情 報 学 基 礎 44－5
（1996．11．21）

単語統計情報と言語情報とを併用した新しい文書検索のモデル

野口直彦　稲葉光昭　野本昌子　菅野祐司

{ noguchi,inaba,nomoto,kanno }@trl.mei.co.jp

松下電器産業(株) マルチメディアシステム研究所

〒140 東京都品川区東品川 4－5－15

近年、WWW上での情報検索システムに代表されるように、部分一致(partial match)モデルに基づく情報検索システムが実用化段階を迎えている。しかし、新聞検索・特許検索・文献検索など、比較的均一で大規模な情報に対する検索については、ランキングの適合率向上が必須である。

本稿では、部分一致モデルに基づく検索システムの適合率の向上を目的として、単語統計情報と言語情報とを併用した新しい検索モデルを提案する。我々の提案する検索モデルでは、文書／検索質問の特徴量の表現、およびそれらの照合過程に、人間の文書検索過程についての内省から得られる知見を反映させる。

また、我々は、新聞記事データを対象として、その検索モデルを実現する実験システムを構築した。実装においては、将来の実用化を意識し、索引作成速度／索引容量／検索速度の点で十分実用的なものになるように工夫した。本稿では、その実験システムの概要と、それを用いて行なった適合率の評価実験についても述べる。実験では、検索条件を注意深く入力することで、ランキングの上位において適合率が改善される可能性があることが確認された。

A New IR Model Utilizing Word Statistics and Linguistic Information

Naohiko Noguchi　Mitsuaki Inaba　Masako Nomoto　Yuji Kanno

{ noguchi,inaba,nomoto,kanno }@trl.mei.co.jp

Mutlimedia Systems Research Laboratory
Matsushita Electric Industrial Co., Ltd.

4－5－15 Higashi-Shinagawa, Shinagawa-ku, Tokyo 140, JAPAN

Information retrieval systems based on the partial match IR models are now in commercial use, especially in the field of World Wide Web search systems. In order to apply such IR systems to large and homogeneous data such as newspaper articles and patent documents, it is necessary to improve the effectiveness of the document ranking they produce.

In this paper, we propose a new IR model which utilizes the linguistic information as well as word statistics information to improve its effectiveness. To conceptualize the model, we examine and analyze the human process of selecting relevant documents.

We have developed a search system for newspaper articles realizing proposed IR model. Although it is an experimental system, it can index and search giga-byte data in practical computational time and spaces. We address the implementation of the system briefly and show the results of some experiments on its retrieval effectiveness. The results show that it can improve the precision of the upper part of the document ranking when the queries are deliberately constructed.

1 はじめに

近年、ベクトル空間モデル、確率モデルなどの非完全一致(inexact match)、部分一致(partial match)モデルに基づく情報検索システムが実用段階を迎えている。特に、WWW上での情報検索システムに代表されるように、計算機ネットワーク上に分散して存在する異種／雑多な文書情報を検索する際には、従来の書誌情報に基づいた検索や、キーワード／全文検索などの完全一致(exact match)モデルに基づく厳格なシステムでは柔軟性に欠けるため、利用者からは簡単な入力(文章やキーワードセットなど)を受けとり、その入力に適合する順に検索結果を出力するランキング機能を備えたシステムが標準的になってきている。(Excite[5], WebCrawler[6], Lycos[7]等々)

WWW上での情報検索の場合は、利用者の欲する情報がどこにあるのかが分からないので、その情報源へのきっかけをつかみたい、といった検索要求が主となる。従って、検索結果の上位にそのような情報がいくつか存在すれば要求がほぼ満たされる。つまり、再現率よりも、上位数件(あるいは数十件)での適合率が重視される。一方、新聞検索・文献検索・特許検索など、比較的均一で大規模な文書データから、ある主題についての情報を半ば網羅的に検索したいといった、再現率／適合率両面の総合的な検索性能(effectiveness)が重視されるタイプの検索要求に対しては、いまだにキーワード検索／全文検索といったブール型のシステムが主流である。これは、ブール型モデルで検索される文書集合が利用者に明確に把握できるのに対して、部分一致モデルで検索される文書集合およびランキング結果は、それほど明確には把握できないこと、また、部分一致モデルでの検索性能(再現率／適合率)がそれほど高くないという理由による。

本稿では、そのような部分一致モデルにおけるランキングの適合率を向上させることを目的として、単語統計情報と言語情報とを併用した新たな検索モデルを提案する。我々の提案する検索モデルでは、文書／検索質問の特徴量の表現、およびそれらの照合過程に、人間の文書検索過程についての内省から得られる知見を反映させる。

本稿の構成は以下の通りである。2では、関連研究について述べ、本研究の位置付けを行なう。次に、3で、本研究で提案する検索モデルの基本的アイデアについて述べ、4にて、提案する検索モデルの詳細について述べる。5では、そのモデルを適用した実験システムの構成と、実装上の工夫について述べる。6では、実験システムを用いて行なったランキングの適合率に関する実験結果について述べ、7で結論を述べる。

