自然言語処理 115－17
情報学基礎 43－9
（1996. 9. 13）

統計的確率に基づくキーワード重要度算出モデル

松田 透　小川 泰嗣
株式会社リコー　情報通信研究所

単語の頻度情報に基づく従来のキーワード抽出／キーワードランキングでは、$tf*idf$ モデルが主流である。しかし、$tf*idf$ モデルは、頻度以外から得られるキーワード性を反映していないので、確率的に表現したキーワード性や語長による補正を反映できるよう $tf*idf$ モデルの拡張を試みた。そして、実例から得られる文字・品詞・構文レベルの統計的確率に基づいてキーワード性を求め、このモデルを評価した。

A Keyword Weighting Model Based on Statistical Probabilities

MATSUDA Toru, OGAWA Yasushi
Information & Communication R&D Center, RICOH Co., Ltd.

In the extraction and ranking of keywords, the tf*idf model has been the most common approach researchers adopt. This model, however, uses only word frequency and neglects other features relevant to determine whether a word is a keyword. To incorporate these other features, we disigned a probabilistic model as an expantion of the tf*idf model and confirmed its validity by conducting a test using a text corpus.

1　はじめに

単語の頻度情報に基づく従来のキーワード抽出／キーワードランキングでは、文書内頻度（tf）と文書頻度（idf）からキーワード重要度を算出する $tf*idf$ モデル [7] が主流であった [8]。

tf

語の文書内出現回数が本来の tf であるが、この本来の tf に $\log x$ や $\frac{x}{K+x}$ などの単調増加関数による変換を施したものも、文書内の出現の多さを表すものと言える。本稿では、これら同類のものも含めて文書内頻度と呼ぶことにする。また、確率的処理ができるように 0 以上 1 以下の値を取るように正規化した文書内頻度を特に正規化文書内頻度と呼ぶことにする。

idf

本来は、語 w の idf は、単純に$\frac{\text{全文書数}}{w\text{を含む文書数}}$ であるが、この本来の idf に単調増加関数による変換を施したものも、同様に「少数の文書にしか含まれない」程度を表すものと言える。また、「全体として少数しか出現しない」特殊さの程度を表す$\frac{\text{全語数}}{w\text{の総出現回数}}$ や、それに単調増加関数による変換を施したものも、idf と同類のものと言える。本稿では、これら同類のものやその近似も含めて文書頻度と呼ぶことにする。また、確率的処理ができるように 0 以上 1 以下の値を取るように正規化した文書頻度を特に正規化文書頻度と呼ぶことにする。

しかし、$tf * idf$ モデルには、以下の不十分な点がある。

1. キーワード候補がキーワードになりやすいか否かは、それを構成する各要素（の特性）に依存している。$tf * idf$ モデルは、構成要素の特性を反映していない。
2. キーワードランキングにおいては、キーワード文字列の長さも重要な要因となっているとの報告もある [4]。$tf * idf$ モデルは、単語長を反映していない。

そこで、キーワードランキングの精度向上を目指して、$tf * idf$ モデルを以下のように拡張した。

2 重要度算出モデル

2.1 方針

モデルを検討する上で、上記問題をかんがみて、以下を方針とした。

1. キーワード候補の構成要素から得られるキーワードになりやすさを確率的な値[1](以下ではキーワード確率と呼ぶ) によって表現し、それをキーワード重要度の計算に反映させる。
2. 単語長 len を正規化し、ランキングに反映させる。

また、以下の点も考慮した。

1. 日本語文には複合語が多数現れるが、キーワードにふさわしい複合語範囲の決定は難しい [5]。したがって、複合語に関しては、適当な単語連鎖の全てをキーワード候補とし、そのなかから適切なキーワードとして選択する。また、包含関係にあるキーワード候補の出現による文書内頻度の補正も考慮する。
2. tf, idf の値が 0 ～ 1 の範囲であると確率的扱いができるので、正規化を導入する。
3. tf の算出において、複合語の重複を考慮する [1]。

