

Web 検索結果の専門家に対する適合度による ランキング手法について

鶴田雅信, 梅村祥之, 酒井浩之, 増山繁

豊橋技術科学大学 知識情報工学系

1 はじめに

Web 上に存在するさまざまな文書を, 一般の利用者が有効利用するには検索技術 [1] が不可欠であり, 多くの Web 全文検索サービスが存在している. それらのサービスの品質は, そのランキング手法の違いによって決まる. しかし従来の検索結果のランキング手法の多くは, 文書の対象となる読者の専門知識の有無を考慮していない [3]. そのため, そのような手法の検索結果には, 専門知識を持っている利用者にとっては簡単すぎる, もしくは初心者にとっては難しすぎる文書が混在していた. 専門知識を持っている利用者にとっては, 初心者向けの文書はノイズである場合が多い.

この問題を解決するためには, 利用者の知識に応じた文書を選択できるような検索を行うことが有効であると思われる. そこで, 「ある Web 文書がどれぐらい専門家向けであるか」を表す尺度, および, それによって Web 検索結果をランキングする手法を考案する.

ある別の手法によってランキングされた検索結果における上位の数記事を再度ランキングし, ユーザビリティを向上させる事が本手法の目的である. 従って, 実装において, 文書とクエリの類似関係によって検索を行なうシステムを, 本手法に基づいたシステムとは別個に用意する必要がある.

1.1 「専門家向けの文書」の定義

本手法では, ある文書の属する分野において, その分野で使用される専門用語を全て知っている人を「専門家」と定義する.

ある文書のみを読んだ際に, 専門家でなければその内容が理解できない場合, その文書は「専門家向けの文書」である. もしある別の文書のテーマが, ある「専門家向けの文書」と同一であったとしても, その文書のみを読んだ際に, 専門家でなくともその内容が理解できた場合, その文書は「専門家向けでない文書」である. また, 「専門家向け」であるかどうかの程度には文書によって差があり, この程度を「専門家に対する適合度」と定義する.

1.2 実際の文書の特徴

最初に「専門家向けの文書」の定義に従い, 検索結果内の文書に対して, 「専門家向けである」, 「やや専門家向けである」, 「どちらでもない」, 「やや専門家向けではない」, 「専門家向けではない」の 5 段階での評価を人手によって行なった. その結果に基づき, それらの文書の特徴を調査した.

「専門家向けである」文書の多くには, 以下のような特徴が見受けられた.

- 略語やジャーゴン, 専門用語といった「重要語」が多数出現する.

- 含まれている「情報」の量¹が多い.

一方, 「専門家向けではない」文書には, 以下のような特徴が見受けられた.

- 「重要語」も出現するが, その説明には一般的な語が使用される.

- 含まれている「情報」の量が少ない.

また, 「やや専門家向けである」, 「どちらでもない」, 「やや専門家向けではない」文書は, 「専門家向けである」, もしくは「専門家向けではない」文書の特徴を部分的に持っていることも分かった. よって, これらの特徴を数値化できる尺度が存在すれば, 文書の専門家に対する適合度によって検索結果をランキングすることが可能となる.

2 重要語密度

最初に, ある文書の「重要語密度」をその文書が専門家向けであるかどうかの尺度として使用する手法を考案した. この尺度は以前「ある文書が専門的であるか, それとも専門的でないか」を二値分類するために考案した [4] ものであるが, ランキング付け手法としても用いることができる.

¹ 数学的に定義された情報量のことではなく, あくまでも直観的な印象としての「情報」の量を意味する.

2.1 重要語の定義

専門家向けの文書には、専門用語が多数出現する。一方、専門家向けではない文書の多くには、専門家向けの文書ほど多くの専門用語は出現しない。また、多くの専門用語が出現する「専門家向けではない文書」には、専門用語の意味を解説する非常に多くの普通語が出現する。これらの観察より、ある文書中における専門用語の出現密度を、その文書が専門家向けであるかどうかの指標として利用できるのではないかと考えた。

しかし、アドホックな Web 検索処理の実行中に「専門用語」を抽出することは、処理時間の制約もあり困難だと思われる。そこで、以下の点で専門用語に似た性質を持った語、すなわち「検索結果に含まれない文書にはあまり出現しないが、検索結果に含まれる文書中には多く出現する語」(名詞、カタカナ語、およびそれらの複合名詞)を「重要語」と定義し、その密度を使用する手法を考案した。