2 関連研究

部分一致モデルでは、文書と利用者の検索質問から特徴量を抽出し、それらを何らかの類似尺度に基づいて比較照合して、検索質問に適合する順番に文書のランキングを行なう[3], [12]。この検索モデルは、文書／検索質問の特徴表現と、それらの照合手法までも含んだモデルであるが、その特徴表現の基礎となる理論として、単語の重み付け理論がある[16]。単語の重み付け理論とは、文書や検索質問を構成する文に出現する単語に注目し、各単語の、各文書／各検索質問に対する重要度を表す指標(重み)を計算して、文書／検索質問の特徴量とするものであり、その計算には、単語の出現頻度や文書全体での出現分布など、単語に関する統計量を通常用いる。単語重み付けの手法は様々であるが、$tf \cdot idf$[14]と呼ばれる重み付けがその代表例である。これは、以下のような式に従って、単語に重み付けするものである。ある単語t_iの、文書d_jにおける出現頻度を$FREQ_{t_i,d_j}$とし、全文書数N、t_iが出現する文書数を$DFREQ_{t_i}$とした時、t_iの文書d_jにおける$tf \cdot idf$による重み$TFIDF(t_i, d_j)$は、

$$TFIDF(t_i, d_j) = FREQ_{t_i,d_j} \cdot (1 + \log_2 \frac{N}{DFREQ_{t_i}})$$

で与えられる。

しかし、文書／検索質問の特徴量を、そのような単語レベルの統計量だけで表現することには限界がある。利用者の検索ニーズは大抵の場合非常に複雑であり、単に単語の集合、あるいは各単語の重みからなる多次元ベクトルだけでは表現しきれない。また、単語の統計量が、文書／検索質問の特徴を的確に表現している保証もない。そのような反省から、部分一致モデルにおいてさらに精密な検索を行なうために、以下のような研究が行なわれてきている。

検索モデルを洗練するアプローチ
: 上記したような、部分一致に基づく検索モデルそのものを洗練していこうとする理論的なアプローチ。通常のブール論理式の解釈を多値的な解釈に緩めていこうとする拡張ブール型モデル、また、ナイーブなベクトル空間モデルから、各索引語間の相関性を排除した特徴空間へと変換を行なう一般化ベクトル空間モデル、また、文書の検索質問に対する関連度を、当該文書から検索質問が推論できるかどうかの確からしさによって判断しようとする確率的推論モデルなどがある[4], [15]。

特徴量を拡充するアプローチ
: 上記の検索モデルとも関連するが、文書／検索質問の特徴量を精密かつ豊富にしようとするアプローチがある。例えば、単語単独でなく、単語のなんらかの共起関係を特徴量として導入する、文書の構造情報／参照情報より特徴ベクトルを拡充する、あるいは精密な言語解析により、構文的な情報や意味的な情報から特徴を抽出するなど、さまざまな試みが行なわれている。言語情報を特徴量として導入しようとするアプローチのサーベイに[13]がある。

本研究でも、これらのアプローチに従い、文書／検索質問からいかなる特徴を抽出するかということと、その特徴を利用して、いかなる比較照合処理を行なうのか、という二面から、新たな検索モデルを考案する。

3 基本的アイデア

3.1 文書／検索質問の特徴量

まず、文書／検索質問から抽出する特徴量について考える。そもそも、従来の部分一致モデルで利用されてきた単語統計情報は、検索対象となる文書中で、ある文書を他の文書と区別するための情報としては重要であるが、それは、利用者の検索ニーズを直接表現するものにはなりえない。なぜなら、利用者が検索を行う際に、その検索対象文書がどのような単語頻度、単語分布を持っているかということは、そもそも意

識にはのぼらないからである。従って、利用者の検索ニーズとの適合性を判断するためには、その検索ニーズをもっと直接的に表現できる情報を利用する必要があるだろう。

例えば、単語同士の共起情報は、単なる一単語の出現よりも特定的 (specific) な情報であり、利用者の検索ニーズを表現するのにより適切な情報だと考えられる。ただ、共起情報と言っても、文書内での共起、段落内での共起、文内での共起など、様々なレベルが考えられる。現在までに、共起情報を特徴量として採り入れようとする試みも多々あるが、文書内共起か、あるいは二単語間の距離をパラメータとして、ある距離内での共起 (近接共起) を共起関係として用いるものが殆んどである。しかし、利用者の検索ニーズを直接的に表現していると考えられるのは、ある構文要素内で係り受けのある共起関係である。例えば、「文書検索」に関する文書を検索したい場合、「文書を検索する...」という一節が文書内にあれば、それだけで検索ニーズとの関連性が判断できる場合もある。ここでは、「文書」と「検索」とが、互いに係り受けの関係を持って共起しているということが重要な情報になっている。

そこで、本稿では、文書／検索質問から抽出する特徴として、次の二通りの単語共起関係を考える。

- 文書内共起
- 構文要素内共起

前者は、検索質問中に与えられる検索語のうちのどれだけが文書内に共起しているかということを表すものであり、後者は、上記したように、ある特定の構文要素の中で、係り受け関係を持つものとして定義される共起関係である。

また、特徴量として単語の統計情報以外の情報を採り入れようとする従来の試みの中には、その新たな情報を、単語頻度情報と同列の、例えば特徴ベクトルの枠組みの中に位置付けている例が多く見られる。照合過程の簡素化のためには、全ての特徴量を同一のレベルの表現形式で表わした方が都合がよいことは確かであるが、そこに力点を置くあまり、そのような異種のレベルの特徴量が持つ意味への配慮がいささか欠けていると思われる。

