類似度評価を用いたキーワード検索

2S−8

小野寺　浩　　細井　正樹

富士通ｴﾌ･ｱｲ･ﾋﾞｰ株式会社

1. はじめに

名称文字列をｷｰﾜｰﾄﾞとした情報検索では、任意のｷｰﾜｰﾄﾞに対して、その同意語が複数存在する可能性を考慮しなければならない。これらには、「ﾎｰﾑﾗﾝ」と「本塁打」などで表される異表記、「ｽｽﾞﾑｼｷﾞｿｺｳ」と「ｽｽﾞｷﾞﾝ」などの省略語、「ｼｭｳﾏｲ」と「ｼｭｰﾏｲ」のような表記のｺﾚ、「ﾀｲﾍｲ」と「ﾀﾞｲﾍｲ」のような読みの誤りまで様々のものがある。このうち異表記については類義語辞書[1],[2]に関係を定義しておき、検索時に展開する方法が一般に知られている。しかしながら、上記の全てについて、それぞれ類義語辞書に登録することはﾒﾝﾃﾅﾝｽ上、効率的でないことが予想される。そこで我々はこれらの同意語のうち、異表記以外に関して、ｷｰﾜｰﾄﾞと被検索文字との間の類似度を動的に算出し、類似度の高い被検索文字群の中から該当ﾃﾞｰﾀを特定する方式を提案し、評価することにした。

2. 類似度評価方式

名称文字列では『文字の先頭付近が意味的重要性が高い』という仮説をたて、以下の”照合重要度”による類似度評価方式を考える。この方式の特徴は、比較文字列と被比較文字列（文字列長の長い方）との意味的距離を評価しようというものである。

2.1 照合重要度

被比較文字列の各々の文字位置に先頭文字を最大として末尾に従って減少するような意味的重み(1)*1を設定する。次に比較文字列と一致した被比較文字の文字位置の重みの和を求め(2)、それを被比較文字列の全文字位置の重みの総和(3)で割る（β_i / γ_i）。

(1) $S_i = e^{-0.1(i-1)}$　　（iは文字位置）

(2) $\beta_i = \sum_{i=1}^{n} \alpha_i S_i$　　（$\alpha_i = 0,1$）[0:不一致　1:一致]

(3) $\gamma_i = \sum_{i=1}^{n} S_i$

3. 実験ｼｽﾃﾑ

本方式を評価するために、以下の銀行振込顧客名の名寄せ処理を考える。

—A社ではﾕｰｻﾞから振り込まれた際に送られてくる入金ﾃﾞｰﾀ（ｶﾅ顧客名、入金額、その他）を入力として、回収ﾏｽﾀ中のｶﾅ顧客名、請求金額とﾏｯﾁﾝｸﾞして対応するﾃﾞｰﾀを特定する処理を行おうとしている。ところが回収ﾏｽﾀへの入力はA社の各顧客担当営業がおこない、入金ﾃﾞｰﾀに関しては顧客が金融機関の窓口などで記入するため顧客名に関しては両者で微妙なﾃﾞｰﾀの不整合が生じる。このため、完全一致など従来型のｷｰﾜｰﾄﾞ検索方式では検索もれが生じる可能性が高い。—

我々はこの処理における顧客名の検索に本類似度評価方式を用いたｼｽﾃﾑを構築し、検索実験をおこなうことで本方式を評価することにした。

3.1 実験方法

試行に用いるﾃﾞｰﾀは、回収ﾏｽﾀ5835件に対して、入金ﾃﾞｰﾀ542件。これらは実際のｼｽﾃﾑからｻﾝﾌﾟﾙ抽出した。

また、実験の条件としては、本方式によって、算出された被検索顧客ﾃﾞｰﾀを類似度の高い順にｿｰﾄし、その上位候補について、絞り込みをおこなう。そして金額が一致したﾃﾞｰﾀが回収ﾏｽﾀにあった場合をその入金ﾃﾞｰﾀにおいて「ﾏｯﾁﾝｸﾞが成功した」とし、その件数を集計した。

3.2 実験結果①

本方式と金額ﾃﾞｰﾀを用いた絞り込みをおこなった結果が以下の通りである。

成功率	90.6 %　(491/542件)
ゴミ	2.2 %　(11/491 件)
CPU 時間	300 min

表3.1 全件探索による類似度評価

Information Retrieval using a Similarity Evaluation
Hiroshi Onodera, Masaki Hosoi
Fujitsu Facom Information Processing co.

