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

2004－ＦＩ－ 76　(2)
2004－ＮＬ－163　(2)
2004／9／16

Web からの情報抽出・検索システムにおける全文検索

濱口佳孝† 池野篤司† 井佐原均‡

†沖電気工業株式会社　〒541-0053 大阪府大阪市中央区本町 2-5-7 丸紅ビル４階
‡独立行政法人情報通信研究機構　〒619-0289 京都府相楽郡精華町光台 3-5
E-mail: hamaguti662@oki.com, ikeno546@oki.com, isahara@nict.go.jp

本稿で報告するシステムは、Web を情報ソースとして文書検索を行い、検索された文書から抽出された人名などの情報をランキングして提示する。このシステムの文書検索について、Web ページを情報ソースとすることと、検索結果が情報の抽出・検索に用いられることに着目した開発を行った。本報告の文書検索では評価式のベースとして OKAPI を用い、1)tf の正規化に単語の繰り返し易さとして cf/df を使うことと、2)文書の評価値に記事数を反映するものとして単語の種類数を反映させる試みを行った。これらについて東京大学殿のホームページを対象にした 10 語～100 語の入力文に対する人名検索の精度を評価した結果、長文の入力における精度低下の軽減が認められた。

Web oriented full text search for information retrieval

Yoshitaka HAMAGUCHI† Atsushi IKENO† Hitoshi ISAHARA‡

†OKI Electric Industry Co. Ltd.
4F MarubeniBuilding, 2-5-7, Honmachi, Chuou-ku, Osaka-city, Osaka 541-0053, JAPAN
‡National Institute of Information and Comunications Technology
3-5, Hikaridai, Seikacho, Soraku-gun, Kyoto 619-0289, JAPAN
E-mail: hamaguti662@oki.com, ikeno546@oki.com, isahara@nict.go.jp

Our entire system is a web-oriented search system which list up Named Entities, such as person names, related to the input text. We tried to raise accuracy of this system by improving the full text search method of this system, in view of that the full text search retrieves “web pages” and the NEs are extracted from those “web pages”. In this report, we introduce 1) cf/df to normalize tf, and 2) number of different words in a page as the parameter representing the number of topics in the page. In those two cases, we experimented which searches person names rerated 10 – 100 words query, by using the web pages of The University of Tokyo as information source, and observed that the fall of accuracy with the longer input is smaller.

1. はじめに

検索対象となる文書が多くなるにしたがい文書の検索結果は膨大な量となり、文書中の情報を直接検索するニーズが高まり、各方面で研究が進められている。我々はこれについて、検索したい NE (固有表現) の種類がユーザによって指定されることを前提として、入力文にマッチした種類の NE（「人名」「技術名」など）を Web ページのような非定形文書から検索するシステムの開発をすすめている。これは、論文・特許などを情報ソースとした人名検索による産学連携支援や、技術名検索による発想支援が可能な支援サービスである株式会社三菱総合研究所殿の Bluesilk® [1](http://www.bluesilk.biz/)において、Web ページも横断的に情報ソース

[1] Bluesilk®は株式会社三菱総合研究所殿の登録商標です。

とすることを目指して適用するなどの、具体的な応用を想定している。

この例のような場合、情報ソースの具体的なターゲットは大学のホームページなどとなる。また、専門の異なるユーザが文書中からキーワードを選択する必要が無いインタフェースを実現するためには、キーワードや、ユーザによる質問文として選ばれた語からなる入力以外のものを受け付ける必要がある。これには、Web ページや資料などからコピー・ペーストを行った入力などを想定し、質問の形を成していない通常の文を入力とした場合の検索精度を上げることが効果的と考えている。このような入力には文中での使われ方が大きく異なる単語が含まれるという観点で、この報告では tf (term frequency)の正規化において文書長に加え、各単語の繰り返し易さを考慮した正規化を試みた。

また、文書中から抽出された情報に統計的にスコアリングする観点からは、Web ページのような不揃いな文書群から、より目的に適した文書が検索結果となるようにすることが望ましい。この観点からは、今回は Web ページは１ページ中に複数の記事を含む場合も多いことに着目し、文書中の記事数と相関があると思われる、単語の種類の数による文書の評価値の修正を試みた。