文書が検索ニーズに合致していたかどうかを決定するのは、最終的には、その検索ニーズを持っている利用者であり、利用者の検索ニーズを直接的に表現するための第一次近似として上記の共起関係を考えたわけであるから、それは、利用者が文書の関連性をどのように判断しているかという処理に即して利用されるべきであろう。つまり、利用者の関連文書選択過程を考察、分析することによって、新たな特徴量の、比較照合処理への利用法が明らかになると考える。

以上、本稿で提案する検索モデルの基本的なアイデアを、以下にまとめておく。

(1) 単語レベルの統計情報以外に、利用者の検索ニーズを直接的に表現しうるような言語的情報 (文書内共起、構文要素内共起) を特徴量として採用する。

(2) 言語的情報については、単語統計情報に基づく特徴表現と同列には扱わず、その特徴量に合った処理を考える。

(3) 比較照合処理のモデルについては、人間の検索活動 (特に関連文書の選択) 過程を反映させる。

3.2 人間の関連文書選択過程について

ここでは、上記の方針に従って、人間が文書の関連性を判断する処理を、特許検索を例にとって考察してみよう。

通常、特許検索においては、最初に適当な検索条件を与えて特許の母集団を求め、その中から真に関連する特許を選択する (絞り込む) という活動を繰り返す。その初期の段階では、検索条件を精密化して母集団のサイズを落とすために上の過程を繰り返すが、最終段階では、ある程度のサイズの母集団の全数チェックを行って、関連特許を漏れなく抜き出す、ということが行われる。

関連特許を抜き出す際に、最初から特許の明細書を読んで理解してから判断する、ということはまれである。むしろ、最初は「発明の名称」や「抄録」など、その特許が何に関するものであるかが端的に表現されている部分だけを読み、そこに自分の検索ニーズに近い表現があればそれだけで関連特許であると判断する、というようなことを行なっている。もちろん、それだけで判断できない場合もあるので、その場合には、明細書の内容をもう少し読んで判断することを試みる。例えば、「産業上の利用分野」や、「従来技術」といった、重要な部分を読んでみる。明細書の全文を読んで内容を理解し、それで関連文書かどうかを判断するという行動は、いわば最後の手段である。

つまり、人間は、最初から特許の全文を読んで内容を理解してから関連特許かどうかの判断をするのではなく、特許文書の中から重要な部分を選択的に、かつ段階的に選んで読み進み、その中に関連特許の証拠となりそうな特定的な情報を見つけた時点で関連特許であると判断する、というような処理を行っていると考えられる。

以上の処理をまとめると、

(A) 文書中の重要部分を、選択的かつ段階的に読み進む。

(B) 各段階で、関連文書／非関連特許の証拠となる特定的な情報を発見したところで判断して処理を中止する。

(C) そのような証拠が見つからない時に初めて、全文を読んで関連文書かどうかを判定する。

本稿では、このような処理をなるべく忠実に再現する検索モデルを考える。ここで、(B) で用いる、関連文書かどうかの証拠となりうる「特定的な情報」に、前節で考えた共起関係 (文書内共起関係、構文要素内共起関係) を採用することにする。また、(C) で最終的に内容を読んで判断する処理では、全文から抽出した単語統計情報から得られる特徴量を用いることにする。

4 検索モデル

本節では、3で述べたアイデアを実現する検索モデルについて述べる。

4.1 文書と検索質問の特徴表現

文書／検索質問の特徴表現は、共に、特徴ベクトル $\vec{V}$ と出現単語集合 T、共起関係集合 C の3つ組 $(\vec{V}, T, C)$ から構成されるものとする。

4.1.1 文書の特徴表現

文書 d_j は、特徴ベクトル $D\vec{V}_j$、文書内出現単語集合 DT_j、文書内共起関係集合 DC_j からなる3つ組

$$(D\vec{V}_j, DT_j, DC_j)$$

で表現する。

$$DT_j = t_1, t_2, \ldots, t_n$$

とした時、$D\vec{V}_j$ は以下のように与えられる。

$$\begin{aligned} D\vec{V}_j &= (W_{t_1}, W_{t_2}, \ldots, W_{t_n}) \\ W_{t_i} &= TFIDF(t_i, d_j) \end{aligned}$$

また、DC_j は、文書 d_j から抽出された構文要素内共起関係の集合であり、その抽出方法については後述する。

4.1.2 検索質問の特徴表現

検索質問 q は、特徴ベクトル $Q\vec{V}$、質問内出現単語集合 QT、質問内共起関係集合 QC からなる3つ組

$$(Q\vec{V}, QT, QC)$$

で表現する。

$$QT = q_1, q_2, \ldots, q_m$$

とした時、$Q\vec{V}$ は以下のように与えられる。

$$\begin{aligned} Q\vec{V} &= (W_{q_1}, W_{q_2}, \ldots, W_{q_m}) \\ W_{q_i} &= 1 \ for \ \forall i \end{aligned}$$

