

SearchSpace: 2次元空間へのキーワード配置を用いる 全文検索システムのインタフェースと検索エンジン

堤 富士雄
(財)電力中央研究所

我々が開発した全文検索システム「SearchSpace」では、従来型の AND/OR 条件式ではなく、複数のキーワードの2次元平面への配置でユーザの検索条件を表現する。平面上の位置は、そのキーワードの優先度と曖昧さを表しており、複雑な条件を直感的かつ容易に表現できる。検索エンジンは、キーワードの出現頻度に基づいて優先度を、文字列上の類似度で曖昧さを処理し、各文書の条件への近さ(適合度)を計算する。結果は適合度で整列された文書リストである。本システムにより、ユーザは検索条件を微妙に修正することで、検索結果を変化させ、目的の文書に次第に近づくことが可能になる。

SearchSpace: a fulltext database system
using keywords arrangement on 2D space

Fujio Tsutsumi

Central Reserch Institute of Electric Power Industry

2-11-1 Iwado Kita Komae, Tokyo 201 JAPAN

We propose a new full text database system "SearchSpace". The interface of SearchSpace has a 2-dimensional space for arranging keywords. Trough the arrangement users can easily communicate their confidence and their preference about keywords to the system. From some experimentation, we found that users get control of retrieval as they intend by the system.

1 はじめに

WAIS に代表される古典的な全文検索システムでは、検索条件として、複数のキーワードの論理積・和を用い、検索エンジンはキーワードの完全一致を基本としている。しかし、用語の統制の緩い一般的な文書を対象とする場合には、言語の表記上のゆれや、類義語・多義語、またユーザーの記憶の不完全さへの対応といった問題が存在するため、時として検索不能となることがある。また、積・和という単純な条件しか使えないため、検索結果が少なすぎたり、多すぎたりする事態を招き、検索が困難となることがある。これらの問題に対して、現在までに様々な手法が提案されてきており、いくつかは実用システムとして構築されている。[3][4][10][12]

しかし 曖昧さの問題に対して、現在までに提案された手法の多くは、データベースシステム側があらかじめ予想される範囲の曖昧さや表記のゆれなどに対する処理機構を備えることで、対応しようとするものであるため、一般的には予測不可能なこの問題への対応として完全なものではない。また、検索式の記述力に関しては、自然言語を用いて検索要求を表す研究は多くあるが、記述力の向上という面では必ずしも成功していない。

我々はこの問題に、ユーザの力をもっと活用することで対処できる部分があるのではないかと考えている。つまり検索要求の曖昧さや意味付け、さらに自身の記憶について最も理解している、検索者自身の力を発揮できるような検索環境を実現することが、検索効率の向上に繋がるのではないかと考えている。

この考えに基づいて、我々は「SearchSpace」と呼ぶ全文検索システムを開発した。SearchSpace の検索条件インタフェースは、キーワードを 2 次元空間に配置させるという独自の機構を持っている。これによって、ユーザは、複数のキーワードの優先度と曖昧さを含めた複雑な条件を、容易にかつ直感的に指示することができる。開発した検索システムでは、入力された条件に基づき、文書 DB 中から適合する文書を抽出する検索機構を供えており、柔軟な検索を可能にしている。

我々は本システムを小規模のネットワークニュース(記事数 11000 本、25MByte)に適用し、アプローチの有効性を確認した。

2 キーワードの空間配置

SearchSpace は、ユーザの質問入力画面として、キーワードを四角形の 2 次元空間に配置させるというインタフェースを採用している(図 1)。画面上のキーワ

