6-6

検索質問の潜在的意味関連に基づく知的情報検索

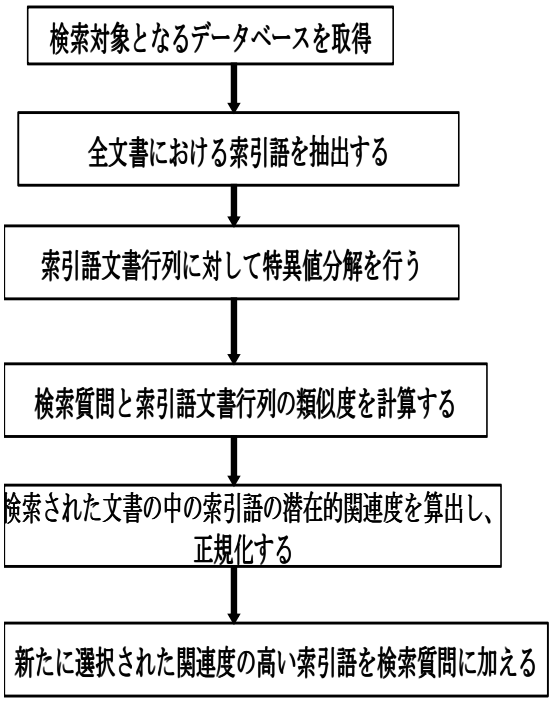
日大生産工（院）　　○閻　飛
日大生産工　　山下　安雄

１．はじめに

近年、膨大な文書やデータベースからある検索語または事柄について書かれた文書を検索する方法として最も一般的に用いられるのは、その検索語または事柄に関する単語を含む文書やデータだけ探し出すという方法である。この検索方法は、検索システムにユーザが検索質問を入力すると、検索システムは検索結果として、その検索質問の検索語を含む文書を表示する。

通常、文書やデータベース管理システムに対する処理要求を文字列として表したキーワードのことを検索質問（ｑｕｅｒｙ）と呼ぶ。しかし、検索要求をユーザが正確に検索質問として表現できる場合もあるが、時としてユーザの意図している検索質問が見つからずに、ユーザが検索したい意味内容を表現できない場合もある。また、情報検索システムでは、検索質問と文書中の検索質問が一致することにより検索が行われ、言い換え表現などのような概念に対して表現の多様性を考えることなしに、パターンマッチングでの検索が行なわれるという問題が生じる。

そこで検索質問について書かれた文書を一つ見つけたとき、その文書と関連した文書を検索することが考えられる。本研究では、入力した検索質問において一度検索を行い、その結果から上位に検索された関連文書を検索質問に加え、さらに関連がないと判断された文書を検索質問から差し引いて再検索する方法を利用した。結果として文書の中に含まれている潜在的な意味関連に基づく検索質問を拡張することにより、検索精度を高めることを図る。



Ｆig.1　提案した検索処理の流れ

２．検索処理の流れ

提案した検索処理の流れは Fig.1 に示している。具体的な内容は以下のようになる。

Intelligent Information Retrieval based on Potential Meaning Connection of Search Question

Fei YAN and Yasuo YAMASHITA

Step-1

検索対象となる文書またはデータベースを入手する。

Step-2

文書から索引語（文書の内容を特徴付けるうえで重要な単語の）を抽出し、索引語文書行列を作成する。

Step-3

索引語文書行列に対して、特異値分解（Singular Value Decomposition, SVD）などの行列変換手法で多次元空間における文書ベクトルの次元を削減する。

Step-4

検索対象となる全文書の検索質問に対する適合度、すなわち、類似度の計算を行う。

Step-5

検索された文書の中から類似度の高い文書いくつか選出し、その抽出された索引語の文書に対しての潜在的関連度を算出し、正規化する。

Step-6

関連度の高い索引語は検索質問との関わりが深くと見られ、検索質問に加え、再検索を行う。

３．検索質問の潜在的関連度

本章では、検索に使われる計算方法について述べる。

３．１　索引語の抽出

一つの文書の中に数多くの単語が含まれている。しかし、文書に含まれている単語すべてが一律に文書の内容と関係しているわけではない。たとえば、日本語の場合には助詞や助動詞、英語の場合には冠詞や前置詞などは、文書の内容とほとんど意味がない。したがって、文書の内容を特徴づける単語を抽出することが重要である。このような文書における重要な意味をもつ単語のことは索引語と呼ぶ。