また、QC は、検索質問 q から抽出された構文要素内共起関係の集合である。

4.2 類似度の計算

4.2.1 単語統計情報に基づく類似度基準

文書／検索質問の特徴表現の第1項である特徴ベクトル $D\vec{V}_j$、$Q\vec{V}$ を用い、通常のベクトル空間法での内積尺度 [3] によって、文書 d_j と検索質問 q の類似度基準を以下のように与える。

$$SIM_v(q, d_j) = D\vec{V}_j \cdot Q\vec{V}$$

ここで、・は2つのベクトル間の内積計算を表わす。4.1.2より、

$$Q\vec{V} = (W_{q_1}, W_{q_2}, \ldots, W_{q_m}) = (1, 1, \ldots)$$

であるから、

$$\begin{aligned} SIM_v(q, d_j) &= \sum_{i=1}^{m} TFIDF(q_i, d_j) \\ &= \sum_{i=1}^{m} FREQ_{q_i, d_j} \cdot (1 + \log_2 \frac{N}{DFREQ_{t_i}}) \end{aligned}$$

となる。

4.2.2 文書内共起関係に基づく類似度基準

QT、DT_j を上記の通りとした時、文書内共起関係に基づく類似度基準 SIM_L は、以下のように与えられる。

$$SIM_L(q, d_j) = \mid QT \cap DT_j \mid$$

ただし、$\mid S \mid$ は、集合 S の要素数を表す。この類似度基準の意味は、検索語 QT として与えられたもののうち、より多くの語が出現している文書の類似度を高く見るものである。

4.2.3 構文要素内共起関係に基づく類似度基準

QC、DC_j を上記の通りとした時、構文要素内共起関係に基づく類似度基準 SIM_C は、以下のように与えられる。

$$SIM_C(q, d_j) = \mid QC \cap DC_j \mid$$

これは、検索共起関係 QC として与えられたもののうち、文書内の構文要素内共起関係 DC_j として出現しているものがいくつあるかということによって類似度を判断するものである。

4.2.4 類似度基準の優先度

3では、人間の関連文書選択処理を考察した。ここでは、本節で与えた、文書／検索質問の特徴表現と、それらを用いた類似度基準で、どのように人間の処理過程が反映できるかを考えてみよう。

まず、上で考えた特徴量のうち、構文要素内共起関係は、最も特定的な情報であり、利用者の検索ニーズを直接的に表現しうるものである。QC がそのようなニーズを表現しているとすると、文書 d_j の重要部分を選択的に読み進む過程で、QC の要素に一致するような DC_j の要素を発見した段階で、その文書 d_j が関連文書であると判断するものと考えられる。従って、文書の重要部分から構文要素内共起関係 DC_j を抽出しておき、検索時には、検索質問から抽出された QC との照合を類似度基準 SIM_C に従って行なうことで、人間の処理を近似することができる。

また、文書内共起関係に基づく類似度基準 SIM_L は、各文書内で、検索語 QT の要素が共起すればするほど、類似度が高まるというものである。これは、ベクトル空間モデルの枠組では、$D\vec{V} = (W_{t_1}, W_{t_2}, \ldots) = (1, 1, \ldots)$ として、$Q\vec{V}$ と内積計算をした値であると見ることができるが、一方、確率的推論モデル (probabilistic inference model)[17] の概念化によれば、文書／検索質問に出現する単語を命題と見た時の、古典論理における論理的含意の確からしさの度合と見ることもできる。つまり、そのような見方をすれば、SIM_L は、人間の論理的思考を反映する類似度基準であり、SIM_C ほど明確ではないが、SIM_V よりは利用者の検索ニーズの表現に近いものであると考えられる。

類似度基準 SIM_V は、文書中の単語統計情報に基づいており、検索対象文書集合全体の中での、文書の特徴を表現している。つまり、利用者の検索ニーズの表現からは最も遠い。

以上の考察より、上記した三種類の類似度の間には、次のような優先度を設定け、段階的に適用してランキングを求めることとする。

$$SIM_C > SIM_L > SIM_V$$

SIM_V よりは SIM_L の方を優先的に、また、SIM_L よりは SIM_C の方を優先的に適用する。この順序は、それぞれの類似度基準が用いている情報が、どれだけ利用者の検索ニーズを直接的に表現し得るかという尺度によっている。

5 実験システムの構築

新聞記事データを対象として実験システムを構築し、4で述べた検索モデルの性能評価を行なった。ここでは、検索モデルの評価が第一の目的であるが、将来の実用化を意識し、以下の要件を満たすように実験システムを構築した。

- 大規模な文書に対応可能であること。
- 実用的な速度で前処理 (索引の作成) を行なうことができ、また、索引容量はコンパクトであること。
- 実用的な速度で検索できること。

現在、新聞記事データ約 840MB に対して、索引作成速度 4 ～ 5 時間程度、索引容量は 650MB 程度、また検索速度が 2 ～ 3 秒程度の実験システムが、SUN 社製ワークステーション ULTRA1(CPU UltraSparc 167MHz, Memory 448MB) 上で稼働している。

なお、現在索引容量の圧縮および索引作成の高速化についてさらに取り組んでおり、頻度索引に限れば、同じ 840MB のデータに対して、索引容量約 200MB、作成時間 1 時間以内を実現している。