ードはマウスによって、自由な位置に移動できる。また生成・削除も容易に行える。

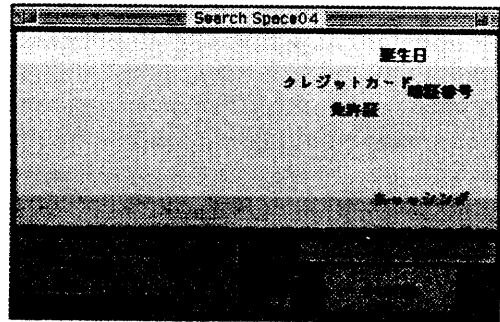


図 1: SearchSpace の検索条件入力インタフェース

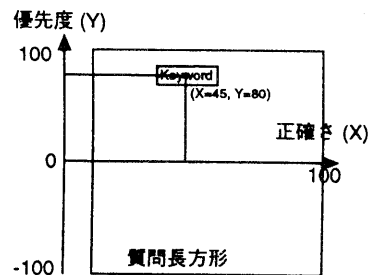


図 2: キーワード位置の解釈

キーワードの画面上の位置は、上下方向がそのキーワードの優先度を表し、左右方向が曖昧さを表す(図 2)。例えば、上にあるキーワードは下にあるキーワードよりも、検索結果として得たい文書に、より含まれていて欲しいことを表す。また、右にあるキーワードは、正確で表記の揺れを考慮する必要がないということを表し、左にあれば、緩やかに扱う必要があるということを表す。優先度や、曖昧さをどのように検索結果に反映させるかは、4 章で述べる。

従来型のメニューやスライドバーによる条件入力では、個々の条件を正確に指示するには適しているが、複雑な条件を入力できるようにするためには、いくつものメニューやスライドバーを備える必要があるため、操作が複雑になる。また、複数の条件間の相互関係がわかりにくい、という欠点がある。それに対して、キーワードを空間配置するという方法は、細かい数値を扱

ような操作は難しいが、条件相互の関係が一目瞭然であり、マウスによって位置を移動するだけで複数の条件を変化させることができる点で容易であると言える。文書検索において個々のキーワードの優先度や曖昧さは、細かい数値的な指示がほとんど意味をもたないのに対して、条件の変化や条件の間の相互関係は大きな意味を持つと考えられる。

3 空間配置による検索例

2つの検索例を用いて、検索システムの動作を説明する。

例 1. (横方向の移動を使った例):

3-(a)
3-(b)

図 3: 横方向の移動を使った例

図3は「震災保険」というキーワードの2つの配置例を示している。

3-(a) の検索結果:

該当するものではありません

3-(b) の検索結果:

記事番号 (得点): ヒットしたキーワード

14009 (53045): 火災保險 地震保險

13985 (42436): 火災保險 地震保險

13973 (42436): 火災保險 地震保險

14043 (31827): 火災保險 地震保險...

検索結果のそれぞれの行は一つの文書を表す。行の最初の数字は記事番号、2番目は文書の質問に対する得点、コロン以降はヒットしたキーワードである。結果は得点によってソーティングされて出力される。

図 3-(a) ではキーワードが右隅に配置されている。これは、キーワードを正確に扱うことを示す。結果は、同じキーワードがないため、該当なし、である。一方、3-(b) では左に配置されているので、キーワードは曖昧

に扱われる。そのため、類似文字列である「火災保険」「地震保険」を含む文書が検索されている。□

例 2. (縦方向の移動を使った例):

Figure 4 consists of two panels, (a) and (b), showing examples of mail survey forms. Panel (a) shows a form with the text '郵便局 調査' (Post Office Survey) at the top. Panel (b) shows a form with the text '調査' (Survey) at the top and '郵便局' (Post Office) on the right side.

図 4: 垂直方向の移動を使った例

4-(a) の検索結果:

13913 (1040401): 調查簡易郵便局

13912 (764873): 調査 071. 取扱郵便局 076. 郵便

13914 (224422): 船内郵便局 付録 1. 郵便局 郵便

13856 (132613): 広島西郵便局 広島中央郵便局

13862 (112211): 広島西郵便局 広島中央郵便局 ...

4-(b) の検索結果:

14641 (71092): 調查 調查依頼 調査不能 郵便局

14670 (51284): 調査 調査不能 郵便局

13912 (34508): 調査 Q71. 取扱郵便局 Q76. 郵便

14640 (30888): 調査 郵便局

14185 (20200): 調查...

この例の 4(a) と 4(b) では、「郵便局」というキーワードの位置が違う。このキーワードを上から下に移動したために、4(b) の検索結果では「郵便局」というキーワードを含む文書が下位に下げられ、残りの「調査」を含むものが上位に上がっている。□

4 検索エンジンでの曖昧さ・優先度の処理

ユーザから与えられたキーワード配置情報をもとに、検索エンジンがどのように適合する文書を検索するかを述べる。

全体の流れは次の通りである。

ステップ1: まずそれぞれのキーワードの指定された曖昧さに基づいて、文書中の適合するキーワードを抽出する。

ステップ2: その後、各文書が適合したキーワードをどれほど含んでいるかで、文書の適合度を計算する。

検索結果は適合度でソーティングした上位の文書リストである。

4.1 文字列共有度

曖昧さは、キーワードの文字列としての類似度で計算している。意味的な類似性を判定する辞書や機構は装備していない。文字列間の類似度を計るために、検索システムは文字種によって異なる次のような文字列共有度を採用している。