2. 実験概要

我々の検索システムは、文書検索を行い、その検索結果の文書中にあるユーザ指定の NE について統計処理を行い、その NE をランキングして提示する。また、各文書中の情報は NE タガーにより事前に抽出され、データベースに収められる。（図１）。なお、この実験では文書検索エンジンとして、情報処理振興事業協会(IPA)殿が実施した独創的情報技術育成事業の研究成果である、汎用連想検索エンジン GETA を利用させていただいている。

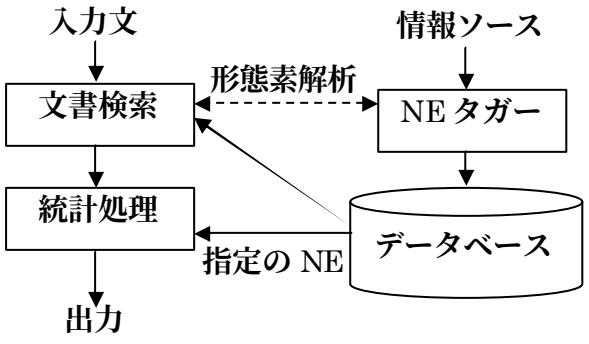

図１

我々は、文書検索の文書評価式には、TREC の Web タスクにおいて精度が高い評価尺度のひとつである Robertson の 2-poison model（OKAPI）をベースとしたものを使用している[1][2]。今回の実験では、文書検索の入力文への適応と統計処理を前提としたスコアリングによる性能向上を目指す目的で、この OKAPI に対して tf の正規化と、文書のスコアリングの修正を加え、比較評価を行った。比較の対象とするベースシステムの評価式は以下の物とした。

$$S(d)=\sum_{t}\left(\frac{TF(d,t)}{k_1 \,l(d)/\Delta + TF(d,t)}\cdot\log\left(\frac{N}{DF(t)}\right)\cdot f(q,t)\right),$$

$$f(q,t)=\frac{TF(q,t)}{k_2+TF(q,t)} \qquad (1)$$

ここで、$TF(d,t)$ はページ d 中の形態素 t の数、$DF(t)$ は全ページ中で形態素 t が出現するページ数、N は全ページ数である。文書長にあたるものとして、Δ は全ページ中の１ページあたりの平均の単語数、$l(d)$ はページに含まれる単語数を用いている。なお、単語としてはこの実験では自立語のみを用いている。この報告中での単語数なども自立語の数となっている。また、k_1 は経験的に決められる値で、後述の評価結果が最良となる 0.7 を用いている。$f(q,t)$ は入力文中の単語 t の数により決まる重みで、経験的に決まる値 $k_2=0.5$ としている。なお、$f(q,t)$ は今回の実験では変更していない。

開発の目的は、文書検索ではなくシステム全体の出力の精度を高めることにある。そのため評価は、東京大学殿の Web ページ(http://www.u-tokyo.ac.jp/)を情報ソースとした人名検索を行った結果得られる人名について、５位以内及び１０位以内に正解の人名が得られた率で評価を行うこととした。

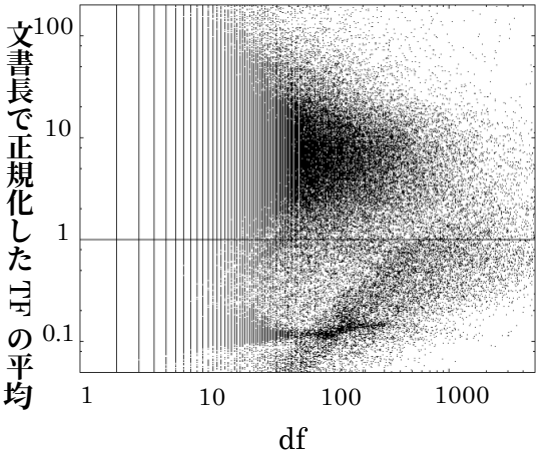

図２

3. 提案手法の説明

3.1. 手法１：単語の平均出現数での正規化

3.1.1. 仮説

Web の記事等の文書からのコピーを入力文とした場合、当然のことながら汎用的な単語などキーワードとして不適当な語も含め、キーワードとしての妥当性が大きく異なる単語が１入力中に含まれる。一般的にはこのような単語の重みの評価には文書の絞込みの観点から IDF 等が使われる。しかし、図２は今回実験に用いた情報ソースの単語ごとにその単語が出現する文書中での出現頻度の平均をプロットしたものを示すが、同じ df（document frequency）を持つ単語であっても１つの文書での使用頻度に大きなバラツキがあることがわかる。これは、Web では特許や論文と比較して平易な表現が繰り返し使われ易く、専門用語などは表題などに使われるに留まることが多いなどの理由が考えられる。