5.1 構成

構築した実験システムの構成を図 1に示す。検索に先だって、辞書を用いて文書から単語を抽出し、その文書内頻度、出現文書数を計測して頻度索引として構成しておく。また、各文書中で、ある特定の構文要素内に共起する二単語を抽出し、その情報を共起索引として格納しておく。今回、辞書は、ＥＤＲ辞書 [2] の見出し文字列集合から本実験システム用に変換したものを用いた。

利用者が質問文、あるいはキーワード列で検索質問を与えると、類似度判定処理部は、まず、文書から単語頻度、共起関係を抽出した手法と同様な手法で検索質問を解析して、検索語集合 QT、検索共起関係 QC を求める。その後、4に記載したような類似度基準に従って文書をランキングして、出力する。その際に、全文検索エンジンを併用し、辞書に単語として登録されていない文字列を入力された際にも、頻度情報が正確に求められるようにした。

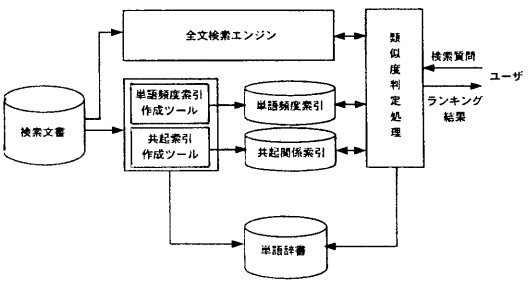

図 1: 実験システムの構成

以下に、実験システムを実装する上で工夫した点を述べる。

5.2 単語統計情報の抽出

日本語などの膠着言語で書かれた文書の場合、単語頻度を抽出するためには形態素解析などの処理を行なって、文書を単語分割する必要がある。しかし、形態素解析を行なっても完全に単語分割できるわけではなく、その精度に依存して、検索もれが起きたり、頻度計測が不正確になってランキング精度が悪化したりする。また、形態素解析はコストのかかる処理であり、今回対象とするような大規模なデータに対しては実用的でない。

我々は、[9],[10] において、形態素解析処理を行なわずに単語索引を構成し、それを用いて高速に全文検索を行なう手法 (完全索引方式／完全延長極大索引方式) を提案した。この手法は、前処理は高速で、構成する索引量もコンパクトであり、また、単語索引を構成するので文字列レベルの全文検索も高速に行なえるものである。今回の実験システムでは、任意文字列検索の機能は必ずしも必要ではないが、上記した方式を用いれば、単語頻度情報を計測して索引を高速に構成できる。本実験システムでは、完全索引方式／完全延長極大索引方式を用いて、頻度索引を構成した。その索引作成方式の詳細については別稿に譲る。

5.3 構文要素内共起関係の抽出

利用者の検索ニーズを直接的に表現できるような構文要素として、今回は以下のようなものを対象とした。

- 「の」関係

  例) 「野茂の大リーグ初勝利」
  → (野茂, 大リーグ),(野茂, 初勝利)

- 名詞連続関係

  例) 「野茂の大リーグ初勝利」
  → (大リーグ, 初勝利)

- 格関係 (A { が | を | に }... Bする)

  例) 「野茂が大リーグに挑戦する」
  → (野茂, 挑戦する),(大リーグ, 挑戦する)

3.2での考察によれば、人間は、文書の関連性を判断する時に、重要だと思われる部分を選択的に読み進む。この処理を反映させるためには、新聞記事の重要部分として、見出し、あるいは本文の先頭から数文だけに限定して共起関係を抽出する、といった方法も考えられる。(今回の実験では、見出し／本文全体を対象として、共起関係の抽出を行なった。)

また、ここでも、形態素解析／構文解析など、コストのかかる処理は行なわず、文字種区切りなどのヒューリスティクスと辞書引きのみで抽出処理を高速に行った。(処理の詳細については省略する)

5.4 未知語への対応

実験システムでは、まず利用者からの入力を解析して、検索条件 $(\vec{QV}, QT, QC)$ を抽出する。その際に、入力された単語が辞書に登録されていない (未知語) 場合がある。そのような語は無視するという方法も考えられるが、それではランキングの精度が落ちてしまう。そこで、本システムでは、図1に示したように、全文検索エンジンを併用して、未知語の入力に対しても動的にその頻度を求められるようにしてある。

5.5 新聞記事データへの対応

新聞記事データは、書誌事項／見出し／本文／キーワードなどいくつかのフィールドから構成されている。通常、見出しはその記事の内容を端的に表現しており、本文に比べて重要度が高いものと考えられる。そこで、本実験システムにおいては、見出しと本文のフィールド毎に頻度索引を構成して、DV_j の計算を以下のように修正することとした。

$$\vec{DV_j} = (W_{t_1}, W_{t_2}, \ldots, W_{t_n})$$
$$W_{t_i} = C \cdot TFIDF_m(t_i, d_j) + (1-C) \cdot TFIDF_h(t_i, d_j)$$

ただし、

$$TFIDF_m(t_i, d_j) = \frac{MFREQ_{t_i,d_j}}{TOTFREQ_{t_i}} \cdot (1 + \log_2 \frac{N}{DFREQ_{t_i}})$$

$$TFIDF_h(t_i, d_j) = \frac{HFREQ_{t_i,d_j}}{TOTFREQ_{t_i}} \cdot (1 + \log_2 \frac{N}{DFREQ_{t_i}})$$