漢字の場合: 漢字の場合は、単純に文字の共有数を文字列共有度とする。例えば「内閣改造」と「改造人間」の共有度は2である。

カナの場合: カナの場合は、連続した2文字の重なり度合いを文字列共有度とする。例えば「システム」と「シスオペ」の共有度は1で、「エンジン」と「エジソン」の共有度は0である。

アルファベット・数字の場合: アルファベットの場合は、連続した3文字の重なり度合いを文字列共有度とする。例えば、「Inform」と「Transform」の共有度はfor, ormの2つで、2である。

インデックスファイルはこのセグメンテーションに基づいて構築されており、類似文字列を高速に抽出することを可能にしている。

4.2 検索結果の計算方法

検索結果の具体的な計算方法を以下に述べる。

質問入力画面は上下方向の真ん中の線(優先度ゼロ線)で半分に区切られており、その線より上方向に正の優先度を、下方向に負の優先度を表す(図2)。優先度は基本的には、各文書中でのキーワードの出現回数を元に計算する。

キーワードの配置される長方形の左辺をY軸、上下に2等分する水平線をX軸とする。原点は左辺の中心となる。この点より上方向に+Y、下方向に-Yとし、四角形の上辺をY=100、下辺をY=-100とする。また、原点より右方向に+Xとし、長方形の右辺をX=100とする。これにより、各キーワードの位置を座標として記述する。

ユーザの指示したキーワードとその配置は、各キーワードの文字列と座標上の位置の組として検索システム

に渡される。与えられたキーワードの集合をKとする。文書DB中の文書の適合度FPを次の2つのステップで計算する。

ステップ1. 適合キーワード集合を求める

K中のあるキーワードkに対して、まず適合キーワードの集合SKを次のように決定する。文書DB中に現れる全キーワード集合AKに含まれるあるキーワードakとキーワードkの類似度sdは、kとakの文字列共有度をsl、kを構成するセグメント数(漢字の場合は文字列数、アルファベットの場合は文字列数マイナス2)をlとして、次のように定義する。

$$sd = sl/l.$$

適合キーワードは、このsdの値が、キーワードの曖昧さを表すX座標の値xに対して次の関係を満たすものである。

$$sd \geq x/100.$$

例えば、4文字の漢字キーワードがX=50の位置に置かれた場合、そのキーワードと適合するキーワードは、 $sl/4 \geq 50/100$ の関係、つまり、 $sl \geq 2$ 、2文字以上の共有文字列を含むキーワードとなる。

ここで、適合キーワードは、ユーザが与えた曖昧さで判定した、類似キーワードの集合となる。曖昧さを高く(キーワードを左に位置させる)すると、共有文字列の少ないキーワードも適合キーワードと判定する。曖昧さを低くすると、共有文字列の多いキーワードのみを適合キーワードと判定する。一番左端におくと、曖昧さは最低になり、X=100で完全に同じ文字を含むもののみを、適合キーワードとするようになる。

ステップ2. 各文書の適合度を求める

ユーザが与えたキーワード集合K中のすべてのキーワードに対して、適合キーワード集合を求める。キーワードkの適合キーワード集合をMK(k)とする。キーワードkのY座標の値、優先度をy(k)とする。文書DB中のある文書Dの適合度FPを、次のように計算する。

文書D中に含まれるすべてのキーワードの集合をDK(D)とする。DK(D)中のあるキーワードdkの適合度P(dk)は、

$$P(dk) = \sum_{dk \in MK(k)} y(k).$$

と定義する。文書Dの適合度FP(D)は

$$FP(D) = \sum_{dk \in DK(D)} P(dk) |P(dk)|.$$

と定義する。

ある文書の適合度はその定義から、優先度が高いとされたキーワードが数多く含まれている場合に高い値を取る。逆に負の優先度のキーワードをたくさん含むと低い値、時にマイナスの値を取る。

優先度の総和でなく、優先度の二乗の総和としているのは、実験から、より直感に合致すると判断したためである。その根拠は薄弱であるが、単純に和とした場合は、優先度が強いと指示したものの効きが弱い、という印象を与えるようである。