一方、式(1)の$TF(d,t)$に関わる項は、変形すると$\dfrac{TF(d,t)/l(d)}{k_1/\Delta + TF(d,t)/l(d)}$となり、$TF(d,t)$を文書長で正規化している。したがって、単位文書長あたりに出現するキーワードが少ないほどスコアが低くなる。しかしながら、単位文書長あたりのキーワードが少ないことが

1）その文書がキーワードと関連が薄いためにキーワードの出現頻度が少ない。
2）そのキーワードが文書中で繰り返されにくいため、いずれの文書でもキーワードの出現頻度が少ない。

いずれの理由であってもスコアが低くなる。図２に見られるように、同じ df でも文書長で正規化された tf の平均が数桁異なるような場合には単語の繰り返され易さ自体が大幅に違うと考えられる。これは、(2)のような場合も無視できない可能性を示している。このため、繰り返し易さを考慮した正規化に効果があると考えた。すなわち、その単語の文書中での繰り返され易さも用いて単位文書長あたりの単語の出現数の期待値を求め、それで正規化を行うことを考えた。

ただ、Web ページは論文・特許などと比較して短く、平易な表現が用いられ易いこともあり、極端に使われにくい単語も少なくないと考えられる。このことは tf が小さいことに繋がり、場合によっては 1,2 個のものも多くなる。これにより、tf により値を決定する評価式の統計的な安定性に疑問が生じる。このことは、逆に、１文書あたりの出現数が多い単語、すなわち tf が大きくなることが期待できる単語のほうが統計的に安定した評価値に貢献することが期待できる。このため、ＩＤＦに相当する項に、１文書あたりの tf の期待値が高いキーワードの重みを増す補正を加えることを試みた。

3.1.2. 提案手法

単語の繰り返し易さを考慮した場合、単語tに関する長さ$l(d)$の文書で期待される tf は、tに関する文書中での平均出現率に文書長をかけたものとする。これは、ここで、ある単語tに関係する文書の集合を$s(t)$、その平均文書長を$\Lambda(s(t))$、$s(t)$中の単語tの数を$CF(s(t),t)$、$s(t)$に含まれる文書数を$DF(s(t))$とした時、

$$\frac{CF(s(t),t)\cdot l(d)}{DF(s(t))\cdot \Lambda(s(t))} \qquad (2)$$

で表される。

しかしながら$s(t)$は未知であるので、単語tが出現する文書全ての集合である$k(t)$で近似できるものとした。この場合、$k(t)$に含まれる単語tの総数と全文書セット中の単語tの総数は等しいことと、$k(t)$中の文書数がすなわち単語tが出現する文書数であるこ

とを考慮すれば、式(2)は

$$\frac{CF(t)\cdot l(d)}{DF(t)\cdot \Lambda(k(t))} \qquad (3)$$

となる。ここで、$CF(t)$は情報ソースとなる文書セット中での単語tの出現数、$\Lambda(k(t))$は$k(t)$に含まれる文書の１文書あたりの平均長（単語数）である。

さらに、$k(t)$中の１文書あたりの平均長をキーワードごとに得るのは計算コストがかかるため、今回はこれをさらに文書セット中の全文書の平均長Δで近似している。

これでtfを正規化し、式(1)のtfに関わる部分の項を作ると、式(4)のようになる。

$$\frac{TF(d,t)}{k_3 \frac{CF(t)}{DF(t)}\cdot\frac{l(d)}{\Delta}+TF(d,t)} \qquad (4)$$

ここでk_3は後述の評価が最適になるように決定し、0.7を得ている。

また、tfの期待値が高い単語が統計的に安定した評価値に貢献することに期待することについては、cf/tfが高いものがtfが高い可能性が高いため、IDF項にあたるものを以下のようにした。

$$\log\left(\frac{N}{DF(t)}\cdot\left(\frac{CF(t)}{a_1 DF(t)}\right)^{a_2}\right) \qquad (5)$$