ここで、$MFREQ_{t_i,d_j}$ は単語 t_i の文書 d_j の見出しにおける頻度、$HFREQ_{t_i,d_j}$ は単語 t_i の文書 d_j の本文における頻度、C は、見出しと本文の重要度の割合を決めるパラメータである。C は、必要に応じて利用者が設定できるようにした[1]。また、類似度基準 SIM_V が高頻度語に引きずられるのを防ぐために、$TFIDF$ の頻度要因は、単語 t_i の文書全体における頻度である $TOTFREQ_{t_i}$ で正規化を行なった。

6 実験結果

6.1 実験方法

実験には、(社) 情報処理学会から提供されている、情報検索システム評価用ベンチマーク Ver 1.0[8](以降、「ベンチマークデータ」と呼ぶ) を用いた。これは、情報処理学会データベース研究会の下部組織「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」により構築されたものである。Ver 1.0 はその試用版であり、日本経済新聞朝刊、経済面の 600 記事と、それに対する検索要求 60 問、ならびに各検索要求に対する正解記事集合とからなる。以下の各実験においては、それぞれの目的に応じて、このベンチマークデータの検索要求 60 問から適宜選択して評価用データセットを構築した。

実験は、本稿で提案している新たな類似度基準 SIM_L と SIM_C の導入効果を測定するために、SIM_V の類似度基準だけを用いたランキングの適合率と、SIM_L、SIM_C を併用した場合のランキングの適合率を比較する形で行なった。ランキングの適合率は、以下の四つの場合に分けて求めた。

(1) 類似度基準 SIM_V のみを用いるもの

(2) 類似度基準 SIM_V、SIM_L を用いるもの

(3) 類似度基準 SIM_V、SIM_C を用いるもの

(4) 類似度基準 SIM_V、SIM_L、SIM_C を用いるもの

類似度基準を併用してランキングを行なう場合 ((2), (3), (4))、これらの優先度は 4.2.4で述べたように、

$$SIM_C > SIM_L > SIM_V$$

であるから、優先度の高い類似度基準を先に適用して検索結果を階層化し、次に優先度の高い類似度基準を用いて各階層を細分化する、といったランキング (層別ランキング) が行なわれることになる。

また、全実験を通じて、適合率の評価は以下のように行なった。まず、評価データの各質問について検索実験を行ない、各質問に対する正解集合と比較して、再現率 – 適合率グラフで評価を行なった。グラフには、再現率が変化したポイントだけをプロットし、鋸状になる部分は直線で近似した。さらに、各質問に対応する再現率 – 適合率グラフから、再現率が 0.0 から 1.0 までの 0.1 刻みでの 11 点に対応した適合率を求め、それらを評価データの全質問にわたって平均して、全体の再現率 – 適合率グラフを構成した。ただし、各質問、各点での適合率は、各質問に対応するグラフを階段状に平滑化して求めた。

6.2 文章入力による実験

まず、ベンチマークデータの検索要求文を検索条件としてそのまま用いて実験を行なった。図 2 に、ベンチマークデータ 60 問全問の平均からなる再現率 – 適合率グラフを示す。グラフ中には、類似度基準の利用に関して上記した四通りの場合の曲線がプロットしてある。このグラフでは、以上四本の曲線の間に差はほとんど見られなかった。

ベンチマークデータでは、正解数が 5 以上 30 以下の質問を用いることを推奨している。そのような質問は、60 問中 47 問である。また、ベンチマークデータの検索要求文は簡潔なものが多く、一単語からなるものも多い。本実験システムでは、検索質問から構文要素内共起関係を自動的に抽出するが、そのような一単語からなる検索質問では、構文要素共起

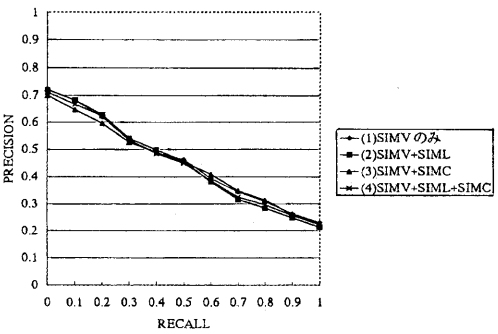

図 2: 再現率 – 適合率グラフ (文章入力,60 問の平均)

[1]今回の実験では、$C = 0.5$ とした。

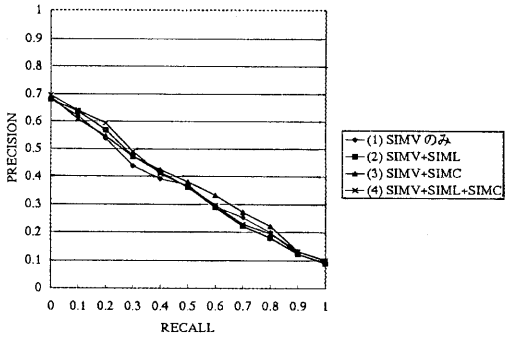

図 3: 再現率 – 適合率グラフ (文章入力,34 問の平均)

