自然言語処理 117－14
（1997. 1. 21）

シソーラスを用いた文書データの自動分類法

塩見 隆一　徳田 克己　青山 昇一　柿ヶ原 康二
松下電器産業（株）マルチメディア開発センター

キーボードに不慣れなユーザーにとって、メニュー選択による検索は、有効な検索手法である。しかし、ネットワークなどを通じて得られるフロー型データに対して、リアルタイムに適切なメニューを作成することは難しい。我々は、シソーラスを用いて文書データを階層的に分類し、適切なメニューを作成する手法を提案する。本手法では、(1) 文書データ中の重要語を用いてシソーラス上に文書データを対応付け、(2) シソーラスを変形しメニューを作成する。シソーラスを用いることにより、意味的な関係を保持した階層的なメニューを作成することができる。評価用データベースＢＭＩＲ－Ｊ１を用いて作成したメニュー検索システムで、７３％のデータを検索することができた。

A Method of Clustering Documents Using Thesaurus

Takakazu Shiomi　Katsumi Tokuda　Shoichi Aoyama
Kouji Kakigahara

Multimedia Development Center, Matsushita Electric Industrial Co., Ltd.

Retrieval by menu is one of the most effective methods for users that don't always type by keyboard. However, it is difficult to make a menu from flow data through a network in real time. We suggest a method of clustering documents using a thesaurus and making a hierarchical menu. In this method, (1) System connects words that appear frequently in the document and words in the thesaurus. (2) System makes a menu by changing the thesaurus. This method enables the creation of a menu having a semantic hierarchy using a thesaurus. On a menu retrieval system using a database BMIR-J1, 73% of data was retrieved.

1　はじめに

近年、WWWを代表とするネットワーク上の大量の電子データを個人が取り扱えるようになり、この中から必要な情報を獲得するためのデータ検索技術がより重要になってきている。有力なデータ検索手法として、キーワード検索がある。しかし、キーワード検索は、情報機器に不慣れな初心者ユーザーにとって、

1. 適切なキーワードを思いつくのが難しい。
2. 文字入力そのものが難しい。

という問題点がある。

これらの問題点を解消するもう１つの検索方法にメニュー検索がある。メニュー検索では、データを一覧表示し、ユーザーはその中の情報を選択するだけでよい。しかしながら、ネットワーク上のデータのようなフロー型データに対して、即時的にメニューを作成することは容易ではない。

そこで、我々はシソーラスを用いてデータを自動分類する手法を提案する。本手法を用いることにより、データ検索のための階層的なメニューを作成することができ、ユーザーはメニュー検索によって容易に情報を得ることができる。本稿では、

1. シソーラスを用いた文書の自動分類手法
2. 本手法を用いたメニュー検索システム
3. 検索実験

について報告する。

関連研究

文章データの自動分類は、大きく分けて2つの方法に大別できる。1つは、予め定義したカテゴリーに対して割り付ける方法 [1, 2, 3] であり、もう1つはクラスタリングアルゴリズム [4, 5] を用いて分類するものである。

前者の一般的な手法は、カテゴリーの代表的な文書ベクトルを用意し、対象文書との距離を用いて分類を行なう。この手法ではユーザーにとって、わかりやすい分類グループにデータを分類できるが、不特定のデータに対して分類を行なう場合、大量のカテゴリーを必要とし、これらのカテゴリーを用意することが困難であるという問題点がある。

後者は、不特定のデータに対応可能であるが、分類されたグループの意味を表現するのが難しく、ユーザーがグループの選択を行なうのが困難とされている。

我々の手法は前者に分類することができる。カテゴリーとして既存のシソーラスを用いることで、ある程度広い範囲のデータを階層的に分類できる。しかし、あくまでも単語ベースであるので、末端の階層でまとめられている文書データは相互に類似しているとはいい難く、指定した記事の類似記事を収集する補助機能などとの組合せが不可欠である。

また、索引の作成という観点で、分類語彙表 [6] を用いた索引作成の取り組みが行なわれている [7]。本手法は、この取り組みを一歩進め、文書中の重要単語の抽出を行ない、積極的なシソーラスの変形及び表示の工夫を行なったものである。また、被験者実験を行ない有効性の評価を行なった。

2 シソーラスを用いた文書の自動分類

2.1 メニュー作成の方針

メニューを作成するに当たって、品質の高いメニューの要件を考える必要がある。文献 [8] によれば、品質の高いメニューは、

1. 意味構造を考慮していること。
   →ユーザーの項目選択が容易。
2. 階層数が少く階層中の項目が（一定範囲内で）多いこと。
   →階層中のデータが多い方が全体把握が容易。

とされている。

まず、メニューに意味構造を持たせるためにシソーラスを利用する。分類すべき各文書からは主題を代表するような単語を抽出し、シソーラスに対応付けする。シソーラスの構造を利用して、各文書の主題を表す単語を階層構造のメニューにすることで、意味構造を持ったメニューを実現する。

次に、階層数を削減するために、シソーラスの変形を行なう。シソーラス中のメニュー作成に必要な項目部分を取り出し、シソーラスの変形を行ない、不要な中間項目を削除することによって階層数を削減する。

2.2 文書データとシソーラスとの対応付け

文書とシソーラスの対応付けは、文書の主題を表すような重要語を抽出し、シソーラス中の単語と照合を行なうことによって行なう。抽出した重要語がシソーラス中に複数の語義を持つ場合は、その両方と対応付けする。また、文書の重要語はユーザの検索要求によって異なることが考えられるので、文書中から複数の重要語を抽出し、シソーラスと対応付けることとする。