索引語を抽出する方法に関して、二つの選択肢が考えられる。まず、人手による索引語抽出である。すなわち、人間が見て重要だと思われる単語が抜き出す方法である。しかし、この方法は大規模な文書集合の索引語を抽出する場合は膨大な作業コストを必要とする。このような問題を解決するために、自動的に索引語を抽出する方法がある。単語の出現頻度や共起関係などに基づき、文書から索引語を抽出する。Fig. 2 は自動抽出した索引語の例を示す。

原文：**an information storage and retrieval system capable of both mechanized and manual access at the option of the user is described . each reference is a stylized abstract . the keywords are assigned role indicators and a permuted index is available . the system achieves mechanical flexibility and both mechanical and manual access on a low budget**

抽出後の索引語：

assigned role indicators , available , both mechanical , both mechanized , described , information storage, keywords, low budget, manual access , option , permuted index , reference, retrieval system capable, stylized abstract, system achieves mechanical flexibility、

Fig. 2　索引語の抽出例

３．２　ベクトル空間モデル

検索質問と文書を同じ索引語の重みを要素とするベクトルで表現することによって、各文書がどれくらい検索質問に適しているかをベクトル間の類似度に帰着することができる。ベクトル空間モデルは、ベクトル間の類似度によって検索質問に対する文書の適合度を計算するモデルである。

検索対象となる文書を$D_1, D_2, \cdots D_n$とし、これら文書集合全体と通して、全部でm個の索引語$T = \{t_1, t_2, \cdots, t_m\}$があるとする。このとき、文書$D_j$は、式(1)ようにベクトル$d_j$で表現される。これを文書ベクトルと呼ぶ。

$$d_j = (d_{ij}, d_{2j}, \cdots d_{mj})^t \qquad (1)$$

ここで、tは転置、d_{ij}は索引語t_iの文書D_jにおける重みである。文書ベクトルを作成するとき、ベクトルの要素には局所的、大域的索引語の分布を考慮するために、索引語の頻度に重み付けした数値を用いる。

また、文書集合全体は、$m \times n$行列Dによって表現することができる。式(2)のように行列Dは索引語文書行列と呼ぶ。

$$D = [d_1\, d_2 \cdots d_n] = \begin{bmatrix} d_{11}\, d_{12} \cdots d_{1n} \\ d_{21}\, d_{22} \cdots d_{2n} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ d_{m1}\, d_{m2} \cdots d_{mn} \end{bmatrix} \qquad (2)$$

検索質問も、文書と同じに表現することができる。検索質問に含まれる索引語t_iの重みをq_iとすると、検索質問ベクトルqは、次の式(3)ように表わされる。

$$q = (q_1, q_2, \cdots q_m)^t \qquad (3)$$

文書の検索質問に対する適合度、つまり類似度は、よく余弦尺度を用いられる。余弦尺度は、２つのベクトルのなす角度で表わし、次の式(4)のように求める。

$$\cos(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \|q\|} \qquad (4)$$

３．３　特異値分解

潜在的意味インデキング(Latent Semantic Indexing; LSI)は、高次元の空間にある文書ベクトルを低次元の空間へと射影することにより、検索精度の改善を図る特異値分解に基づく手法である。索引語の文書行列Dが与えられた場合、この行列Dは特異値分解によって次のように表すことができる。

$$D^{(k)} = U^{(k)} \Sigma^{(k)} V^{(k)T} \qquad (5)$$

rは行列Dのランク(n, m)の小さい方の数とすると、$U^{(k)}$は$n * r$行列、$V^{(k)}$は$m * r$行列。$\Sigma^{(k)}$は$r * r$対角行列となる。さらに特異値の大きい順にk個だけを使って再構成した行列$D^{(k)}$を求める。