なお、これについては、定性的な仮説を作る参考とするため、今回の実験では何乗のオーダーで効くかを探るために、このような式とした。

a_1、a_2はそれぞれ経験的に決定し、それぞれ$a_1=2.0$、$a_2=0.6$を得ている。

この手法の$TF(d,t)$に関わる式(4)は、$CF(t)/DF(t)$が小さい単語の重みを式(1)より重くする方向に働く。反面、IDFに相当する式(5)では$CF(t)/DF(t)$が大きい形態素の重みを式(1)より重くする方向に働く。

このように、双方の項が相反する方向への変更となっている。

将来的には双方の項への$CF(t)/DF(t)$の寄与を一つの項にまとめることが考えられる。

3.2. 手法２：単語の種類数による文書評価

3.2.1. 仮説

Webページには、名簿や研究会のプログラムなど、１ページに複数の記事が含まれていることも多い。

このように複数の記事が含まれるページが検索された場合、そこからの情報の抽出・統計には以下のような影響が考えられる

- より多種類の情報が加わることによる再現率の向上
- 入力文と関係が薄い記事中の情報が加わることによる、適合率の低下
- 不必要に多様な情報が加わることによる、NEの統計処理時のパフォーマンス低下

したがって、文書検索においてページの重みにページ中の記事数を反映させることで、これらの影響の度合いを制御し、最終的な出力であるNEのランキングの精度向上を図ることができると期待できる。

Webページ内の記事数を推定するにはHTMLタグを利用する手法も考えられるが、タグの使用方法は多様であり、また、HTML以外のテキスト文書には適用できないという点で汎用性に欠ける。

このため、より単純なアプローチとして、ページ中の記事数が多いほど話題が多様になり、単語の種類が多様になるという予想に基づき実験を行った。

3.2.2. 提案手法

１ページ中の記事数と相関が予想されるパラメータとして、単語の種類数$I(d)$を全ページの単語の種類数の平均Γで除算した$I(d)/\Gamma$を用いた。そして式(1)を、以下の式(6)で除算したものをページの重みとした。

$$1+b_1\cdot g(d)^{b_2} \qquad (6)$$

$$g(d)=\begin{cases} l(d)/\Gamma & if \quad l(d)/\Gamma > b_3 \\ b_3 & if \quad l(d)/\Gamma \le b_3 \end{cases}$$

b_1、b_2、b_3はそれぞれ実験的に決定し、それぞれ$b_1=0.67$、$b_2=0.16$、$b_3=0.4$を得ている。

この項は、単語種数が何乗で効くかを見積もるために、式(6)のようにして実験を行った。また、記事数が１以下では記事数と単語

種数の相関はなくなるので、b_3で単語種数の下限を設けている。

単語種数は、記事数が変わらず文書長が長くなった場合でもある程度まで増えると考えられ、本来はそれと比較する形で式を作る必要がある。

今後、それも含めたモデル検討及び実験を進めたい。

4. 評価

4.1. 問題設定

入力文及び正解とする人名は、東京大学殿の産学連携テーマデータベース[4]の情報・通信分野より、データ収集時直近の 205 データを用い作成した。

また、情報ソースとしては、東京大学殿の全 Web ページから、この産学連携テーマデータベースの部分を除いた約 36 万ページを用いた。

文書検索はここから上位 50 文書を検索結果としている。人名は、その文書中の人名の出現数及び入力文中の単語との距離を用いて統計的に重み付けしている。

評価１では、各テーマの内容及びタイトルから、主観評価で妥当と思われる単語を最大５個まで選出し、入力とした。

評価２では、各テーマの内容から自立語 10〜40 語を含む主要文を主観評価で抽出し、入力文とした。

評価３では、各テーマのタイトル及び内容すべて(30〜110 語)を入力文とした。

いずれの場合も、正解はそのテーマの担当の方の名前とし、入力文に対して正解が 5 位以内、あるいは 10 位以内に入る率の向上を目指すこととした。

なお、実験的に決定されるパラメータは、評価２での 10 位以内に正解が入る率が最も高くなるようにチューニングした結果を用いている。

4.2. 実験結果

提案手法の、手法 1, 手法 2 及び 手法 1 と 2 双方を用いた場合の人名検索の 10 位正解率と 5 位正解率を、tf の正規化が文書長のみによる式(1)と比較して表１に示す。

このように今回提案手法により、主要文を入力とした評価２と全文を入力とした評価３で人名検索における精度が向上することが確認できた。

評価１，２，３で比較してみると、主観評価で選ばれたキーワードを入力とした場合には精度に有意な違いは見られない。しかし入力が主要文、全文と変わるにつれて式(1)では大きく精度を下げている。これと比較すると、提案手法ではいずれもそれほど大きな精度低下は見られなかった。

