情 報 学 基 礎 47－13
（1997．9．12）

シソーラス掲載語の重要性を考慮した文書スコアリング

高木 徹† 木谷 強† 関根 道隆‡ 出口 信吾‡

†NTTデータ通信 情報科学研究所
‡日本経済新聞社 データバンク局

ユーザの検索作業を支援する観点から、全文検索の検索結果に対して重要度(スコア)を付与する研究が実施されている。一方、シソーラスは重要な概念を表わす語で構成されており、一般的に文書の主題を表わす語が多く含まれている。そこで本論文では、検索語がシソーラスに掲載されている語である場合、その語が出現する文書の重要度を変更する方法を提案する。日本語新聞記事のテストコレクション BMIR-J1 と日経シソーラスを使用し、シソーラス掲載情報を利用する場合と利用しない場合の検索精度を比較した。検索文字列とシソーラス掲載語の文字列の一致度合、および検索文字列が含まれるシソーラス掲載語のカテゴリ頻度をパラメータとして文書の重要度を変化させた結果、再現率が5%向上することを確認した。

Relevance Ranking of Documents Using Importance of Thesaurus Words

Toru Takaki† Tsuyoshi Kitani† Michitaka Sekine‡ Shingo Deguchi‡

†Laboratory for Information Technology, NTT Data Corporation
‡Databank Bureau, NIHON KEIZAI SHIMBUN, INC.

To facilitate users' retrieval work, it is necessary to rank documents according to their relevance. A thesaurus is composed of words which can be main subjects of the documents. This paper describes a relevance ranking method that utilizes importance of query words appearing in the thesaurus. The traditional frequency-based method alone and combined method are compared using the Nikkei thesaurus and a test collection of Japanese newspaper articles called BMIR-J1. Experimental results show that the proposed method, using the thesaurus-term frequency and the degree of string matching between the query and thesaurus word, improves retrieval recall by 5%.

1 はじめに

新聞記事、特許やインターネット上などのテキスト情報が電子化されており、大量の情報をデータベースに蓄積できるようになってきた。大量のテキストから情報を入手する手段として、全文検索が一般的になってきた。全文検索の利点として、検索語が文書中に出現しているものを漏れなく見つけ出せる点があげられる。その反面、検索語を含む文書数は膨大となることが多く、利用者が真に入手したい情報にたどりつくためには、絞り込み検索などの試行錯誤が必要である。そのため、利用者の労力を最小限に抑える検索システムの必要性が高まっており、近年、検索条件に対する文書の重要度をスコアとして示し、スコア順に検索結果を利用者に提示する手法が盛んに研究されている[1][2]。しかし、現状ではスコアの精度は必ずしも高いとはいえない。

本論文は、スコア順に検索結果を提示するシステムにおける適合率の向上を目的とし、スコア算出の要素として、シソーラスに掲載されている語の重要性を利用した手法を検討する。本手法は、従来から広く利用されている検索単語の文書内

出現頻度と出現文書頻度を使用したスコアリング手法をベースとしている。

まず、2章でシソーラスの一般的な利用方法と本研究で用いた「日経シソーラス」について述べ、3章でスコアリング手法、4章でシソーラス情報を用いたスコアリング手法を説明する。5章で従来手法と提案手法を実験により比較評価し、考察を行う。6章では、本論文のまとめと、今後の課題を述べる。

2 全文検索に対するシソーラス利用

本章ではシソーラスの一般的な利用方法と本研究で用いる日経シソーラスについて簡単に述べる。

2.1 検索時でのシソーラスの利用

シソーラスは、語の階層関係（上位語、下位語）、同義関係（同義語）、その他の関係（関連語、参照語）を示している。シソーラスの編集は対象分野の専門家などの人手によることが多く、高度の専門性が要求される。一度作成したシソーラスも時間の変化に応じて、新たな語の追加や不要な語の削除、分類の追加や見直し等、定期的な改訂が必要となる。シソーラスはテキスト検索において、主に再現率を高めるために、検索語を追加する目的で利用されることが多い。システムによっては利用者が指定した検索語の上位語や下位語を自動的に展開し、検索漏れを防ぐものもある。また、シソーラスの階層関係を利用し、意味的な語の類似度を用いて検索精度を向上させる研究も行われている[3]。