5 実験概要

本システムを、ネットワークニュースに対して適用した結果を述べる。

対象とした文書群:

分野: 特定のパーソナルコンピュータに関する
ネットワークニュース記事 (fj.sys.mac)

言語: 日本語

分量: 11000 本, 25MByte

インデックスファイル: 34MByte(実サイズ)
59MByte(ファイルホール含む)

開発したシステムの概要:

構成: ワークステーションとパソコンをイーサ
ネットで結んだサーバクライアント型

開発言語: GNU C++(サーバ) と
HyperCard(クライアント)

通信プロトコル: TCP/IP

インデックスファイルは ndbm を使用しているため、ファイルホールを含む見かけのサイズが、実際のサイズより大きくなっている。

5.1 検索速度に関する結果

検索速度は、適合度計算の時間と検索結果として返されるデータの転送時間によって決まる。DB 中から選んだ、1 から 5 個のキーワードを与えて検索したところ、検索結果は最悪 7 秒、平均 4 秒で返された。DB 中に頻出する単語や、キーワード数を多くするに従い、適合度計算に時間がかかり、遅くなる。

検索システムの効果に関しては現在以下に示す定量的な結果のみである。より定量的な結果を示すための実験を現在行っている。

5.2 「曖昧さ」の効果

検索を行う際に、多くの場合、最初に与えるキーワードには自信がないため、また実際正確さにかけるため、左に位置させることにより、広めに検索範囲を指定できるという効果があった。これにより、従来の検索をした結果がゼロという事態を、避けることができる事例が多くあった。また、最初に正確だと思って右に配置したら、結果があまりに少ないため、あらためて左に動かして検索することにより、目的の文書を発見できた事例があった。具体的には、記事中で単語がミスタイプされていた事例があった。

しかし本システムの曖昧さ処理によって補えるのは、文字列の類似のみであり、例えば、「漢字トーク」と「KanjiTalk」のような言い替えにはまったく対応できないため、検索不能となる場合があった。

5.3 「優先度」の効果

正の優先度と、負の優先度を組み合わせることにより、検索範囲を広げたり絞り込んだりすることが容易になった。実際には、最初に与えたキーワードがカバーする範囲が広すぎる場合に、その中の一文を見てみて、関係ない文書に共通して出てくるキーワードを負のキーワードとして与えてゆくことで、だんだんと目的の文書に近づいて行くという操作が可能になった。また、曖昧さと組み合わせることにより、最初に広めにとって、細かく刻んで行くなどの操作が容易に行えるようになった。

5.4 空間配置インタフェースの効果

これら「曖昧さ」「優先度」の情報を使った、検索戦略は、空間配置というインタフェースによって、迅速かつ容易に実現することができた。また、一度使わなくなったキーワードを後で再利用するなどの行為も、容易に行えることから、検索戦略の幅を広げることにつながった。しかし、検索戦略の幅が広がったことにより、一つの目的を達成するまでに行う検索回数 (Search ボタンを押す回数) は多くなった。それによって目的が達成される場合は良いが、時に検索不能の場合もあり、その場合は検索不能を確認するまでに、従来システムに比べてより検索回数を必要とした。

6 関連する研究

6.1 キーワードの文字列上の類似性を扱う研究

文字列上の類似を扱う手法にはすでに様々な提案がなされている [5] [6] [7]. ただしこれらは、あらかじめ予想される揺れを想定してそれに対して対処を考えた物であるため、自由な曖昧さに対応することはできない. 対して、曖昧さの指定をユーザに解放した NSEARCH[11] という検索システムがある. NSEARCH はニューラルネットワークを用いて類似関数を構築しているが、内部処理に関しては明らかにされていないため、本手法が採用している手法とどちらが優れているかは判定できない. なお、本手法が採用している文字列類似性の処理手法は [8] などで採用されている n-gram の多言語版への拡張ととらえることができる.

6.2 全文検索のユーザインタフェースの拡張

全文検索において、ユーザとシステムのインタラクションが重要な役割を果たしていることは古くから指摘されていた [3]. 中でもユーザが検索結果を採点し、その情報をシステムが検索結果にフィードバックさせるという適合性フィードバック (relevance feedback) に関しては多くの研究がなされており、有効性も認められている. このアプローチと SearchSpace との違いは、関連性フィードバックでシステム側が受け持っていた部分を、ユーザ側により多く受け持たせようとしている点である. この違いによって、SearchSpace は目的がより明確で能動的なユーザを支援するのに向き、適合性フィードバックはより曖昧な目的を持つ受動的なユーザに適していると考えられる. これに関しては現在、関連性フィードバックを拡張したシステムを構築中であり、それとの比較の中で明らかになってゆくものと考えられる.