2.2 重要語密度による手法

重要語密度による手法を以下に示す。

Step 1. 検索結果である文書集合 S に含まれる語 $t_i, i = 1, 2, \dots, n$ を抽出する²。

Step 2. 語 t_i に対して、語の重み $W(t_i, S)$ を計算する。

$$W(t_i, S) = \left(A + \frac{Tf(t_i, S)}{\max_i Tf(t_i, S)} \right) \times \log \frac{|N|}{df(t_i, N)} \quad (1)$$

A : 定数³。

$Tf(t_i, S)$: 文書集合 S における語 t_i の出現頻度

$$Tf(t_i, S) = \sum_{d \in S} tf(t_i, d) \quad (2)$$

$tf(t_i, d)$: 文書 d における語 t_i の出現頻度。

$df(t_i, N)$: 全文書集合 N において、語 t_i を含んでいる文書の頻度。

Step 3. 重み $W(t_i, S)$ がすべての語における重みの平均値 T 以上である語 t_i を重要語として採用する。

$$T = \frac{\sum_{i=1}^n W(t_i, S)}{n} \quad (3)$$

Step 4. 文書 d における重要語密度 $Den(d)$ を求める [2]。

$$Den(d) = \frac{\sum_{t \in KW(d)} W(t, S)}{ad(d)} \quad (4)$$

²形態素解析には Mecab (<http://cl.aist-nara.ac.jp/%7Eetaku-ku/software/mecab/>) を使用した

³予備実験により $A = 0.4$ とした。

$$ad(d) = \frac{\sqrt{\sum_{k=2}^{|KW(d)|} (dist_k)^2}}{|KW(d)| - 1} \quad (5)$$

$KW(d)$: d に出現する重要語の集合。

$dist_k$: k 番目に出現した重要語と $k-1$ 番目に出現した重要語の距離 (語の数)。

Step 5. $Den(d)$ を文書 d のスコアとし、値の大きい順に S に含まれる文書をランク付けする。

2.2.1 重要語密度

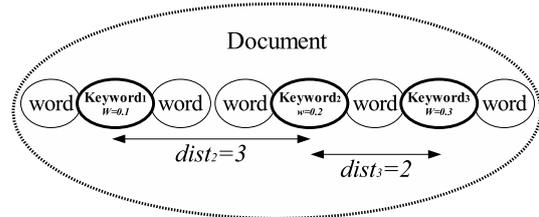


図 1: 重要語密度の例

重要語密度 $Den(d)$ を計算する式として、式 4 を用いる。この式には、式 5 で示した、 d に出現するキーワード間のユークリッド距離 (2 乗平均距離の平方根) を使用している。図 1 の例では、式 5 における $|KW(d)| = 3$, $dist_2 = 3$, $dist_3 = 2$ となり、この時、重要語密度 $Den(d)$ は $\frac{0.1+0.2+0.3}{1.8} = 0.33$ となる。このユークリッド距離が小さいということは、キーワードが文書において密集して出現することを意味する。

2.3 評価実験 1

2.3.1 手法

評価実験は以下の方法で行う。対象となる検索結果は、NTCIR4 WEB タスク D⁴ で使用されたターゲットデータセット (あるクエリによるコーパス中の検索結果) のうち、無作為に選択した 6 個を用いる。また、評価するのは 1 つのクエリにつき、それぞれクエリとの適合度 r における上位 20 件までとする。クエリとの適合度 r は、ターゲットデータセットにおいてあらかじめ与えられている。

正解データ 正解データは、以下の条件で作成する。

- 6 名の評定者 (理工系学生) $v_j, j = 1, 2, \dots, 6$ の評価をもとに作成する。

⁴<http://research.nii.ac.jp/ntcweb/>

- 文書の評価前に、それぞれのクエリが表現するであろう文書集合に対する、評定者 j の有する専門知識の程度 $k_j(S)$ を自己評価させる。 $1 \leq k_j(S) \leq 5$ であり、 $k_j(S)$ が大きいほど、その文書集合に対してより多くの専門知識を有しているとする。
- 文書 d に対して、「専門家向けである (4)」、「やや専門家向けである (3)」、「どちらでもない (2)」、「やや専門家向けではない (1)」、「専門家向けではない (0)」の 5 段階で専門家に対する適合度の評価を行う ($p_j(d)$)。技術的な理由で評価できない文書 (HTML タグの記述ミスでページが表示されない、画像しかなく ALT 要素も記述されていない等) は「評価不能」とする。
- 文書 d が専門家向けであるかどうかの評価値 $a(d)$ は、評定者の評価 $k_j(S)$ の重みつき平均で求める。