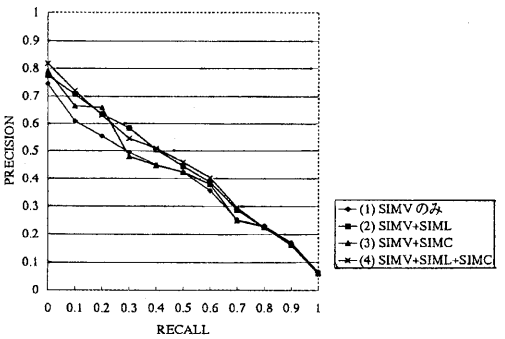

図 4: 再現率 – 適合率グラフ (統制入力,19 問の平均)

関係を抽出することは不可能である。上の 47 質問の検索要求文を調査したところ、本実験システムによって一つでも構文要素内共起関係が抽出できたものは 41 質問であった。また、構文要素内共起関係が抽出されたとしても、それが文書中に一度も出現しないような共起関係である場合には、類似度計算に反映されない。上の 41 質問の検索結果を調査したところ、文書中に一度は出現するような有効な構文要素内共起関係を抽出できているものは 34 質問であった。図 3は、その 34 質問について平均をとったものである。図 2と比較すると、SIM_L、SIM_C を導入した方が、SIM_V のみを用いる場合に比べて、僅かだが上回ってきていることが分かる。しかし、図3でも、SIM_L、SIM_C を併用したものが、SIM_V のみを用いたものを、全ての点において上回っているわけではなく、それらを導入する効果があるかどうかについては、はっきりとした結論は導けなかった。

以下に、各グラフにおける 11 ポイントによる平均適合率を示す。

60 問の平均	
(1)SIM_V のみ	0.504
(2)$SIM_V + SIM_L$	0.495
(3)$SIM_V + SIM_C$	0.497
(4)$SIM_V + SIM_L + SIM_C$	0.495
34 問の平均	
(1)SIM_V のみ	0.400
(2)$SIM_V + SIM_L$	0.404
(3)$SIM_V + SIM_C$	0.418
(4)$SIM_V + SIM_L + SIM_C$	0.414

6.3 統制入力による実験

6.2では、ベンチマークデータの検索要求文をそのまま検索条件として与え、本実験システムの入力解析部にて抽出された検索語と共起関係を修正することはしなかった。

しかし、抽出された検索語と共起関係が、必ずしも検索要求に合致しているとは言えないため、入力解析の精度が実験結果に影響を及ぼしていると考えられる。従って、共起関係に基づく類似度基準導入の純粋な効果を測るためには、入力をある程度統制する必要がある。また、ベンチマークデータでは、対象とする検索システムの機能を、基本機能／数値・レンジ機能／構文解析機能／内容解析機能／知識処理機能の五つとし、60 問の検索要求文を、これらの機能により分類している。しかし、今回の実験システムでは、同義語辞書、シソーラスなどを利用しておらず、また、数値・レンジ機能については全く考慮していないため、それらの機能のみを測定するための検索要求に対しては、結果が悪化することが予想される。

以上のことを考慮に入れ、本実験では、次のような手順で実験を行なうこととした。

(1) 対象とする質問の選択
全 60 質問中、以下の基準により、実験対象とする質問を選択する。

(a) 同義語拡張機能が本質的な質問、および数値範囲指定機能が本質的な質問は除く

(b) 正解が 10 件以上 30 件以下
ベンチマークデータでは、正解数が 5 以上 30 以下の質問を用いることが推奨されているが、今回は、再現率が 0 から 1 までの 0.1 刻みでの 11 点についての平均適合率にて最終的な評価を行なうため、各質問についても、再現率 - 適合率グラフに 10 以上の点がプロットできるように、正解数が 10 件以上のものに限ることとした。

(2) 統制検索条件の決定
(1) で選択された各質問に対して、検索条件 (検索語の集合、共起関係の集合) を決定する。ただし、各質問に対して検索条件として与える検索語、共起関係については、各質問の検索要求文、ならびに付記されているコメントから想像できる範囲のものに限るものとする。(正解記事を見て条件を設定することはしない)

以上の基準により選択された質問は全部で 19 質問であった。統制入力条件の一例を以下に示す。

質問番号 25　：「メーカーの減益に対する対策」
質問内出現単語：メーカー, 減益, 対策, 業績悪化, 赤字, 減収, 業績, 悪化, リストラ
質問内共起関係：(メーカー, 減益),(メーカー, 対策), (減益, 対策)

実験結果を図 4に示す。これを見ると、6.2の結果に比べて、SIM_L、SIM_C の導入効果がはっきりしてきているこ

とがわかる。SIM_V のみを用いたものに比べて、SIM_L を導入したもの、SIM_L、SIM_C を両方導入したものは、再現率が 0 ～ 0.7 のポイントで適合率が上回っており、再現率が低いポイントほど、適合率の改善度合が高くなっている。特に、再現率が 0 ～ 0.5 の範囲では、SIM_L、SIM_C を両方導入したものは、SIM_V 単独のものよりも適合率で 0.03 ～ 0.1 程度上回っている。

一方、SIM_C を導入したものは、再現率 0 ～ 0.2 のポイントでは SIM_V のみを用いたものに比べて適合率が上回っているが、0.3 以降のポイントでは差が見られなかった。