2.2 日経シソーラス

「日経シソーラス」は日本経済新聞社が提供する記事データベースの索引付けと検索のために用いる統制キーワード集である[4]。この記事データベースでは、あらかじめ「日経シソーラス」に掲載されている語を記事に対する統制キーワードとして付与してあり、指定された検索語との一致により検索を実行している。日経シソーラスには総数16,830語が次の4種類のキーワードとして収録されている。

(1) 品目キーワード (8,579語)
原材料・素材、半製品、部品、完成品など製品名を中心に製法や工法、技術、情報、サービス、レジャー、スポーツ、組織や施設、自然物、学術、職業など幅広い範囲の用語
(例) 形状記憶繊維、バレンシアオレンジ、マルチメディア

(2) 業界キーワード (445語)
産業界を業種ごとに表現する用語
(例) 音楽業界、機械業界、都市銀行業界

(3) 項目キーワード (7,343語)
行為、状況、関係などのほか、制度、法律、政策などに関する用語（企業や団体の活動状況、政治・経済・産業・社会・国際の動向や状況を表現）
(例) アントイドローン、学校統廃合、銃犯罪、戦争責任、日ロ関係

(4) 地域キーワード (463語)
日本国内の地域ブロックおよび都道府県名、海外の地域ブロックおよび国名。国内外の海洋、海峡、湾、河川、湖沼などの水域名や山脈、砂漠などから主要地名を選択。
(例) 神奈川、九州、アジア、ナイル川

各キーワードは33の大分類、155の小分類に分けられており、小分類ごとに最大6階層の上位・下位関係を保持している。また、一部のキーワードは同義語をもっている。図1に日経シソーラスの例を示す。

大分類：食品　小分類：飲料、酒類

飲料　　(第1階層)
・飲料水　(第2階層,「飲料」の下位語)
・・ミネラルウォーター
　　(第3階層,「飲料水」の下位語)
・栄養補助飲料
・缶入り飲料
・・缶コーヒー
・・缶ジュース
・・缶ビール
・健康飲料
・コーヒー
・・アイスコーヒー
・・インスタントコーヒー
・・缶コーヒー
・・レギュラーコーヒー
・ココア

図 1 「日経シソーラス」の例
（・は階層レベルを示す）

3 単語頻度情報によるスコアリング

全文検索の場合、文書量の多いデータベースを検索したり、出現頻度の高い単語で検索すると、一般的にヒットする文書数が多くなる。このとき、利用者がすべてのヒット文書を参照して所望の文書を漏れなく探し出すことはかなりの労力を

必要とする。スコアリングは、利用者の検索要求に対して検索にヒットした文書にスコアを付与することであり、利用者はスコアの高い文書から参照することにより、素早く必要な文書を参照できる利点がある[5]。

スコアリングの方法として、単語の頻度情報を用いたアルゴリズムが広く検索システムで用いられている[1]。基本的なアルゴリズムは、単語の文書内出現頻度(tf: Term Frequency)、出現文書頻度(df:Document Frequency)、および検索要求内出現頻度を用いて重要度を算出するものである。ここで、ある検索要求QをM個の単語で表わしたとき、データベース内の文書D_iに対する文書内出現頻度TF^{D_i}、出現文書頻度DF、検索要求内出現頻度TF^{Q}はそれぞれ次のようなベクトルで表わせる。

$$TF^{D_i} = (tf_1^{D_i}, \cdots, tf_M^{D_i}) \quad (1)$$

$$DF = (df_1, \cdots, df_M) \quad (2)$$

$$TF^{Q} = (tf_1^{Q}, \cdots, tf_M^{Q}) \quad (3)$$

また、文書D_iのスコア$score^{D_i}$は、

$$score^{D_i} = f(TF^{Di}, DF, TF^{Q}) \quad (4)$$

で表わせる。実際のスコアリングではdfの逆数(idf:InverseDF)を利用することからこの手法はTF/IDF法と呼ばれている。

4 シソーラス掲載語情報による検索語重要度の変更

前章で述べたようにTF/IDF法に基づくスコアリングでは、TFとDFがスコア算出の重要な要素となる。DFは値が小さい(検索語がデータベース内の文書に出現することがまれな場合)ほど、スコアが高くなる。これでは、重要な検索語であってもDFが大きい場合には算出されるスコアは小さくなる場合もある。本研究では、検索語の重要度をDF以外の要素を考慮してスコアリングを実施する手法を考える。