この結果から、ｱﾙｺﾞﾘｽﾞﾑ自身は90.6% という高い成功率であるので満足できる。しかし、本方式では、文字列の比較になんら制約が加わらないので、全件探索が繰り返し発生することになり、ﾏｼﾝﾊﾟﾌｫｰﾏﾝｽは非常に悪い。

3.3 実験結果②

そこで、意味的重要度の高い「先頭文字だけは最低限一致しなければならない」ことを条件に同一環境で再度実験をおこなった。その結果は以下の通り。

成功率	93.4 % (506/542件)
ｺﾞﾐ	0.5 % (3/506件)
CPU 時間	2 min

表3.2 先頭文字列一致による類似度評価

この方式を用いると、『ﾖｺﾊﾏﾏｷﾁｮｳ』と『ﾊﾏｷﾁ』のような類似検索は成功しなくなるが（含542 件中 5 件）、ｺﾞﾐは 3 件に減り、ﾏｼﾝﾊﾟﾌｫｰﾏﾝｽは劇的に向上した。

4. 他のｱﾙｺﾞﾘｽﾞﾑとの比較

本方式の有効性を評価するために、以下の方式について、同様の実験をおこない、結果を比較することにした。

・完全一致
・前方一致率
・照合重要度

完全一致は文字列同志が完全に一致したものをすべて抽出して評価対象とする。前方一致率は文字列の先頭から連続的に一致する文字数をｶｳﾝﾄしていきその数を被比較文字総数で割ったものを類似度とする。

4.1 比較結果

以上の方式を用いて実験をおこなった結果が以下の通り。

方式	ﾏｯﾁﾝｸﾞ 成功率
①完全一致	49.1% (266/542件)
②前方一致率	92.3% (500/542件)
③照合重要度	93.4% (506/542件)

以上の比較結果から、完全一致による検索方法は50%を割ってしまい、やはり文字列ﾃﾞｰﾀ の検索には適さないことがわかる。

一方、前方一致率による実験は高い成功率が得られており、2.で示した仮説『文字の先頭付近が意味的重要性が高い』が有効であることを示している。その反面、例えば 1 文字でも異なれば、その時点までの評価値しか算出しないため、『ｼｮｳﾅﾝﾌﾞｯｻﾝ』と『ｼｮｰﾅﾝﾌﾞｯｻﾝ』では先頭から 3 文字目以外はすべて同じなのに類似度評価は22%(2/9)という結果となってしまう。

これに対して、照合重要度を用いた方式では、文字列の意味的距離を評価するのが特徴であるため先頭に近い文字の重要度も評価するが、文字列全体としての評価値を算出するので、たとえ 3 文字目が一致しなくても、それ以降の一致率が高ければ全体の評価値は極端に影響されない。（ちなみに本方式を用いたときの上記の例での類似度は93.9%であった）

5. 評価

本実験の結果を総合すると、名称文字列をｷｰﾜｰﾄﾞとした検索処理では、先頭 1 文字がｲｺｰﾙであるという条件下において、照合重要度を用いた類似度評価方式を適用した場合に、最も高い確率で、定められた閾値内に対象ﾃﾞｰﾀが検出され、かつﾏｼﾝﾊﾟﾌｫｰﾏﾝｽの点でも良好であることが証明された。

6. 今後の課題

今後はﾏｼﾝﾊﾟﾌｫｰﾏﾝｽを考慮した全件探索手法やｺﾞﾐﾃﾞｰﾀの原因調査などをおこなって更なるﾏｯﾁﾝｸﾞ成功率の向上をはかりたい。

参考文献

[1] 和田 他：ｷｰﾜｰﾄﾞの意味的類似度判定を段階的におこなう検索ｼｽﾃﾑの提案
情報処理学会 第43回全国大会 6M-7, 1991

[2] 中谷 他：事例ﾍﾞｰｽ型設計支援ｼｽﾃﾑSUPPORTにおにおける類似例探索方式
情報処理学会 第41回全国大会 1F-1, 1990

*1 本稿では意味的重みの指標値の 1 例として、指数関数式 $e^{-0.1(i-1)}$ を用いている。