3Z－04

頻度情報に基づくキーワードの洗練化を利用したテキストデータ検索手法の提案*

中嶋　章　　樽松　理樹†
岩手県立大学　ソフトウェア情報学部‡

1. はじめに

近年、大量のテキストデータ（半構造データ）から目的にあったデータを検索する機会が増えている。そのような場合、一般的にキーワードを与え、そのキーワードとの関連性から、目的にあったテキストデータの検索を行う。これまでに、様々な検索手法が提案されており[1]、検索の性能も向上してきている。しかし、検索結果には依然として、誤った情報が含まれることや、検索できなかった情報が存在する場合が多い。そのため、ユーザはキーワードを変更するなどして、再度検索を行う必要があり、検索に係る負荷は依然として、大きい。

このような問題に対し、本研究では、検索の精度を向上させることを目的とし、ユーザが与えたキーワードを頻度情報に基づき洗練化（選択）し、それらを用いて検索を行う手法を提案する。さらに、本手法の有効性を、評価実験を通して検証する。

2. 提案手法

本提案手法は、(1) ユーザが与える検索質問に含まれるキーワードを、検索対象となるテキストデータ中の出現頻度を用いて洗練化する。(2) 洗練化で得たキーワードに基づき、検索対象となるテキストデータをベクトル表現(文書ベクトル)に変換する。各文書をベクトル表現に変換することで、検索の効率化を図ることが期待できる。(3) 同様に検索質問を洗練化されたキーワードでベクトル化し、文書ベクトルとの類似度を計算することで、目的の文書の検索を行う。以下、それぞれのステップについて、詳細な説明を行う。

2.1. キーワードの洗練化

本手法では、最初にユーザが与えたキーワードを、次のような手順で洗練化する。

はじめに、ユーザが与える検索質問に含まれるキーワード（以下、索引語）に対し、検索対象となるテキストデータにおけるそれぞれの idf の値を式(1)を用いて求める。idf とは、ある索引語が全文書中に出現する頻度を表す尺度であり、この値が高いほど、その索引語の出現する文書が少ないことを意味し、索引語としての利用価値が高くなると考えられる。

$$idf(term) = \log \frac{N}{df(term)} + 1 \quad (1)$$

ここで式(1)中の各記号の意味は、次の通りである。

term ：索引語　　*N* ：全文書数
df ：索引語が1回でも出現する文書数

次に、一定の割合を閾値として設定する。この閾値と idf の最大値との積以上の idf 値をもつ索引語のみを、以後、検索語として利用する（洗練化）。

2.2. 文書のベクトル化

次に、各文書中に含まれる洗練化の結果から得た索引語に対して、tf·idf を求める。それらの値を各要素の値とするベクトルに変換する。ここで、tf·idf とは、ある索引語に対する tf と idf との積である。tf とは、その索引語が一つの文書中に出現する頻度を表し、tf·idf が高い索引語は、その語が出現する文書が少なく、かつ、出現する文書中においては頻繁に現れる語ということを意味する。よって、この値が高いほど、その文書を特徴づける語と言え、索引語として有効であると考えられる。

以上のような手法で、全文書をベクトル化する。求まったベクトル（文書ベクトル）に対し、検索質問をベクトル化したものとの類似度を求めることで、検索を行う。

2.3. ベクトル空間における類似度の計算

最後に、一つ前の工程で求めた文書ベクトルと検索質問から変換したベクトルとの類似度を計算し、テキストデータの検索を行う。

検索質問は、それに含まれる索引語のうち、洗練化によって抽出された索引語に対する値を要素とするベクトルに変換する。ここで、各要素の値としては、どの索引語にも同じ重みを与える手法と、それぞれの索引語のテキストデータにおける idf を与える手法を用いる。

文書ベクトルと検索質問のベクトルとの類似度を計算する方法としては、既存の手法であるベクトル空間モデルと拡張ブーリアンモデルを用いる。

2.3.1. ベクトル空間モデル

ベクトル空間モデルでは、文書ベクトル d と検索質問のベクトル q に対し、内積（式(2)）、または余弦（式(3)）を求めることで、類似度の計算を行う。ここで算出された値が大きいほど類似していると判断する。本手法は、二つのベクトルの成す角の角度が小さいほど、類似していると捉えることができる。

$$sim(d,q) = \sum_{i=1}^{T} w_i \cdot q_i \quad (2)$$

$$sim(d,q) = \frac{\sum_{i=1}^{T} w_i \cdot q_i}{\sqrt{\sum_{i=1}^{T} (w_i)^2 \cdot \sum_{i=1}^{T} (q_i)^2}} \quad (3)$$

---

* A text data retrieval using keywords selected based on frequency
† Akira Nakajima and Masaki Kurematsu
‡ Faculty of Software and Information Science, Iwate Prefectural University