３．４　検索質問の潜在的関連度

文書における索引語の重みとその文書の検索質問に対する余弦の積が潜在的関連度となる。文書に含まれている索引語Tの中のt_lに対する潜在的関連度Pが以下のように求める。

$$P(T, t_l) = \sum_{j=1}^{n} d_{lj} \cdot \cos\left(v_T, d_j^{(k)}\right) \qquad (6)$$

d_{lj}は文書D_jに出現する索引語t_lの重み、v^Tはt_lに対応する要素の値のみ１となり、その他は０であるベクトル、$d_j^{(k)}$は$D^{(k)}$から得た文書ベクトルとなる。

また、検索質問と索引語に対する意味関連のある単語を選出するため、正規化する必要がある。検索質問に対する索引語について正規化した関連度は以下の式(7)ように求める。

$$P_n(T, t_l) = \frac{P(T, t_l)}{\sum_{j=1}^{n} d_{lj}} \qquad (7)$$

算出された関連度の高い検索質問に関連のある索引語と考えられ、高い順からいくつかの索引語を選び、検索質問に加える。

4　実験方法

実験で情報検索システムの評価用テストコレクションを用いる。テストコレクションは「情報検索システムの検索有効性評価に用いる、正解データを含めた実験用データセットを指す。これは（1）データベース、（2）利用者の検索要求を記述した「検索質問」、（3）検索質問を満たす「正解文書の網羅的なリスト」から構成されている。

まず、前処理として索引語の抽出を行う。索引語の重み付けには索引語の頻度で得た数値を用いる。このようにして得られた索引語から索引語文書行列を作成し、その行列に対して特異値分解により次元削減を行い、近似行列を計算する。近似行列と検索質問との類似度を算出し、類似度の上位の文書を選んでその文書の索引語の潜在的関連度を求め、正規化する。関連度の高い索引語が検索質問に加え、再検索を行う。

検索システムの評価には、一般的に適合率（Precision）と再現率（Recall）を用いる。

$$\text{Recall} = \frac{\text{出力した正解文書数}}{\text{全正解文書数}}$$

$$\text{Precision} = \frac{\text{出力した正解文書数}}{\text{出力した全文書数}}$$

5．まとめ

本論文では、一度行った検索の結果において類似度の高い方の文書の索引語を選出し、新たに検索質問に加える方法を選んだ。この方法は従来の情報検索の方法に比べ、検索質問を拡張することによって検索精度の改善を期待できる。

しかし、今回の方法はカテゴリー化された文書からの検索を行ったため、今後検索方法の対応性についてさらなる検討が必要と思われる。そして、索引語の抽出に関する問題もある。英語などのヨーロッパ系の言語の場合には、単語が空白により区切られているため、単語の同定はきわめて容易である。しかし、日本語や中国語などでは単語間がつながっているため、単語の同定が難しくなる。その問題の解決には形態素解析などの技術を用いて索引語を抽出する方法を考える必要がある。さらに、単語の同義性や多義性を考慮する検索も今後の課題となる。

「参考文献」

1)藤井啓彰，小島一秀，渡辺広一，河岡司，概念間の関連度に基づく情報ランク付けを用いた知的検索手法，人工知能論文誌，17 巻 6 号 D，(2002)，pp.684～689.

2)川前徳章，青木輝勝，安田浩，統計的潜在的意味空間の抽出，情報処理学会自然言語処理研究会報告，(2002)，pp.25～30.

3)鷹野孝典，増田圭祐，内田亜美，陳幸生，動的フィードバック機能をともなった特徴語抽出方式における文書ベクトル空間改善プロセスに関する評価実験，情報処理学会データベースシステム研究会報告，(2007)，pp.61～66.

4)松村敦，高須淳宏，安達淳，情報検索における単語間の関係効果，情報処理学会データベースシステム研究報告，(2001)，pp.257～264.

5)http://research.nii.ac.jp/ntcir/outline/prop-ja.html.