2.2 モデルの式

各キーワード候補に対する重要度 rel は、文書内頻度・キーワード確率・文書頻度・単語長から算出されることとする。具体的算出方式としては、以下の二つを考えた。

- モデル 1 ：
$$rel = kp \cdot ntf \cdot nidf \cdot nlen \tag{1}$$

ここで、kp はその単語のキーワード確率、ntf は正規化文書内頻度、$nidf$ は正規化文書頻度、$nlen$ は正規化語長であり、演算 $\cdot$ は通常の算術的な乗算を意味する。

[1] 総和が 1 になるとは限らないので厳密には確率とは言えないが、確率に類似のものである。

- モデル２：

$$rel = nkptf \cdot idf \cdot nlen \tag{2}$$

ここで、$nkptf$ はキーワード確率を考慮した正規化文書内頻度である。

2.3 QJP について

キーワード候補の切り出しと品詞情報・構文情報の取得には、簡易日本語解析系 QJP [3] を用いた。QJP は、品詞ごとの表記に用いる文字種・接続上の制約などに着目して形態素解析を行ない、解析用辞書は極めて小規模で、解析処理が高速であるという特長を持っている。また、QJP は、同一文字種から構成される場合、複合語らしいものをなるべく連結して長い単位での単語切り出しを行なう。

2.4 キーワード確率

キーワード確率は、文字・品詞・構文の３つのレベルを考え、それらを合成したものとした。文字レベル (kp_{chr}) 、品詞レベル (kp_{pos}) 、構文レベル (kp_{syn}) の３つの合成は、以下の式で行なう。

$$kp(w) = \bigoplus_j kp_{chr}(w_j) \otimes kp_{pos}(w_j) \otimes kp_{syn}(w_j) \tag{3}$$

ここで、w_j はキーワード候補 w の j 回目の出現を意味する。また、演算$\oplus$ と演算$\otimes$ は、それぞれ和と積のイメージの演算であるが、具体的には以下の中から選択して使用する。

(1) 平均値

(2) 確率的論理積
独立な事象の積事象の確率の計算方法 $x \cdot y$

(3) 確率的論理和
独立な事象の和事象の確率の計算方法 $x + y - x \cdot y$

(4) 最小値

(5) 最大値

(6) 算術的な和
ただし、$\oplus$ としてのみ使用する。

2.4.1 文字レベルのキーワード確率

人間がキーワードを抽出した学習用データから、各文字がキーワードに含まれる統計的確率を求め、辞書化しておく。キーワード候補が構成される各文字についての統計的確率を演算$\otimes$ で合成して、そのキーワード候補の文字レベルのキーワード確率とする。合成方法の選択肢は kp を求める際の演算$\otimes$ の選択肢と同じである。

2.4.2 品詞レベルのキーワード確率

同様のことを品詞レベルで行なう。人間がキーワードを抽出した学習用データから、各品詞の語がキーワードに含まれる統計的確率を求め、辞書化しておく。

2.4.3 構文レベルのキーワード確率

構文情報については、文節の性質がキーワード性に与える影響と、文節対の係り受け関係がキーワード性に与える影響を重みに反映させることを試みた。

2.5 文書内頻度

2.5.1 正規化

モデル1の場合

モデル1では、文書内頻度を Robertson らと同じ方式 [6] によって 0 ～ 1 の範囲になるように正規化する。

$$ntf = \frac{tf}{k_{tf} + tf} \tag{4}$$

モデル2の場合

モデル2では、キーワード確率を考慮して正規化文書内頻度を以下のように計算する。

$$nkptf = \frac{\sum_j pk(w_j)}{k_{tf} + \sum_j pk(w_j)} \tag{5}$$

2.5.2 キーワード候補の包含関係による補正

キーワード候補を長い単位で切り出すので、以下の問題が発生する。

(1) 長い単語の意味は部分的には包含される短い単語によっても表現されているのに、長単位語の出現しかカウントしないと、出現頻度が低くなってしまう。

(2) 長単位語中に含まれる短い単語の出現がカウントされない。

この問題の解決策の一つが疑似キーワード相関法 [4] であるが、今回は2つの重要度算出モデルの枠組みの中で別のアプローチを取り、必要に応じて文書内頻度の補正で対応している。