11 ポイントによる平均適合率を、以下に示す。

19 問の平均

(1)SIM_V のみ	0.434
(2)$SIM_V + SIM_L$	0.477
(3)$SIM_V + SIM_C$	0.455
(4)$SIM_V + SIM_L + SIM_C$	0.483

6.4 考察

以上の実験結果より、類似度基準 SIM_L、SIM_C の導入効果については、次のようなことが言える。

- 単純な文章入力では、それらの導入効果が見られない。
- 統制入力では、SIM_L、SIM_C 導入の効果が見られた。また、それら二つを同時に用いた場合が最も適合率が改善された。また、特に、再現率が低い部分 (ランクの上位) について、適合率の改善が顕著に見られた。

従って、SIM_L、SIM_C の重要度基準を用いた場合、検索条件を注意深く入力することによって、適合率が改善される可能性があることが分かった。また、その際に、特にランキングの上位についての改善率が顕著になる傾向が見られることから、上位 10 件、20 件だけが問題になるような実用システムにおいては、より効果が大きくなることが期待できる。

7 おわりに

本稿では、単語統計情報と、言語情報とを併用した、新しい検索モデルを提案した。言語情報の特徴量としては、共起関係 (文書内共起関係、構文要素内共起関係) を用い、それを用いて、人間の関連文書選択過程を再現するように検索モデルを構成した。また、提案した検索モデルの評価と、将来の実用化を目的として実験システムを構築し、ベンチマークデータを利用して評価実験を行なった。

今回行なった実験は小規模なものであり、定性的な評価はできたが、言語情報を利用した類似度基準を導入した場合とそうでない場合とで、統計的に有意な差は見られなかった。今後、評価データを大規模化して、再実験を行なっていきたい。また、今回は新聞記事を対象として実験システムを構築したが、現在特許明細書データについても実験を行なっている。その一次評価は [11] にて報告したが、その後の結果についても別稿にて報告する予定である。

また、本実験で用いたベンチマークデータは、あくまで試用版であり、種々の課題があると思われるが、我々が実験を通じて強く感じたのは、以下のような課題である。

- 評価データの絶対量 (文書量、質問数) が少ないため、入力条件の違いによって、結果のゆれが大きくなる。
- 各質問に対して正解とされる記事の基準が時に曖昧であり、また、時に難しすぎる。

これらの点については、今後のベンチマークデータの修正・拡充を期待したい。

謝辞

本研究での実験システムの構築、ならびに評価実験において、日本経済新聞社データバンク局様にご協力をいただきました。ここに深く感謝いたします。

また、評価実験には、株式会社日本経済新聞の協力によって、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) を利用させていただきました。ここに感謝いたします。

参考文献

[1] Belkin, N.J. and Croft, W.B.: *Retrieval techniques*, Annual Review of Information Science and Technology, Vol.22, pp. 109–145 (1987).

[2] (株) 日本電子化辞書研究所: EDR 電子化辞書仕様説明書 (1993).

[3] Frakes, B. and Baeza-Yates, R. (editors): *Information Retrieval: Data Structures and Algorithms*, Prentice-Hall, Englewood Cliffs, New Jersey 07632 (1992).

[4] Fuhr, N. : *Probabilistic Models in Information Retrieval*, The Computer Journal, Vol. 35, No.3, pp.243–255 (1992).

[5] http://excite.com/

[6] http://webcrawler.com/

[7] http://www.lycos.com/

[8] 芥子育雄 他 : 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.139–146 (1996).

[9] 稲葉光昭, 野口直彦, 菅野祐司, 倉知一晃: 日本語文書に対する新しい索引検索方式 – 試作・実験および評価 –, 情報処理学会第 50 回 (平成 7 年前期) 全国大会予稿集, pp. 4-43 – 44 (1995).

[10] 倉知一晃, 野口直彦, 菅野祐司, 稲葉光昭: 日本語文書に対する新しい索引検索方式 – 索引作成と検索の原理 –, 情報処理学会第 50 回 (平成 7 年前期) 全国大会予稿集, pp. 4-41 – 42 (1995).

[11] 野本昌子, 野口直彦: 文書構造と共起表現を用いた文書ランキング手法, 情報処理学会第 52 回 (平成 8 年前期) 全国大会予稿集, pp. 4-203 – 204 (1996).

[12] Salton,G. and McGill,M.J.:*Introduction to Modern Information Retrieval*, McGraw-HillPublishing Company (1983).

[13] Smeaton, A.F.: *Progress in the Application of Natural Language Processing to Information Retrieval Tasks*, The Computer Journal, Vol.35, No.3, pp.268–277 (1992).

[14] Sparck-Jones, K.: *A Statistical Interpretation of Term Specificity and its Application in Retrieval*, Journal of Documentation, Vol.28, No.1, pp. 11–21 (1972).

[15] 谷口 祥一.: 80 年代における情報検索モデル研究の展開 : 文献レビュー, Library and Information Science, No.30, pp. 59 –76 (1992).

[16] 海野 敏.: 出現頻度情報に基づく単語重みづけの原理, Library and Information Science, No.26, pp.67 – 88 (1988).

[17] Wong,S.K.M. and Yao,Y.Y.: *On Modeling Information Retrieval with Probabilistic Inference*, ACM Transactions on Information Systems, Vol.13, No.1, pp. 38–68 (1995).