ISS-P-317

文書検索における重要度の重み付け

奥村直也† 三浦孝夫††
†法政大学理工学部創生科学科 ††法政大学院理工学研究科電気専攻

1. はじめに

近年、インタネットで電子書籍（ニュース記事）が一般的になっているが、膨大な量の情報からユーザの興味ある情報を正確に抽出する必要がある。これまで主に使われるのが文書検索である。ここでは、文書の意味を語の集合で代用し、その語の集合に語の出現頻度 TF (Term Frequency)となる重みを付け、余弦類似度で近似する。出現頻度 TF だと、すべての語が平等に重みを有し、語の重要性を失う。本研究では、ニュース記事集合を前提として、重要度を考慮した重み IF (Importance Frequency)を提案する。

2. 重要度と情報検索

ニュース記事における重要度の定義を考えるとき、ニュース記事は結論を先行させるケースが多い。すなわち、ニュース記事では先行すればするほど語の重みが重要であるといわれている。重要度は、段落番号 P と文番号 S の逆数を取ることにより，先行されている語の値を大きくする役割を与える。重要度は位置情報により決定される。

$$IF = w \times \log L \times \sum \left(\frac{1}{P} \times \frac{1}{S}\right), w : 0.01$$

L は記事の長さ, P は段落番号,S は文番号を表す。

図 1 は重要度 IF の例である。「曇り」という単語が第 1 段落の第 2 文目と第 2 段落の第 3 文目に出現しているため位置情報は P＝1 ,S＝2 と P=2 ,S＝3 となる。重要度 IF（“曇り”）は、以下のようになる。

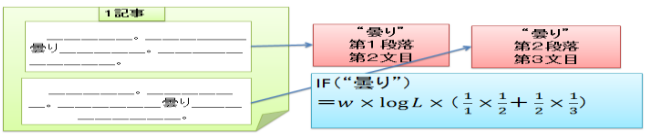

図 1. 重要度 IF の計算例

3. 実験

本研究では、ニュース記事に対して出現頻度 TF より重要度 IF の正確性があるかを確認するために以下の実験を行う。実験には、ニュース記事として毎日新聞 2012 年度 1 月分（計 8046 記事）データを使用する。ニュース記事の本文に対して予め Mecab を使用して形態素解析し、自立語のみを抽出しておく。なお自立語として、名詞の数詞、代名詞は一般的な語は本研究では除外し、動詞、形容詞、副詞をすべての形態を基本形にする。 次に、自立語に重み TF、TF*IDF、IF、IF*IDF を与え、『自立語：重みの値』の順でそれぞれベクトル生成を行う。

本実験ではテストを行う。テストでは、しきい値以上の語の重みの値で余弦類似度検索を行う。実際、10 件の記事の見出しをランダムに選択し、その見出しに対して本文と同様の内容を行う。最後に、それぞれの重みに対してテスト結果をランキングする。評価のため、テストではその見出しの記事が、トップ 10、20、50 に含まれれば正解とする。

4. 評価

図 2 から、テストでは、10 位以内にランクインしているのが重要度 IF：7 割であり、頻度 TF：6 割であるため正確性が優れていることがわかる。

テスト1	TF	IF	TF*IDF	IF*IDF
10位以内	60%	70%	40%	40%
20位以内	60%	70%	50%	60%
50位以内	60%	80%	60%	70%

図 2. テストの実験結果

5. むすび

本研究ではニュース記事における語の重要度の重み付けを提案した。これにより重要度の精度が 7 割を超えているため正確な文書検索を行えることができたと言える。今後の改善点は、ニュース記事だけでなく Web ページなど異なる分野の文書も含む一般的な文書に対応させる重要度の定義を考察することである。今後の文書検索の更に正確性を見込める。

参考文献

[1] 金子 満生 恵谷 淳一郎 松澤 由梨枝 韓 東力：「重要語句抽出を利用した要旨作成システム」

[2] 松尾 豊 石塚 満：「語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム」