親間の距離，d_2 は第三の親と両親を結ぶ軸との距離，$\mathbf{e}_1$ は両親を結ぶ軸方向の単位ベクトル，$N(0,\sigma^2)$ は平均0，分散 σ^2 を持つ正規分布乱数を表す。α，β は任意の正の定数とし，$\mathbf{e}_i$ は $\mathbf{e}_1$ に直交した線形独立な単位ベクトルである。ここで，第三の親は正規分布乱数の分散を決定するために用いられる。

2.3 適合度関数

遺伝的アルゴリズムを用いてユーザプロファイルを探索するために，適合度関数を定義する必要がある。本手法では，文書は離散型確率分布として表現されているので，距離関数として KL 情報量を用いる。KL 情報量 $I(Q||P)$ は，式 (5) のように定義され，2つの確率分布間の違いを表す統計情報量である。

$$I(Q||P) = \sum_{k=1}^{n} q_k \log \frac{q_k}{p_k} \tag{5}$$

KL 情報量は，2 つの確率分布が等しいときには，$I(Q||P) = 0$ となり，それ以外の時には，$I(Q||P) > 0$ となる。

今，KL 情報量を文書とユーザプロファイルの類似性を評価する距離関数として利用することを考える。このとき，ユーザプロファイルは興味のある文書群の確率分布に類似していて，興味のない文書群の確率分布と異なっていることが好ましいと思われる。したがって，KL 情報量を用いて適合度関数 $Fit(gene_i)$ を以下のように定義する。

$$Fit(gene_i) = \frac{\sum_{doc_j \in U} I(doc_j||gene_i)}{C * \sum_{doc_k \in I} I(doc_k||gene_i)} \tag{6}$$

ここで，$gene_i$ は i 番目の個体，doc_j は確率分布表現された j 番目の文書，I は興味のある文書集合，U は興味のない文書集合を表す。C は興味のある文書数と興味のない文書数の偏りを補正する重みであり，任意の正の実数である。この適合度関数より，興味のない文書群と異なり，興味のある文書群に似た確率分布を持つユーザプロファイル (個体) が高い適合度を持つこととなる。

2.4 遺伝的アルゴリズムの処理

上記で説明した UNDX と適合度関数を用いてユーザプロファイルの作成を行う。以下に，処理の流れを説明する。

1. 初期集団の生成
   ランダムな値を持つベクトルを遺伝子として，初期集団を構成する。
2. 親の選択
   個体の集団の中から交叉に利用する異なる 3 つの親をランダムに選択する。
3. 子の作成
   選択された親を用いて，UNDX を n_c 回適用し，$2n_c$ 個の子を作成する。このとき，UNDX における正規分布乱数の分散を求めるため，第三の親を利用する。

表 1: 検索タスクの構成

Task	Document	Relevance	Keyword
108	371	64	3,798
109	453	21	4,467
110	624	34	5,490
111	540	40	5,592
114	616	20	6,002
115	931	94	8,758
119	387	20	4,980
121	501	63	4,441
126	279	22	3,695
132	210	22	2,570
138	253	24	3,218
139	386	142	4,561
140	309	23	3,284
147	591	80	6,598

4. 生存個体の選択
   子の生成に利用した 2 個の親の個体と，生成されたすべての子の集団（$2n_c$ 個）から 2 個の個体を選択し，親の個体と置き換える。本手法では，最良個体と最良個体を除いた個体群による重み付きルーレット選択により生存個体を決定する。
5. 終了条件まで (2)〜(4) の処理を繰り返す。

3. 実験と考察

提案手法の有効性を確認するために行った評価実験について説明を行う。

3.1 実験環境

実験には，テストコレクション NTCIR2 の検索タスクを用いた。NTCIR2 では，検索質問ごとに A，B，C 判定が割り当てられた文書群が用意されている。今，判定が付いた文書すべてを実験に用いる文書とし，A，B 判定の文書を正解文書（興味のある文書），C 判定の文書は不正解文書（興味のない文書）とみなして実験データを構築した。実験では，トピックごとに 100 件の文書をランダムに抽出し，ユーザプロファイルを作成するための学習文書とし，残りの文書はユーザプロファイルの評価用文書とした。このとき，学習文書に正解文書が含まれないことを防ぐため，正解文書が 20 件以上含まれている検索タスクを用いることとした。これにより，14 個の検索タスクを評価実験に用いることとなった。実験で用いた検索タスク，全文書数，正解文書数，キーワード数を表 1 に示す。