2章で述べたようにシソーラスに掲載される語は専門家により選択され、文書の主題を的確に示すものである。本検討では、シソーラスに掲載されている語は主題（トピック）となる言葉であるため重要であると考え、検索語がシソーラス掲載語である場合には、その検索語の重要度を変更することを試みる。また、検索語とシソーラス掲載語は文字列が完全一致する場合と部分一致する場合がある。文字列の一致度合と検索語が含まれるシソーラス掲載語のカテゴリ頻度によって検索語の重要度を変化させる検討を行う。

5 評価実験

本論文で提案するシソーラス掲載情報を用いたスコアリング手法について、検索精度を評価する。

5.1 評価対象データベース

評価には、情報検索システム評価用テストコレクション(BMIR-J1)を利用した[1] [6]。BMIR-J1は、新聞記事600件、検索要求文60件、および各検索要求文に対する正解集合から構成されている。正解集合にはA,B,Cの3つのランクが付与されており、Aランクは検索要求を主題とする記事、Bランクは記事の主題は検索要求と異なるが、検索要求の内容を少しでも記述している記事、Cランクは全く関連のない記事（不正解）を示している。本評価ではAランクとBランクを正解とした。

5.2 検索語の生成

BMIR-J1では、検索要求は単語ではなくフレーズの形で提供されている。本手法の前提条件は検索要求を検索語に分割して与えるものであるため、検索要求文から検索語を生成する必要がある。本評価では、検索要求文から「茶筌」[7]を用いて形態素解析を行い、抽出された普通名詞、固有名詞、サ変名詞、未定義語を基本検索語とした。また、抽出された基本検索語が、複合してシソーラス掲載語となる場合は、複合語を検索語とした。表1に検索語抽出例を示す。

表1 抽出された検索語の例

検索要求文	抽出基本検索語	検索語
国内航空大手3社	国内、航空、大手	国内航空、大手
海外企業の日本への進出	海外、企業、日本、進出	海外企業、日本、進出
業績悪化を原因とする企業合併の事例	業績、悪化、原因、企業、合併、事例	業績悪化、原因、企業、合併、事例

[1] 株式会社 日本経済新聞社の協力によって、社団法人 情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが、1993年9月1日から12月31日の日本経済新聞記事を基に構築した情報検索評価用データベース(テスト版)を利用

5.3 シソーラス掲載情報を利用したスコアリング

日経シソーラスを参照して、検索語がシソーラスに掲載されているか調べる。本評価では、まずシソーラス掲載語と検索語の文字列の一致度合により検索語を分類した。検索語 $word_i$ の分類を C_i としたときの分類を表 2に示す。検索語がシソーラスに掲載されている場合、該当する文書のスコアを変更する処理を行う。スコアの算出方法は次節で説明する。

表 2 シソーラス参照による検索語の分類

分類 C_i	検索語 $word_i$ とシソーラス掲載語の一致度合
0	一致しない
1	完全一致
2	部分一致

5.4 スコアの算出方法

ベースとなるスコアリングアルゴリズムとして、Comell 大学の SMART システムで採用されている TF/IDF アルゴリズムを用いた。 SMART システムのアルゴリズムを用いたのは検索精度が高いことが TREC で確認されているためである[8]。SMART システムのアルゴリズムは、ベクトル空間モデルと呼ばれている[5]。

ここでは、スコアの算出方法について説明する。式(5)のように検索要求 Q_j が M 個の検索語で表わせる場合、文書 D_i のスコア $score^{D_i}$ は式(6)～(10)で表わすことができる。なお、ここでは、検索要求内に同一単語は重複して出現しないこととする。そのため、式(4)での TF^Q の各成分は 1 となるため TF^Q の項は無視できる。

$$Q_j = (word_1, word_2, \cdots, word_M) \quad (5)$$

$$score^{D_i} = f(TF^{D_i}, DF) \quad (6)$$

$$= f'(\rho(TF^{D_i}) \times \sigma(DF)) \quad (7)$$

$$= \sum_{k=1}^{M} \left[\rho(tf_k^{D_i}) \times \sigma(df_k) \right] \quad (8)$$