$$a(d) = \frac{\sum_{j=1}^{|V|} p_j(d) \cdot k_j(S)}{4 \sum_{j=1}^{|V|} k_j(S)} \quad (6)$$

評価方法 実行結果と正解データの比較を行い、検索結果ごとにそれぞれ評価を行う。評価実験の対象となる尺度を以下に示す。

- $Den(d)$: 文書 d における重要語密度。
- $Sum(d) = \sum_{t \in KW(d)} W(t_i, S)$: 語と語の距離を考慮しない形での、単純な $W(t_i, S)$ の和。これをベースラインとする。
- $p_j(d)$: 評定者 j の評価値
評定者の評価値の個人差を調べるために実験。

それぞれの尺度によって求められた文書 d のスコアを $V(d)$ 、それを 0,1 に正規化した値を $V'(d)$ とする。評価値として、 $V'(d)$ と $a(d)$ の、すべての $d \in S$ における平均誤差 $Ev(S)$ を使用する。この $Ev(S)$ が 0 に近いほど、その尺度は S において専門性の尺度として有効である。

$$Ev(V'(d), S) = \frac{\sum_{d \in S} |a(d) - V'(d)|}{|S|} \quad (7)$$

$$V'(d) = \log \frac{V(d) - \min_d V(d)}{\max_d V(d) - \min_d V(d)} \quad (8)$$

2.3.2 予備実験結果

6 名の評定者によって決定された正解データを用い、予備実験を行なった。 $Ev(S)$ の全ターゲットデータセットにおける平均を表 1 に示す。

表 1: 各尺度における $Ev(V'(d), S)$ の平均

尺度	$Ev(V'(d), S)$ の平均
$Ev(Den(d), S)$	0.33
$Ev(Sum(d), S)$	0.37
$Ev(p_1(d), S)$	0.21
$Ev(p_2(d), S)$	0.25
$Ev(p_3(d), S)$	0.16
$Ev(p_4(d), S)$	0.24
$Ev(p_5(d), S)$	0.24
$Ev(p_6(d), S)$	0.25

2.4 予備実験考察

重要語密度による手法は、ベースラインである $Sum(d)$ よりも良い結果となった。しかし、重要語を用いた手法の $Ev(V'(d), S)$ は、評定者の $Ev(V'(d), S)$ には及ばなかった。

3 重要文密度

重要語密度による上記の手法を踏まえて、ある文書の「重要文密度」をその文書が専門家向けであるかどうかの尺度として使用する手法を考案した。重要文とは、重要語を 2 語以上含む文のこととする。

3.1 重要文密度による手法

重要文密度による手法を以下に示す。

Step 1. 重要語密度による手法の Step 3. までを行ない、文書集合 S に出現する語 t_i の重み $W(t_i, S)$ を求め、重要語を抽出する。

Step 2. 文書集合 S に含まれる文書 d に出現するすべての文 s に対し、文の重要語密度 $Den_w(s)$ を求め、重要語密度がゼロでなかった文を重要文とする。

$$Den_w(s) = \frac{\sum_{t \in KW(s)} W(t_i, S)}{ad_w(s)} \quad (9)$$

$$ad_w(s) = \frac{\sqrt{\sum_{k=2}^{|KW(s)|} (dist_k)^2}}{|KW(s)| - 1} \quad (10)$$

Step 3. 文書 d における重要文密度 $Den_s(d)$ を求める。

$$Den_s(d) = \frac{\sum_{s \in KS(d)} Den_w(s)}{ad_s(d)} \quad (11)$$

$$ad_s(d) = \frac{\sqrt{\sum_{k=2}^{|KS(d)|} (dist_k)^2}}{|KS(d)| - 1} \quad (12)$$

$KS(d)$: d に出現する重要文の集合.
 $dist_k$: k 番目に出現した重要文と $k-1$ 番目に出現した重要文の距離 (文の数).