ユーザプロファイルの作成は，学習文書の構成により影響を受けるものと考えられるので，100 件の学習文書をランダムに抽出する作業を 10 回繰り返し，各検索タスクごとに 10 回の実験を行った。表 2 に，学習文書の平均正解文書数，平均キーワード数を示す。

学習文書に含まれる単語数は，全文書に含まれる単語数に比べて小さいものとなっている。このため，以下の実験では学習文書に含まれている単語のみを用いてユー

表 2: 学習文書の構成

Task	Relevance Ave.	Keyword Ave.
108	19.0	1244.2
109	3.9	1310.8
110	5.0	1223.6
111	7.2	1320.1
114	3.1	1331.0
115	9.6	1368.3
119	4.6	1598.6
121	11.4	1229.1
126	6.8	1547.2
132	11.2	1371.3
138	10.6	1452.5
139	39.0	1428.4
140	7.7	1314.9
147	13.1	1524.3

表 3: 遺伝的アルゴリズムのパラメータ

Population	5,000
Generation	50,000
Crossover	20
α	0.5
β	0.35

ザプロファイルの作成を行い，学習文書に含まれていない単語については，類似度の計算時に用いないものとした。これは，サンプルデータである学習文書に出現しない単語は確率的に発生しないという厳しい条件である。しかし，学習文書に含まれない単語は利用者の評価がなされていない単語であることを考慮して，補正を行うことは好ましくないと考え，この条件を採用した。

3.2 ユーザプロファイルの作成方法

提案手法の有効性を確認するため，提案手法で作成したユーザプロファイと関連フィードバックにより作成したユーザプロファイルを用いて，評価用文書の選別能力を比較する。以下では，各手法の設定について説明を行う。

3.2.1 提案手法

提案手法における遺伝的アルゴリズムのパラメータを表 3 に示す。適合度関数である式 (6) 中の，C の値は以下により計算した値とした。

$$C = 10 * \frac{|U|}{|I|} \tag{7}$$

ここで，$|I|$ は正解文書数，$|U|$ は不正解文書数を表す。表 2 より分かるように，学習文書中に正解文書の割合が小さいため，適合度関数の補正項である C の値を大きいものとした。C の値が小さい場合，不正解文書と異なる確率分布を持つユーザプロファイルが優先的に探索され，正解文書の影響が小さくなりすぎてしまう危険性を防ぐためである。ユーザプロファイルとして用いる個体は，最終世代における最良の個体とした。

3.2.2 関連フィードバック

関連フィードバックは，ベクトル空間法において，ユーザの検索要求の明確化など，情報検索の分野でよく使われる手法である。このため，本論文では文書を離散型確率分布として表現しているので，関連フィードバックを確率分布として表現された文書に，そのまま用いることは好ましくないと思われる。このため，比較手法である関連フィードバックでは，文書を tf*idf 法を用いてベクトル表現した。ただし，文書ベクトルの単語は，確率分布表現で用いていた単語と同一とし，以下の式で重みを決定した。

$$w_j^i = tf_j^i \cdot \log \frac{N}{df_j} \tag{8}$$

w_j^i : i 番目の文書でのキーワード T_j の重み
tf_j^i : i 番目の文書でのキーワード T_j の出現頻度
df_j : キーワード T_j を含む文書数
N : 文書の総数

これより，i 番目の文書 Doc_i は $(w_1^i, w_2^i, \cdots, w_n^i)$ と表現される。

関連フィードバックは，式 (9) で表される。

$$Prof = a * \sum_{Doc_i \in I} Doc_i - b * \sum_{Doc_j \in U} Doc_j \tag{9}$$

I は興味のある文書の集合，U は興味のない文書の集合，Doc_i，Doc_j は文書ベクトル，a，b は任意の正の数を表す。

式 (9) を用いるには，a と b の値を適切な値に決定する必要がある。本実験では a，b を決定するため，学習文書に対して leave-one-out 法を用いて最適な値を決定した。ただし，a，b の値は，$a+b=1.0$ の条件の下で，a を 0 から 1.0 まで変化させ探索を行った。

3.3 実験結果

各手法より得られたユーザプロファイルを用いて評価用文書を選別したときの精度を 11 点平均適合率 [7] により評価する。適合率と再現率は以下の式より求められ，11 点平均適合率は，再現率レベル $0, 0.1, \cdots, 1.0$ の 11 点における適合率の平均値として計算される。

$$\text{適合率} = \frac{\text{検索された文書中の適合文書の数}}{\text{全文書中の適合文書の数}} \tag{10}$$

$$\text{再現率} = \frac{\text{検索された文書中の適合文書の数}}{\text{検索された文書の数}} \tag{11}$$

表 4 に各検索タスクにおける 11 点平均適合率を示す。表中の 11 点平均適合率に下線が付いたものは，5% の有意水準で t 検定を行い，有意差が認められたものを示している。