4.3. 考察

Web ページを情報ソースとし、質問文として使うことを想定していない、コピー・ペーストされた入力文からの検索精度を上げるという目的に関しては一定の効果があった。

これがキーワード入力と比べて、文章での入力での精度が下がり難いという形で現れたことは、長文の入力文に関する手法１については予想通りであったとも言える。しかしながら、手法２についても同様の傾向が見られる。

この理由は、文書検索の評価式は入力文の全キーワードについての評価値の和を取るので、検索対象のページにそれらのキーワー

入力	評価1：キーワード		評価2：主要文		評価3：全文＋タイトル	
	≦5位	≦10位	≦5位	≦10位	≦5位	≦10位
式(1)	49.3%	57.6%	46.8%	50.7%	37.6%	44.4%
手法1	49.8%	55.6%	47.8%	54.1%	45.4%	54.1%
手法2	49.8%	58.0%	48.8%	52.7%	44.4%	52.7%
手法1+2	49.8%	58.8%	48.8%	58.8%	48.8%	55.6%

表1: 評価結果

ドがより多種入っている場合のほうが、そのページの重み付けが高くなる傾向にあるためと想像できる。すなわち、単語種数が多いページほど有利となると考えられる。これにより、長文の入力文、すなわちキーワードとなる自立語が多い入力に対しては、長文や記事数が多いページが検索されやすくなると予想される。

これに対し、手法２は単語種数が多い文書の重みを減らす効果があり、その傾向を相殺しているものと予想される。これが手法２においても、長文の入力での精度低下が式(1)より小さく収まっている理由ではないかと考えている。

このことから、手法１におけるこの傾向も想定したような正規化の結果ではない可能性を考える必要がありそうである。すなわち、手法２に関する考察同様に、長文におけるキーワードとなる語の多さの問題がある程度解消されていることに起因する可能性も考えられる。つまり、入力文中の単語の重みの強弱が強い評価値となり、結果として長文でも少ない単語のみがスコアに貢献していることの効果である可能性を考える必要がある。

4.4. 課題

我々のシステムは人名などの抽出結果のスコアリング、リストを作ることを目的とするため、今回の評価は人名検索の精度での評価となっている。このため、文書検索自体の精度や傾向の分析はまだ進んでいない。

ただ、この分析を行うためには文書検索の正解となる Web ページを決める必要があり、これが大きな課題となっている。

今後、なんらかの手段で正解ページと正解 NE が共に得られるテストセットを作成し、より詳しい分析と提案手法の裏づけとなるモデルの提案に繋げたい。

また、式(4)と式(5)は、共に cf/df の項を含んでいる。これは文書によらず一定であるので、将来的には式(4)から何らかの近似で外に出し、式(5)と同じ項にまとめたい。同様に cf/df が現れるものとしては、Amati らの Bernouli's normalization[3] があるが、合わせてこれとの比較検証も行いたい。

5. まとめ

本報告では Web ページからの情報のリストアップの前段階に用いられる文書検索に以下の２点の改良を提案した。１つは単語の文書中での繰り返し易さを、tf の正規化と IDF に相当する項の補正を目的に導入すること。もう一つは情報抽出・統計処理に都合がいいページを検索結果とする目的で、ページ自体の重みとしてページ中の単語種数の関数を導入することである。

これらについて実際に Web ページから人名を検索する実験を行った結果、いずれの手法においてもコピー・ペーストを行った文章を入力とした人名検索での精度の向上が認められた。

また、入力文が長文となった場合の精度の低下が少ないことが見出された。

文　献

[1] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M. and Gatford, M. "Okapi at TREC-2", *Proc. Text Retrieval Conference (TREC-2)*. (1993)

[2] 村田 真樹, 馬 青, 内元 清貴, 小作 浩美, 内山 将夫, 井佐原 均, "位置情報と分野情報を用いた情報検索", *自然言語処理 Vol.7, No.2*, pages 141-161 (2000)

[3] Amati, G., Carpineto, C. "FUB at TREC-10 Web Track: A probabilistic framework for topic relevance term weighting", *Proc. Text Retrieval Conference (TREC2001).* (2001)

[4] 東京大学産学連携提案テーマデータベース , *http://www-db.ccr.u-tokyo.ac.jp/*