類似文書検索のための単語重要度の共出語分布分析による計算

1P－6

寺本 陽彦, 宮原 豊, 松本 俊二
Youhiko TERAMOTO, Yutaka MIYAHARA, Shunji MATSUMOTO
富士通（株）計算科学技術センター

1．はじめに

類似文書検索においては，一般に，文書に含まれる単語の重要度を利用し，ベクトル空間法によって文書間類似度を計算する手法が採用されている．単語重要度計算の一般的手法として，TF/IDF法などがある．TF/IDF法には，例えば，短い文書においては重要度が単語の出現頻度のみにより決定されるために出現頻度が同程度の単語は重要度も同程度となるなどといった欠点がある．それに替わる新たな手法として，単語と同じ文書に出現する単語（共出語）の分布を基に単語重要度計算を計算する手法を提案する．

2．単語の重要度

文書検索という観点からすると，文書を特徴づける特定的な単語が重要で，一般的な単語は重要でない．そこで，以下のように考えて単語重要度計算の手法を考案した．

1. 一般的な単語は様々なトピックの文書において出現する．したがって，全文書を通じては，その単語と同じ文書に出現する他の単語（共出語）の種類（共出語種数）は多岐に渡る．
2. 特定的な単語は，限定されたトピックの文書において出現する傾向がある．したがって，共出語の種類は限定される傾向がある．

この考え方に基づき，共出語の種類が限定される度合いを単語の重要度として計算することとした．出現文書数が同程度であれば，共出語種数が少ないほど特徴的な単語であり重要な単語であると見做す．単語の重要度を計算するために，単語の出現文書数に対する共出語種数の比率（以下，共出語種比率）を基本項とし，これに，単語の出現文書数の増減による影響を除去するための調整項(A_1)を加えた式を作成した(式１)．単語wの出現文書数を$N(w)$，単語wの共出語種数を$C(w)$ とし，それらを以下の式に代入して単語wの重要度を計算する．

$$I(w) = a_1 - (G(w)/a_2) \quad \cdots(式1)$$

$$G(w) = (C(w)/N(w)) * A_1(w)$$

$$A_1(w) = 1/(1+b_1*(exp(-N(w)/b_2)))$$

a_1, a_2は重要度のダイナミックレンジを決定するためのパラメータ（正値），$G(w)$ は単語wの共出語種比率($C(w)/N(w)$)に調整項(A_1)を加えたもので，単語の一般性をあらわす指標で，一般性が高い単語ほどこの値が大きくなる．A_1は，ある単語に着目したとき，その単語の出現文書数が増加するのに伴い共出語種比率が減少することによる重要度計算に対する影響を減少させるための補正項で，b_1, b_2はその補正の強度を決めるパラメータである．

実際には，実用的な類似文書検索システムへの適用を考慮して，さらに以下の調整項(A_2)を加えた式(式２)を使用している．

$$G(w) = (C(w)/N(w)) * A_1(w) * A_2(w) \cdots(式2)$$

$$A_2(w) = 1/(c_1+c_2/N(w))$$

A_2は，出現回数の少ない単語（稀出単語）の重要度を高く見積もるための補正項で，c_1, c_2はその強度を決めるパラメータである．この項を導入した理由は以下の通りである．

1. **稀出単語は，類似文書検索において重要である可能性が高いという経験的判断．**
2. **稀出単語は，統計量が少ないため一般に統計情報に対する誤差が大きい．類似文書検索において，稀出単語の重要度を誤って高く設定してもノイズの増加という悪影響はそれほど大きくない．逆に，重要度を誤って低く設定すると，重要な事例が他の事例に埋もれてしまうという無視できない影響が発生する．**

A_2 の導入により，出現回数の少ない単語の重要

Word weight calculation for document retrieval by analyzing the distribution of co-occurrence words
Yohiko Teramoto, Yutaka Miyahara, Shunji Matsumoto
Fujitsu Ltd. 1-9-3, Nakase, Mihama-ku, Chiba-City, Chiba, 261-8588, Japan
E-mail: tera@strad.se.fujitsu.co.jp

度を意図的に高くすることで 1.の判断を取り込みつつ，2.のリスクを避けることができる．

3．テストケースへの適用

計算機のサポートに関する 12,000 件の質問応答事例の質問文に含まれる 1666 個の単語の重要度を計算した．質問文は 20 語程度からなる短文であり，冒頭で述べたように，TF/IDF 法により重要度を計算すると出現事例数が同程度の単語に関して重要な単語とそうでない単語を識別できないという問題が生じる．