短い単語を包含する長い単語の文書内頻度の補正

モデル1の場合、あるキーワード候補が出現した場合に、そのキーワード候補を包含する長いキーワード候補の tf 値も、語長比の定数倍だけ増加させ、ntf を求める際に、補正後の tf 値を利用する。例えば、「新幹線」が出現した場合に「東海道新幹線」や「山陽新幹線」の文書内頻度を補正するというものである。モデル2の場合、出現した短いキーワード候補のキーワード確率も $nkptf$ の計算に組み込む。

長い単語に包含される短い単語の文書内頻度の補正

モデル1の場合、あるキーワード候補が出現した場合に、そのキーワード候補に包含される短いキーワード候補の tf 値も、ある定数分だけ増加させ、式 ntf を求める際に、補正後の tf 値を利用する。例えば、「東海道新幹線」が出現した場合に「新幹線」の文書内頻度を補正するというものである。モデル2の場合、出現した長いキーワード候補のキーワード確率も $nkptf$ の計算に組み込む。

2.6 文書頻度

単語単位の文書頻度としては以下のものを用いている。

$$idf = \log\left(\frac{\text{全文書の総語数}}{\text{文書内語数の平均値}\cdot\text{総出現回数}}\right), \qquad nidf = \frac{idf}{K_{idf} + idf} \tag{6}$$

2.7 語長

語長を $nlen = \frac{\text{語長}}{K_{len}+\text{語長}}$ のように正規化して補正因子 $nlen$ とした。

3 評価

3.1 評価実験の方法

学習用データと評価実験用データ

毎日新聞 1993 年版の CD-ROM から 5〜8 月分の記事 550 件を無作為に選択し、2 人の評価者 (以下、"kuwa" と "haru" と呼ぶ) が独立にキーワードを抽出した。キーワードは、評価者の判断した重要度で、上位 5 個の A ランク、A ランク以外で上位 10 個の B ランク、その他の 3 段階にランク付けされている。"kuwa" の 6〜8 月分 (374 件) から統計情報を求め "kuwa" と "haru" の 5 月分 (176 件) を評価に利用した。

評価値の算出

特定のキーワードランキングに関して、ある文書 D の "x" (x = kuwa or haru) の A ランクキーワードの総数を $N_A^x(D)$、そのうちで i 位までに抽出されたものの個数を $n_A^x(i;D)$ とすると、その文書の上位キーワード i 個における "x" の A ランクキーワードの再現率は $R_A^x(i;D) = \frac{n_A^x(i;D)}{N_A^x(D)}$、適合率は $P_A^x(i;D) = \frac{n_A^x(i;D)}{i}$ となる。D に関してこれらの平均を取った $\bar{R}_A^x(i), \bar{P}_A^x(i)$ の値が大きいほど、重要キーワードのランキング順が "x" に近いと言える。同様に、$\bar{R}_{A+B}^x(i), \bar{P}_{A+B}^x(i), \bar{R}_{all}^x(i), \bar{P}_{all}^x(i)$ を定めると、これら全体が大きいものが良いランキング方法と言える。そこで、$V_A(i) = (\bar{R}_A^{kuwa}(i))^2 + (\bar{P}_A^{kuwa}(i))^2 + (\bar{R}_A^{haru}(i))^2 + (\bar{P}_A^{haru}(i))^2$ とし、$V_{A+B}(i), V_{all}(i)$ も同様に定め、以下の式で評価値を求めることにした。

$$\sqrt{\frac{1}{48}\cdot(3\cdot V_A(5) + 2\cdot V_{A+B}(5) + V_{all}(5) + 3\cdot V_A(10) + 2\cdot V_{A+B}(10) + V_{all}(10))} \tag{7}$$