Step 4. $Den_s(d)$ を文書 d のスコアとし, 値の大きい順に S に含まれる文書をランク付ける.

3.2 評価実験 2

重要語密度の手法の場合と同じ方法で評価実験を行った.

3.2.1 評価実験の対象となる尺度

- $Den_s(d)$: 文書 d における重要文密度.
- $Sum_s(d) = \sum_{s \in KS(d)} Den_w(s)$: 語と語の距離を考慮しない形での, 単純な $Den_w(s)$ の和.

および, 評価実験 1 で評価した尺度.

3.2.2 実験結果

6名の評定者によって決定された正解データを用い, 実験を行なった. その結果を表 2 に示す.

表 2: 各尺度における $Ev(S)$

尺度	平均	$Ev(V'(d), S)$ ターゲットデータセット ID					
		0003	0006	0009	0012	0015	0018
$Ev(Den(d), S)$	0.33	0.22	0.31	0.45	0.36	0.27	0.35
$Ev(Sum(d), S)$	0.37	0.28	0.32	0.53	0.43	0.32	0.35
$Ev(Den_s(d), S)$	0.25	0.23	0.21	0.25	0.21	0.24	0.38
$Ev(Sum_s(d), S)$	0.26	0.23	0.19	0.30	0.22	0.25	0.37
$Ev(p_1(d), S)$	0.21	0.20	0.26	0.22	0.21	0.19	0.20
$Ev(p_2(d), S)$	0.25	0.25	0.26	0.37	0.13	0.23	0.24
$Ev(p_3(d), S)$	0.16	0.14	0.26	0.12	0.14	0.11	0.18
$Ev(p_4(d), S)$	0.24	0.28	0.21	0.17	0.32	0.17	0.27
$Ev(p_5(d), S)$	0.24	0.28	0.23	0.33	0.30	0.14	0.17
$Ev(p_6(d), S)$	0.25	0.17	0.25	0.27	0.18	0.30	0.31

4 考察

重要文密度による尺度の平均 $Ev(V'(d), S)$ は 0.25, 重要語密度による尺度の平均 $Ev(V'(d), S)$ は 0.33 となり, 重要語密度による尺度よりも, 重要文密度による尺度の方が正解データに合致しているという結果となった. 重要語密度はもともと重要文抽出手法に使用される指標 [2] であり, この手法を文書全体に適用するのは妥当ではないと思われる.

個々の評定者による評価と正解データの間にも, 平均 0.225 の誤差が存在していた. これは「専門家に對

する適合度」自体が個人差のあるあいまいなものであるためだと考えられる.

ターゲットデータセット ID: 0018 の検索結果において, 重要文の概念を使用した 2 手法の評価が, 他の検索結果の場合に比べ極端に悪い. この理由として, 正解データの精度が低いことが挙げられた. そこで, ID: 0018 における $k_j(S)$ の平均を調べてみたところ, 4 であった. 他の検索結果では 1.5 から 2.7 の間に収まっていた. これは大半の評定者が, 自分はこの分野に対する専門知識を持っていると判断したためである. この場合, 正解データの精度は高くならなければならない. しかし, 個々の評定者の正解データとの誤差を見るかぎり, 他の検索結果のものとは比べて, この検索結果の正解データの精度が特に高いわけではないと思われる. この理由として, 以下が考えられる.

1. 評定者の想定した分野と, 実際の検索結果の分野がかけ離れていた.
2. 評定者が, 自身が想定していたほどには専門知識を保有していなかった.

従って, 今後の研究では専門知識保有量の自己評価を文書の評価後にも行なう事が必要である.

謝辞

本研究の一部は文部科学省 21 世紀 COE プログラム「インテリジェント ヒューマンセンシング」および文部科学省科学研究費特定領域研究 (B)(2)16092213 の援助により行なわれた.

参考文献

- [1] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.
- [2] H. P. Luhn, The Automatic Creation of Literature Abstracts, IBM Journal, pp. 159-165, 1958.
- [3] C. D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, 1999.
- [4] 鶴田雅信, 梅村祥之, 酒井浩之, 増山繁, 専門性の程度に基づく Web 検索結果の分類手法, 言語処理学会第 9 回年次大会発表論文集 B6-2, 言語処理学会, pp. 725-728, 2004.