4．考察

前項テストケースに対して本手法を適用した計算結果を，出現事例数に対する重要度の分布図として図１に示す．表 1 は，出現頻度が同程度の単語に対する重要度の高い単語と低い単語の例である．表 1 から，同程度の出現頻度の単語に関して，例えば，「2051（エラーコード）」「SQL（データベース用語）」「ORA（製品名の一部）」の重要度が高く評価され，逆に「再度」「状態」「設定」の重要度が低く評価されていることがわかる．

また，この結果を以下の要領で分析することで，本手法によって一般的な単語とそうでない単語とを識別できていることを確認した．

英和・和英辞典（「研究社新英和・和英中辞典」）に含まれる見出し語を一般的な単語であると見做し，重要度の高い語と低い語がそれぞれどれだけそれらの辞書に見出し語として掲載されているかを調査した（表 2）．ただし，「10」「20」などの，数値をあらわす単語は調査の対象外とした．表 2 のように，重要度の高い上位 50 単語よりも重要度の低い下位 50 単語の方が辞書の見出し語となっている割合が高いので，一般的な単語とそうでない単語が識別できていると言える．なお，出現事例数が 50 未満の語(1314 語)の場合と 50 以上の語(352 語)の場合とを別々に分析しているのは，A_2 項により，出現頻度が低い語の重要度と多い語の重要度との間に差異が存在するためである．

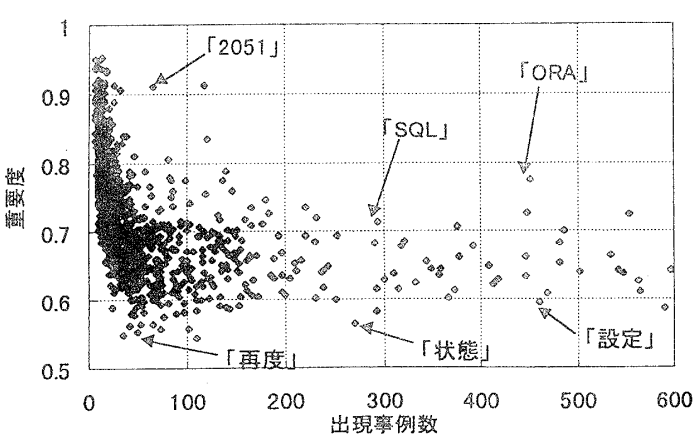

図1 単語重要度の分布

5．まとめ

共出語の分布（共出語種数と出現文書数との対比）を分析して単語の重要度を決定する手法を短文データに対して適用し，この手法が有効であることを確認した．この手法によれば，単語の出現頻度が同程度の単語に関しても，重要な単語とそうでない単語を識別することができる．また，言語依存の知識を利用しないため，日本語以外の言語への適用が容易という利点がある．手法自体の研究課題としては計算式の洗練，重要度計算結果の定量的な評価方法の検討などが挙げられる．

謝辞

研究の実施にあたり，データの提供などで御協力いただきました当社フィールドサポート本部信夫部長，原田課長他関係者の方々に感謝致します．

参考文献

長尾真編：岩波書店 ソフトウェア科学 15「自然言語処理」（１９９６）

表 1 単語重要度計算結果の例

重要度の高い単語の例		
単語	出現事例数	重要度
FAULT	39	0.839
2051	65	0.911
ｿﾌﾄｳｪｱﾀﾞｳﾝ	118	0.912
通訳御	232	0.721
SQL	294	0.714
V4.1	376	0.708
ORA	451	0.776
WSMGR	554	0.724

重要度の低い単語の例		
単語	出現事例数	重要度
以降	35	0.549
再度	50	0.553
問題	110	0.543
変更	230	0.601
状態	270	0.564
処理	373	0.613
設定	461	0.595
データ	564	0.611

表２ 辞書の見出し語となっている語の数と比率

出現事例数	重要度	見出し語数	見出し語率
50 未満	上位50語	12 語	24%
	下位50語	33 語	66%
50 以上	上位50語	12 語	24%
	下位50語	44 語	88%

※各分類において，辞書の見出し語となっている単語の比率（見出し語率）が高い程一般的な単語が多いということを意味する．