6G−3 文章中のキーワード評価値を用いたキーワードN次元空間における類似性検出

中村文子
東京電力㈱　システム研究所ＡＩ研究室

1．はじめに

当社お客さま相談室での相談業務支援を目的とし、キーワードネットワークを用いた文書データの知的検索システム[1][2]を開発し、実証試験中である。

相談事例と回答事例をセットにした文書データ「受付票データベース」は、原子力、屋内配線等の１０数個の分野毎にまとめられ、ユーザは、分野毎に新しい相談事例を随時追加する。文書蓄積量の増大に伴い、類似の文書が重複して含まれる可能性が生じ、データの洗練性や、検索効率低下の原因ともなる。これを未然に防ぐには文書登録時に既存のデータベースに類似の文書が存在しないことの確認が必要である。

そこで、文書間の類似性を検出する等の文書分析機能が必要となってきた。

本論文では、同じ様な内容を持つ文章かどうかという観点からの類似性検出に対し、文書中のキーワード評価値を用いる手法を開発し、評価を行なったので、それについて述べる。

2．キーワード評価値の計算方法

文書の類似性検出には、文書に関する文章特徴量を用いて文書間の距離を求める方法がある。[3]

この場合、いかに適切な文書特徴量を求めるかが重要となる。一般的には文章中の語句の出現頻度値を元に抽出したキーワードを文章特徴量に用いる方法がある。[4]　しかし、これだけでは

①機械抽出語句と人間の判断による選定語句とが完全には一致しない。
②文書を表現するのに最適な語句が必ずしも検索用の語句とはならない。

等の問題があり、人間の検索意図を含んだ文章特徴量を算出することができない。

そこで、機械抽出の評価の他、人間の判定や検索履歴を加味した「キーワード評価値」を用いることとした。

以下にキーワード評価値計算に用いた３つの要素及びその計算方法について述べる。

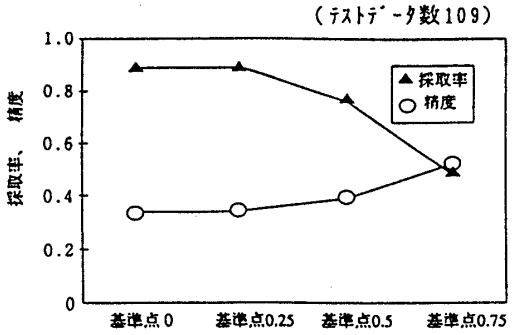
＜要素①＞重要キーワード得点

タイトル文や、主語、目的語を含む重要文節中の語句には高得点を与える等によりキーワード得点（0.00～1.00の連続値）を決定し、基準点以上のものを重要キーワード候補としてユーザに提示する。[5]

基準点を４つ選んで抽出した結果と人間の抽出結果とを比較した結果を図１に示した。

採取率も精度も1.0に近いほど優れた機械抽出と言えるが、基準点が上がるにつれ、採取率は低下し、一方精度は向上する。ここでは採用基準に0.5を採用し、その後人間の判断を加えること（要素②）で適切なキーワードの設定が実現できるようにした。

図１　自動抽出と人間の設定との比較



採取率：人間の設定したキーワードのうち機械で抽出できたものの割合
精度　：機械で抽出したキーワードのうち人間が設定したものの割合

（キーワードの得点−0.50）×２を重要キーワード得点とし、この値を　t1 とする

＜要素②＞人間の判定

重要キーワード候補の全てに対してユーザが採否の判定をする。この際、採用または修正キーワードには 0.4　を、追加キーワードは機械抽出後に人間が必要と判断した上であえて設定したものであるから 0.8 の高得点を与える。不採用のキーワードには 0 を与える。この値を t2 とする。

＜要素③＞検索履歴情報

要素①における得点はその文書の示す内容に応じた得点であり、必ずしも情報検索に適したキーワードではない。その補完として要素②があるが、さらに検索履歴情報を加え、より一層人間の感覚に近い得点を与えることとした。実証試験での実際の検索履歴をカウントし、その文書検索に最多使用されたキーワードに1.00を、他のキーワードにはその使用頻度に応じて0.00～1.00の得点を与え、この要素に対するキーワード得点　t3 とする。

Similarity measurement of documents using the N-dimensional Keyword Evaluation space
Fumiko Nakamura
AI Technology, Computer & Communication Research Center, Tokyo Electric Power Company