ただし、$$\rho(tf_k^{D_i}) = 1.0 + \log(tf_k^{D_i}) \quad (9)$$

$$\sigma(df_k) = \log(\frac{N}{df_k}) \quad (10)$$

ここで、

$tf_k^{D_i}$ は文書 D_i 内に $word_k$ が出現する頻度、

df_k はデータベース内で $word_k$ が出現する文書の頻度（文書数）、

N はデータベースの総文書数である。

また、シソーラス掲載情報による重要度をスコアに反映させるため、式(6)に新たにシソーラス掲載情報 $THES$ を追加する (式(6'))。さらに、5.3節で定義した分類 C_i によりシソーラス掲載情報の影響を変化させるために、分類 C_i に対する出現文書頻度の変化割合 α_{C_i} を設定し、α_{C_i}=0.2, 0.5, 0.8, 1.0, 1,2, 1.5, 2.0, 5.0 と変化させて評価を行った。本評価では式(7'),(8') により各文書のスコアを算出する。

$$score^{D_i} = f(TF^{D_i}, DF, THES) \quad (6')$$

$$= f'(\rho(TF^{D_i}) \times \sigma(DF) \times \psi(THES)) \quad (7')$$

$$= \sum_{k=1}^{M} \left[\rho(tf_k^{D_i}) \times \sigma(df_k) \times \alpha_{C_k} \right] \quad (8')$$

なお、本評価では、出現文書頻度の正確さを向上させるために $\sigma(df_k)$ を算出する際の N と df_k を検索対象のデータベースから取得するのではなく、より大規模な同種のデータベース（日本経済新聞 1993 年一年分、全文書数 186,709 件)より取得した。

5.5 評価方法

各検索要求文から、次節で示す方法で検索語を抽出し検索処理を行い、スコア順に出力された検索結果に対し評価を実施する。なお、スコアリングに用いた記事領域は見出しと本文部分に限定した。

各検索要求ごとに、再現率(recall)が 0, 10, 20, 30, …,100% の場合の適合率(precision)を算出し、検索精度は検索結果の平均値として求めた [5]。評価では、

[評価(A)] 単語頻度情報のみによる検索精度

[評価(B)] シソーラス掲載情報を利用した検索精度

(1) 分類 C_i=1 の検索語の重要度を変更

(2) 分類 C_i=2 の検索語の重要度を変更

を比較することにより、シソーラス掲載情報の効果を評価する。シソーラス掲載語と一致しない分類 C_i= 0 の検索語の重要度は変更しない（α_0 =1 とする)。なお、評価は BMIR-J1 の全検索要求 60 件のうち、単一検索語であるものやすべての検索語がシソーラスに掲載されていないなど、シソーラス情報による精度向上の効果が見込めないものを除いた 39 件の平均を示してある。検索語の

総数は129個であり、分類C_i=0,1,2の検索語はそれぞれ19,57,53個である。また、評価対象の各検索要求では、選択された検索語は最小2語、最大7語、平均3.3語であり、これらの検索語のOR条件により検索を行なった。

5.6 評価結果と考察

本章では実験から得られたデータを分析し、シソーラス情報利用の効果を考察する。

検索語がシソーラス掲載語に完全一致するC_i=1の場合にα_1を変化させたときの評価(A),(B)の検索精度を表 3に示す。α_1が大きい(シソーラス掲載語の重要度を高めたとき)ほど適合率向上の効果がみられる。α_1=2.0のときシソーラス掲載情報の利用により、平均で5.0%の適合率の向上が見られ、シソーラス掲載情報の利用による有効性が確認できた。まず、検索要求ごとの適合率の変化をみる。各検索要求ごとの再現率0, 10, 20, 30, …,100%の計11個の適合率の平均により、評価(A)と評価(B)の適合率の変化を調べる。評価対象とした39検索要求のうち、適合率の向上がみられたものは18検索要求（適合率向上の平均38.4%)、低下したもの9検索要求（適合率低下の平均9.5%)、変化しなかったもの12検索要求であった。また、適合率向上、低下の度合が大きい検索要求を表 4に示す。