ユーザとのインタラクションに関して、検索結果の視覚化を行うことによって全文検索の効率を上げようとする研究がある [9][13]. ユーザの制御力を増やすという観点からは、ユーザの操作によって結果がどのように変化したかをしめす情報が特に求められる. SearchSpace の結果出力にどのような視覚化が適しているかは今後の課題である.

6.3 ユーザの制御感覚の増幅を促す研究

全文検索ではないが、ユーザの要求の時事刻々の変化に対し、ダイナミックに結果を変化させる研究として VIS[1] があり、SearchSpace と基本的な考え方に多くの共通点がある. ただし VIS が操作と視覚化の連続的な相互作用をより重視しているのに対し、SearchSpace は集約された一貫性のある操作系を構築することに重みを置いている点に違いがある. その違いは VIS が高速なリアルタイム処理を要求し、数多くのスライドバーやボタンを配置することを容認しているのに対し、SearchSpace は比較的低速な処理を容認し、条件指定が 2 次元空間配置に集約されている点に現れている.

7 まとめ

本論文では、検索条件の表現としてキーワードの 2 次元空間配置を使う全文検索システム「SearchSpace」のインタフェースと検索エンジンについて述べた. 小規模のネットワークニュースへの適用実験から、本システムのアプローチの有効性を確認した. 我々のアプローチは、曖昧さなどへの対応を行う従来の様々な検索システムのアプローチと排他的なものではなく、互いに協力することで、より効果的な検索環境を実現できるものと考えられる.

なお SearchSpace が採用した「曖昧さ」「優先度」というパラメータ以外の有効なパラメータとして、1. 正規表現などに代表されるキーワード間の位置関係、2. 章立てやヘッダ情報など文書の構造に基づく情報、3. 文書の量やレイアウト [13]、4. 記事の発生した時間的な位置 [2]、などがある. これらを操作系全体の一貫性を損なわないで、組み込むことは今後の課題である.

参考文献

- [1] Ahlberg, C. and Shneiderman, B.: Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *Proc. of CHI94*, pp. 313-317, 1994
- [2] Allen, R. B.: Interactive Timelines as Information System Interface. *Proc. of International Symposium on Digital Libraries 1995*, pp. 175-180, 1995
- [3] Ellis, D.(細野 公男監訳): 情報検索論 認知的アプローチへの展望. 丸善株式会社, 1994.
- [4] 藤澤 浩道, 絹川 博之: 情報検索における自然言語処理. 情報処理, Vol.34, No.10, 1993

- [5] 平出 基一 他.: 誤ったキーでも検索できるファイル構成法. 情報処理学会データベース研究会報告, 93-8, pp. 65-74, 1993
- [6] 加藤 宏次 他.: 自由語による全文検索のためのテキストサーチマシン TSM-I. 情報処理学会第39回全国大会予稿集, pp.1075-1076
- [7] 菊池 忠一, 飯島 豊.: キーワードのコード化による一検索方式. 情報処理学会第39回全国大会予稿集, pp.1073-1074
- [8] Liang, T. and Yang, W.: Signature Methods in Chinese Text Retrieval. *Proc. of International Symposium on Digital Libraries 1995*, pp. 97-104, 1995
- [9] Morohashi, M., Takeda, K. Nomiyama, H. and Maruyama, H.: Information Outlining - Filling the Gap between Visualization and Navigation in Digital Libraries. *Proc. of International Symposium on Digital Libraries 1995*, pp. 151-158, 1995
- [10] 根岸 正光.: フルテキスト・データベースの応用動向. 情報処理, Vol.33, No.4, pp.413-420, 1992
- [11] 日経 BP 社 (北郷 達郎): 急増する全文検索システムの動向を探る. 日経インテリジェントシステム, 172号, pp. 16-21, 1993
- [12] 小川 隆一, 菊池 芳秀, 高橋 恒介.: フルテキスト・データベースの技術動向. 情報処理, Vol.33, No.4, pp.404-412, 1992
- [13] Rao, R. et. al.: Rich Interaction in the Digital Library. *Communication of the ACM*, Vol.38, No.4, pp. 29-39, 1995