＜キーワード評価値＞

要素①～③を次式で合成し、各文書毎の各キーワード評価値 Kを求める。

$$K = 1 - (1 - t1)(1 - t2)(1 - t3) \qquad (1)$$

3．類似度の算定

上記の計算により、各文書には文章特徴量として、複数のキーワード評価値が与えられる。

即ち、文書aの文章特徴量は、

（Ka1, Ka2, Ka3, ..., Kan）で示され、この値を用いて文書相互間の距離を求めることで類似性を判定することができる。この際、同意語同士のキーワードは同じものとして計算する。類似度は次式で求め、値が小さいほど類似性は高くなる。

$$D_{ab} = \sqrt{\sum_{i=1}^{n} (K_{ai} - K_{bi})^2 / n} \qquad (2)$$

Dab ：文書aとbのキーワードn次元空間上距離
Kai ：文書aのキーワードiの評価値
Kbi ：文書bのキーワードiの評価値
n ：文書aとbのいづれかで使用されているキーワードの数（キーワード次元数）

4．評価実験

既存の文書データを用い、本手法による類似文書の抽出と人間による類似文書の抽出結果を比較検討した。

＜機械による類似文書の抽出＞

キーワード評価値算定の3要素の組み合わせから4種類の類似度を算出して各々で類似文書を抽出した。

実験には既存の１２２の文書を用い、その中から選んだ4文書に対し、他の１２１文書から類似文書を抽出した。

	要素①	要素②	要素③
類似度A	○	○	○
類似度B	○	○	×
類似度C	×	○	○
類似度D	×	○	×

＜人間による類似文書の抽出＞

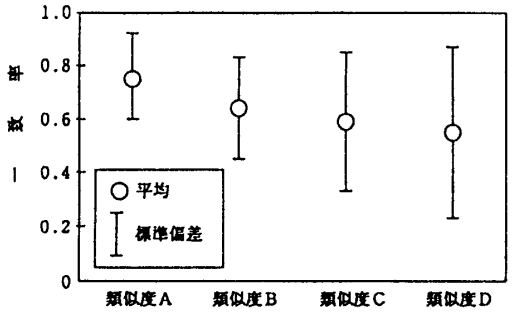
１１人の評価者は、機械抽出に用いたのと同じ4文書に対し、次の基準から類似文書を抽出した。

基準：文書を新規登録する時、類似文書として、是非提示すべき（類似性が高い）文書である

＜評価結果＞

評価は、各類似度において抽出された類似文書の中に人間が類似文書として抽出した文書が含まれる比率で行った。図2に、その比率（一致率）の平均及び文書間のばらつき結果を示した。全ての要素を用いた場合で最も一致率が高く、文書間のばらつきも小さく良い結果となった。

図２　４種の類似度の評価結果



一致率：人間が抽出した類似文書のうち機械抽出の類似文書が含まれる割合

今回の評価方法では、既存の文書を用いたため、極めて類似性が高いと言える文書が必ずしも含まれていなかったこともあり、人間の判定にもかなりのばらつきが生じていた。人間が明らかに類似性が高いとする評価実験用文書を意図的に作成して実験する等、評価方法の改善を検討している。

5．おわりに

文章中の語句の位置情報や人間のキーワードとしての採否情報、検索履歴情報をうまく組み合わせて用いることで、人間の判定に近い、類似文書の抽出を行なうことが出来た。

しかし、実際には、新しく登録する文書には検索履歴情報はなく、また既存文書も充分な検索履歴を持たない場合がある。さらに、今回はシソーラス関係のキーワードも独立関係と考えて計算しているが、理論的にはこれも考慮すべきと考える。一方で、実用面を考えるとシソーラス辞書そのものの構築が難しいことから、同意語関係のみを考慮することでも充分な類似性検出精度が得られるものと考える。

また、本手法は文書データの自動分類等への適用も考えられる。

今後は重要キーワードや各要素得点のより適切な設定、検索履歴情報以外の情報の活用等により、より一層精度の高い類似性検出手法としていく予定である。

参考文献

[1]竹島他、連想ﾈｯﾄﾜｰｸを用いた相談支援ｼｽﾃﾑ　第38回情処学会全国大会 1989-3
[2]竹島他、連想辞書を用いた知的検索ｼｽﾃﾑ　電気学会情報処理研究会 IP-89-3 1989-8-29
[3]長尾真著　昭晃堂刊　人工知能ｼﾘｰｽﾞ2「言語工学」
[4]寺野、浅居、菅野共編　オーム社刊「ファジイ工学入門」
[5]中村正規他、重要ｷｰﾜｰﾄﾞ抽出方式とその活用方法　情処学会ﾃﾞｰﾀﾍﾞｰｽｼｽﾃﾑ研究会　HO3-IP-0003 1991-7-18