ここで、式中の各記号の意味は次の通りである。

w_i ： 文書ベクトル d の各索引語の値

q_i ： 検索質問ベクトル q の各索引語の値

T ： 索引語の総数

2.3.2. 拡張ブーリアンモデル

拡張ブーリアンモデルでは、文書のベクトルdと検索質問のベクトルq間の距離を式(4)を用いて計算する。本手法は、二つのベクトル間の距離が近いほど類似していると捉えることができる。

$$sim(d,q) = 1 - \sqrt[p]{\frac{\sum_{i=1}^{T}\{(q_i)^p \cdot (1-w_i)^p\}}{\sum_{i=1}^{T}(q_i)^p}} \quad (4)$$

ここで式中の各記号の意味は次の通りである。

w_i ： 文書ベクトル d の各索引語の値

q_i ： 検索質問ベクトル q の各索引語の値

T ： 索引語の総数　　p：パラメタ（[1, ∞]）

3. 評価実験

3.1. 実験概要

本手法の有用性を評価するために、実験を行った。実験データとして、検索の性能を評価するために作成されたテストコレクション（Medlars）[2]を使用する。Medlarsは、医学関係の文献データベースで、1033の文書と30の検索質問とそれに対する適合文書の対応表が用意されている。最初に、文書と検索質問から、SMARTシステム[3]で標準的に使われている571個の不要語リストを用いて不要語を取り除く。また、検索質問は文章で書かれているため、文中の名詞と形容詞をキーワードとし、それらの間の関係をAND関係として考える。また、今回扱うテストコレクションは、英語であるため、文章と索引語の接辞処理する場合としない場合についても実験を行う。なお、ベクトル化する際に与える同一の値としては、1を与えることとする。

検索によって得られた結果のうち、上位30位（同順を含む）に対して、再現率（全適合文書のうち、検索された文書の割合）と精度（検索された文書中の適合文書の割合）により、本手法の有用性を評価する。

3.2. 実験結果

表1に、検索質問のベクトル化の際に同一の値を与え、類似度の計算に拡張ブーリアンモデルを用いた結果を示す。ここで洗練化の欄の値は、洗練化における閾値を示し、0%の場合は洗練化を行わないことを意味する。上昇、同じ、下降は、それぞれ0%の場合との比較結果を、接辞処理した場合の0%は、接辞処理しない場合の0%との比較結果を示す。

実験の結果、接辞処理をしない場合は洗練化を65%、接辞処理をする場合は洗練化を70%に設定した時に、最も精度が高く、再現率も高い結果を得た。

また、類似度の計算にベクトル空間モデルを使用した場合も同様の結果を得た。

表1：実験結果（同一の値、拡張ブーリアンモデル）

接辞処理をしない場合								
洗練化		0%	50%	60%	65%	70%	75%	80%
精度	上昇	－	2	9	11	10	10	11
	同じ	－	26	16	13	10	6	5
	下降	－	2	5	6	10	14	14
	平均値	0.37	0.37	0.39	0.40	0.36	0.35	0.33
再現率	上昇	－	2	8	9	9	9	6
	同じ	－	26	19	13	12	7	6
	下降	－	2	3	8	9	14	18
	平均値	0.52	0.52	0.55	0.54	0.50	0.40	0.32
接辞処理をする場合								
洗練化		0%	50%	60%	65%	70%	75%	80%
精度	上昇	11	3	11	14	19	16	17
	同じ	6	23	12	8	3	6	3
	下降	13	4	7	8	8	8	10
	平均値	0.36	0.36	0.39	0.39	0.40	0.40	0.43
再現率	上昇	12	2	11	15	14	12	11
	同じ	4	24	14	8	5	5	3
	下降	14	4	5	7	11	13	16
	平均値	0.52	0.52	0.57	0.57	0.53	0.50	0.41

3.3. 考察

実験の結果、索引語を洗練化することによって、再現率を保ちながら、検索精度は向上することに成功した。これは、頻出するため、利用価値が低いと考えられる索引語を取り除くことで、不適切な文書の検索を回避できたためと考えられる。

しかし、キーワードの洗練化を行った場合に、精度が落ちる場合もあり、特に洗練化時に閾値を高く設定した場合に多い。これは、索引語を絞り込みすぎ、利用価値が高い索引語が失われ、検索されているためであると考えられる。

4. おわりに

本稿では、ユーザが与える検索質問中の索引語を、頻度情報に基づいて洗練化し、それらを用いて検索する手法について提案した。実験の結果、洗練化した場合のほうが、再現率を保ちながら、精度が高くなることが示された。しかし、精度が落ちる場合もあり、今後は、実験結果を検証、反映することで、本手法の改善を図っていく予定である。

参考文献

[1] 徳永健伸, “情報検索と自然言語処理”, 東京大学出版会, 1999

[2] Glasgow IDOM, Test collections, http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/

[3] G.Salton, Automatic Text Processing, Addison-Wesley, 1988