つぎに、検索語がシソーラス掲載語に部分一致するC_i=2の場合にα_2を変化させたときの評価(A),(B)の検索精度を表 5に示す。α_1を変化させたときとは対照的に、α_2を大きくした場合に適合率の低下がみられ、α_2を小さくした場合に適合率の向上がみられる。検索語「企業」は「企業イメージ」「多国籍企業」「企業責任」などシソーラス掲載の77語に部分一致している語であり、シソーラス語が属している分類も24個の小分類と多岐にわたっている。

このように、部分一致している検索語はシソーラス掲載語の多くに一致することが多く、検索語としては抽象度が高いため、検索語としての重要度を低くしたほうが適合率の向上が得られたと考えられる。これを確認するため、シソーラス情報に関する式(8')を、検索語$word_i$が一致するシソーラスの小分類のカテゴリ頻度cf_iにより決定するよう式(8"),(11)を定義し、評価を試みた。

表 3 検索精度の測定結果
（完全一致の検索語C_i=1の重要度変更）

再現率 (%)	適合率 (%)			
	(A) 単語頻度情報のみ α_1=1.0	(B) シソーラス情報利用 α_2=1.0 固定		
		α_1=2.0	α_1=1.5	α_1=1.2
0	75.7	79.5 (+5.0%)	77.5	76.2
10	69.4	71.7 (+3.3%)	70.9	70.1
20	61.9	64.2 (+3.7%)	63.9	63.3
30	50.5	53.9 (+6.7%)	53.9	52.2
40	45.4	46.9 (+3.3%)	47.9	47.4
50	42.8	44.2(+3.3%)	44.8	44.6
60	31.1	31.3(+0.6%)	31.7	32.1
70	27.0	29.3(+8.5%)	29.3	28.8
80	23.5	25.7(+9.4%)	25.6	25.4
90	20.3	22.1(+8.7%)	22.1	22.0
100	16.6	18.6(+12.0%)	18.6	18.3
平均	42.2	44.3(+5.0%)	44.2	43.7

()内の値は評価(A)と比較したときの適合率の変化を示す

表 4 適合率変化の大きい検索要求

検索語	適合率の平均(%)		適合率の差(%)
	評価(A)	評価(B)	[変化率]
*第三セクター,事業,運営	84.9	100.0	+15.2 [+17.9%]
企業,*情報,共有,導入,事例	6.4	20.7	+14.3 [+225%]
*半導体,製品,生産	49.4	58.8	+9.4 [+19.1%]
異,業種,会社,共同,*経営	61.0	55.7	-5.4 [-8.8%]
*リエンジニアリング,リストラ,定義	70.7	65.6	-5.0 [-7.1%]

(* はC_i=1の検索語を示す。)

表 5 検索精度の測定結果
（部分一致の検索語C_i=2の重要度変更）

再現率 (%)	適合率 (%)			
	(A) 単語頻度情報のみ	(B) シソーラス情報利用 α_1=1.0 固定		
		α_2=1.5	α_2=0.8	α_2=0.5
0	75.7	69.6	75.9	78.2
10	69.4	63.4	69.9	71.6
20	61.9	58.5	62.3	63.4
30	50.5	44.8	52.7	53.4
40	45.4	42.0	47.6	46.8
50	42.8	39.4	44.7	44.2
60	31.1	26.8	32.8	32.5
70	27.0	23.0	29.2	30.3
80	23.5	20.6	25.8	26.6
90	20.3	18.0	22.3	22.7
100	16.6	14.7	18.9	19.2
平均	42.2	38.2	43.8	44.5

$$score^{D_i} = \sum_{k=1}^{M}\left[\rho(tf_k^{D_i}) \times \sigma(df_k) \times \psi'(cf_k)\right] \quad (8'')$$

$$\psi'(cf_k) = \left(\frac{\beta}{cf_k}\right)^{\gamma} \quad (\beta, \gamma \text{ は 定数}) \quad (11)$$

cf_kがβより大きいときにスコアを大きくし、小さいときスコアを小さくするものである。式(11)で$\beta=2.6, \gamma=0.3$としたとき、表6に示すような高い検索精度が得られた。検索語の分類C_iごとにcf_iの平均をみると、$C_i=1,2$でそれぞれ4.4, 9.8分類である。検索語が一致するシソーラスの分類頻度も重要な要素であることがわかる。