キーワードの一致として、完全一致と部分一致 (両者に包含関係があれば一致したと見なす) の 2 種類が考えられるので、その 2 つの一致判定で求めた評価値の自乗平均を用いている。

3.2 結果

疑似キーワード相関法と比較して、以下の評価結果を得た。なお、"統計情報なし" は、モデル 1 で補正無しの文書内頻度のみ使う場合で、評価の基準になる。

	疑似キーワード相関法	"統計情報なし"	"モデル 1"	"モデル 2"
評価値	0.396	0.409	0.474	0.473

各情報を利用するかどうかや合成方法など、実際の計算方法にはいくつかのバリエーションがあるが、上の表の "モデル 1" と "モデル 2" は、それぞれのモデルで最良の評価値を得た設定によるもので、その各設定は、以下の通りである。なお、ここで「文字レベルのキーワード確率」のように合成方法以外

で合成方法が書いてあるものは、その情報を利用して、複数の合成方法はその合成方法を用いることを意味する。

		“モデル1”	“モデル2”
合成方法	文字・品詞・構文情報の合成方法	平均値	最小値
	複数出現の場合の合成方法	平均値	確率的論理和
文字レベル	文字レベルのキーワード確率	確率的論理和	確率的論理和
品詞レベル	品詞レベルのキーワード確率	利用せず	利用せず
構文レベル	文節属性	利用	利用
	係り受け	利用	利用
	構文情報同士の合成方法	確率的論理和	確率的論理和
文書内頻度	短い単語を包含する長い単語の文書内頻度の補正	利用せず	利用せず
	長い単語に包含される短い単語の文書内頻度の補正	利用 (係数 0.3)	利用
文書頻度	単語単位の文書頻度	利用せず	利用せず
語長	キーワード候補の長さ	利用せず	利用せず

3.3 考察

品詞レベルのキーワード確率の利用は効果がなかった。これは、キーワード候補を長い単位で切り出すので、キーワード候補の品詞がほとんど「名詞」になってしまうためと考えられる。

語長による補正は逆効果であった。これは、文字レベルのキーワード確率の合成でも語長を反映できて、さらに補正を加えると過剰補正になるためと考えられる。

文書頻度の利用は効果がなかった。これは、評価に用いたデータ量が小さいためと考えられる。

4 結論

重要度算出モデルの拡張によって精度の向上が得られることを確認できた。ただし、今回の評価に用いたデータ量が新聞記事550件と小さかったので、大量のデータを用いて評価し直す必要がある。

参考文献

[1] 原 正巳, 中島 浩之, 木谷 強. 単語共起と語の部分一致を利用したキーワード抽出法の検討. 研究会報告 *NL106*, pp. 1–6. 情報処理学会, 1995.

[2] M.A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In *Proc. of 16th ACM SIGIR Conf.*, pp. 59–68, 1993.

[3] 亀田 雅之. 軽量・高速な日本語解析ツール「簡易日本語解析系 QJP」. 第 *1* 回年次大会, pp. 349–352. 言語処理学会, 1995.

[4] 亀田 雅之. 疑似キーワード相関法による重要キーワードと重要文の抽出. 第 *2* 回年次大会, pp. 97–100. 言語処理学会, 1996.

[5] 小川 泰嗣, 望主 雅子, 別所 礼子. 複合語キーワードの自動抽出法. 研究会報告 *NL97*, pp. 103–110. 情報処理学会, 1993.

[6] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proc. of 17th ACM SIGIR Conf.*, pp. 232–241, 1994.

[7] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.

[8] 海野 敏. 出現頻度情報に基づく単語重みづけの原理. *Library and Information Science*, No. 26, pp. 67–88, 1988.