3.4 考察

表 4 より，8 つの検索タスクにおいて，提案手法の 11 点平均適合率が改善している。また，それらの検索タス

表 4: 11 点平均適合率

Task	Feedback	GA
108	0.281	0.284
109	0.100	0.190
110	0.076	0.110
111	0.188	0.177
114	0.085	0.115
115	0.169	0.216
119	0.087	0.170
121	0.281	0.362
126	0.287	0.183
132	0.208	0.203
138	0.641	0.507
139	0.772	0.653
140	0.192	0.211
147	0.376	0.299

クのうち，4 つの検索タスクにおいて t 検定により有意差が確認できた。一方，2 つの検索タスクにおいて，関連フィードバックが提案手法より優れており，かつ有意差が確認された。今，関連フィードバックが優れている検索タスク 126 と 139 について特に検討する。今，表 2 における平均正解文書数と表 4 の結果を比較すると，正解文書数が少ない状況において，提案手法が良いユーザプロファイルを作成できていることが確認できる。特に検索タスク 139 においては，学習文書に正解文書が平均 39.0 件含まれており，バランスの良い学習データとなっている。したがって，識別境界を構成するためのデータが多いと考えられるので，関連フィードバックにより適切な識別境界を構築できたと考えられる。これより，適合度関数を改良する必要があると考えられる。

また，表 2 における平均単語数と表 4 の結果を比較すると，平均単語数が多い場合にも提案手法と関連フィードバックの差が小さくなっていたり，関連フィードバックが優れていることが多いことが確認できる。平均単語数の増加は探索空間の次元の増加に関係するものであり，遺伝的アルゴリズムでは収束に影響を及ぼすものである。したがって，探索が完了していない可能性が考えられる。これを検討するため，世代数を 100,000 に増加して実験を行ったところ，提案手法の 11 点平均適合率は 0.237 となり 2 手法間で有意差を認めることができなかった。このため，学習文書の構成により，遺伝的アルゴリズムのパラメータを調整する必要があると考えられる。

以上のような問題があると言えども，実際に情報の選別を行う状況を考えると，正解文書が少ない段階で良い選別性能を出す提案手法は有効であると考えられる。

4. おわりに

本論文では，Kullback Leibler Divergence を距離関数として用いて，遺伝的アルゴリズムによりユーザプロファイルを作成する手法を提案した。この際，学習文書中の正解文書数と不正解文書数の偏りを考慮した適合度関数を定義することで，利用者の興味を表したユーザプロファイルが作成できることを確認した。特に，正解文書が少ない状況において，関連フィードバックより精度のよいユーザプロファイルが作成できることが分かった。

今後は，さらなる適合度関数の検討を行い，精度のよりユーザプロファイルを作成することを目指す予定である。

最後に本研究の成果は，文部科学省 科学研究費補助金 奨励研究 (B)(15700104) の助成研究によるものである。記して謝意を表したい。

参考文献

[1] J. J. Rocchio :「Relevance Feedback in Information Retrieval」, SMART Retrieval System Experimens in Automatic Document Processing, Prentice Hall Inc., pp.313–pp.323, (1971)

[2] G. Salton & C. Buckley :「Term-Weighting Approaches in Automatic Text Retrieval」, Readings in Information Retrieval, Morgan Kaufman Publishers, pp.323-328, (1997)

[3] 北研二 :「確率的言語モデル」,, 東京大学出版会 (1999)

[4] I. S. Dhillon & Y. Guan :「Information Theoretic Clustering of Sparse Co-Occurence Data」, Proc. IEEE Conf. on Data Mining, No.3, pp.19-22, Florida, USA (2003-11)

[5] I. S. Dhillon, S. Mallela % R. Kumar :「Enhanced Word Clustering for Hierarchical Text Classification」, Proc. ACM SIGKDO Conf. on Knowledge discovery and data mining, No. 8, pp.191–200, Alberta, Canada (2002)

[6] 小野功, 佐藤浩, 小林重信 :「単峰性正規分布交叉 UNDX を用いた実数値 GA による関数最適化」, 人工知能学会, Vol.14, No.6, pp.1146–pp.1154, (1999)

[7] 北研二, 津田和彦, 獅子堀正幹 :「情報検索アルゴリズム」, 共立出版 (2002)