また、検索語の分類C_iごとに文書出現頻度(DF)の平均を比較すると、$C_i=0,1,2$でそれぞれ4699, 4083, 10893である。$C_i=2$の場合、明らかに検索語自体の文書出現頻度が大きく、単語頻度情報とシソーラス掲載情報の相関が高い可能性がある。しかし、$C_i=0$と$C_i=1$の単語出現頻度の差は小さく、単語頻度情報とは別の要素としてシソーラス掲載情報の重要性があるとみられる。

なお、検索語がシソーラス掲載語に完全一致する$C_i=1$の場合にα_1を大きくし、部分一致する$C_i=2$の場合にα_2を小さくした場合も表6に示すように高い検索精度が得られた。

6 おわりに

本論文では、文書スコアリングの精度向上のため、シソーラス掲載情報を用いたスコアリング手法を検討した。本手法では、文書の重要度を決定するため、従来の文書内出現頻度(TF)や文書出現頻度(DF)のほかに、検索語とシソーラス掲載語の文字列一致度合や検索文字列が含まれるシソーラス掲載語のカテゴリ頻度により文書の重要度を変更した。また、単語頻度情報を用いたスコアリング手法と組み合わせることにより、検索精度の向上がはかれることを日本語新聞記事テストコレクション(BMIR-J1) によって示した。

本手法では、シソーラス掲載情報を利用したが、通常、シソーラス語は階層関係をもっている。上位語にあたるものは、下位の語を一般的に言い表す語であり、抽象的な語であることが多い。一方、下位の語は具体的であり、検索語としての重要性は高いと考えられる。この性質を利用してシソーラス階層情報による検索語の重要度を変更する方法についても今後検討を進めていきたい。

表 6 検索精度の測定結果 (シソーラス分類頻度利用、検索語C_i = 1,2の重要度変更)

再現率 (%)	適合率 (%) (A) 単語頻度情報のみ	シソーラス分類頻度利用 $\beta=2.6, \gamma=0.3$	(B) シソーラス情報利用 $\alpha_1=1.5, \alpha_2=0.8$
0	75.7	81.3	79.5
10	69.4	72.6	71.7
20	61.9	62.9	64.1
30	50.5	51.3	53.7
40	45.4	44.9	46.7
50	42.8	41.9	44.3
60	31.1	32.8	32.1
70	27.0	28.5	30.0
80	23.5	24.8	26.4
90	20.3	20.9	22.7
100	16.6	17.2	19.2
平均	42.2	43.6	44.6

参考文献

[1] D.Harman editor: The 3rd Text Retrieval Conference (TREC-3), National Institute of Standards and Technology, 1995.
http://potomac.ncsl.nist.gov/TREC/

[2] 高木徹,木谷強: 単語出現共起関係を用いた文書重要度付与の検討, 情報処理学会研究会報告, Vol. FI 41-8, pp.61-68 1996.

[3] 大井耕三,隅田英一郎,飯田仁: 単語間の意味的類似度に基づく文書検索手法, 言語処理学会第 2 回年次大会発表論文集, pp.109-112,1996

[4] 日本経済新聞社データバンク局: 日経シソーラス, 日本経済新聞社, 1995.

[5] G. Salton and M. J. McGill: Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[6] 芥子育雄ほか: 情報検索システム評価用ベンチマーク Ver1.0 (BMIR-J1) について, 情報処理学会研究会報告, Vol. DBS106, pp.139-145, 1996.

[7] 松本裕治,北内啓,山下達雄,平野善隆,今一修,今村友明: 日本語形態素解析システム「茶筌」version1.0 使用説明書, NAIST Technical Report, NAIST-IS-TR97007, 1997.

[8] C. Buckley, A. Singhal, M. Mitra and G. Salton: New Retrieval Approaches Using SMART: TREC4, Proc. of The 4th Text Retrieval Conference (TREC-4), 1995.