文書中の重要語抽出には、以下の2通りの基準 [9] について試した。

- **tf(文書内単語頻度)**
  単語 i の重要度 tf_i は、

$$tf_i = W_i$$

  とする。ここで W_i は単語 i の出現回数である。
- **$tf \cdot idf$(文書内単語頻度 * 文書頻度)**
  文書 j 中の単語 i の重要度 $tf \cdot idf_{ji}$ は

$$tf \cdot idf_{ji} = W_{ji} * log\frac{N}{n_i}$$

  とする。ここで、W_{ji} は、文書 j 中の単語 i の出現回数、N は分類対象文書数、n_i は分類対象文書の中で単語 i を含んでいる文書数である。

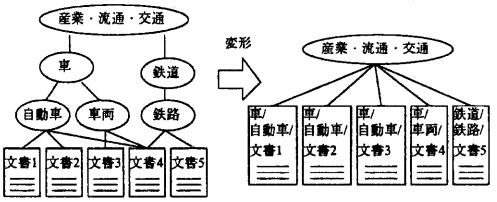
2.3 シソーラスの変形

階層数を削減するために、（1）シソーラス中のメニュー作成に必要な項目部分を取り出し、（2）シソーラスの変形を行ない不要な中間項目を削除する。ただし、階層数が少ければ、1階層あたりの項目数が多くてもよいというわけではない。ここでは、1階層当たりのメニュー項目数の上限 E_{max} を設定する。1階層あたりの項目数 E が E_{max} を越えるようなシソーラスの変形は行なわないものとする。

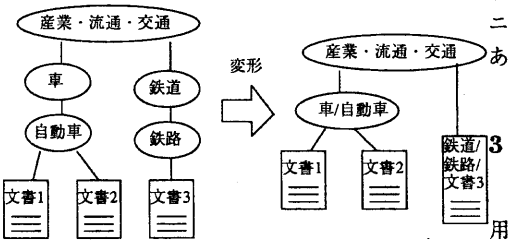
具体的には、以下の操作を行なうことによって、意味的な構造を保持しながら階層数の削減及び項目数の削減を行なう。

操作1　対応文書のない項目の除去。

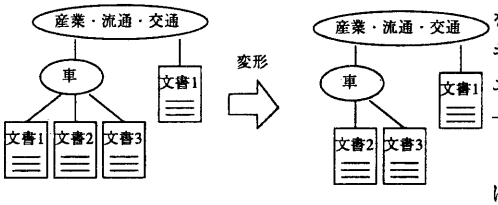
操作2　下位層の全項目に対応する文書総数がE_{max}以下なら、下位層の中間項目を全部削除する。



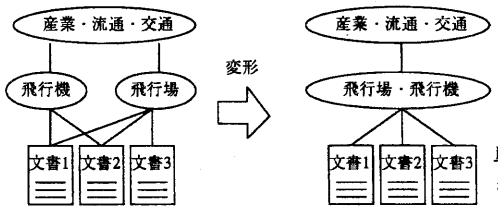
操作3　対応文書が1つだけの中間項目を削除する。



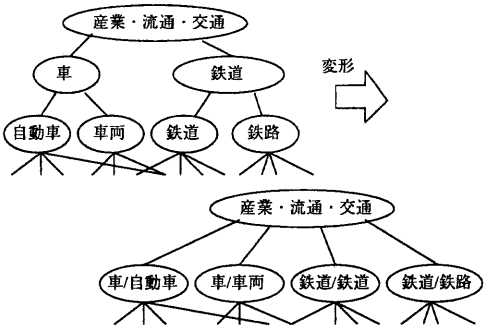
操作4　項目直下にある文書と同じ文書を他の下層項目下から削除する。



操作5　同じ親項目を持つ子項目で、対応文書が他項目の対応文書に包含されている項目を合成する。



操作6　階層内の項目数が少い時、下位項目を上位に格上げする。



なお、各変形によって削除される中間項目の項目名は上位／下位の項目名と合成して残す。これはメニュー表示の際のユーザーの選択を助けるためである。

3　メニュー検索システム

実際に、シソーラスを用いた文書の自動分類法を用いたメニュー検索システムをWWWのＣＧＩプログラムとして作成した。ユーザーがWWWブラウザで新聞記事データ群を指定すると、システムは各記事から重要語を抽出する。次に、各メニュー画面をＨＴＭＬドキュメントとして自動生成する。画面中の選択項目にはアンカータグを付け、他のメニューを呼び出す構成としている。これらのＨＴＭＬドキュメントをユーザーはブラウザで表示しメニュー検索を行なう。図 1は、本システムの画面の一部分である。

表示画面に関しては、ユーザーが項目選択を容易にできるように３つの工夫を行なった。

- ユーザーが各項目を選択するための手がかりとして、項目下の項目名を同時に表示する。
- ページ内に記事へのリンクと次のメニューへのリンクが存在する場合は、これらを分離表示し見易くする。
- 記事タイトルの一覧では記事とメニューを対応付けたキーワード別にまとめて表示する

また、本分類手法では、基本的に共通／類似する単語を含む文書を１画面のメニューにまとめあげたものにすぎない。よって、メニュー画面に含まれている記事群が類似した文書の集合になっているとはいい難い。そこで、目的とする記事を１つ選択すると、その記事の類似記事を収集し、ランキング表示する補助機能を付ける予定である [10]。ここでは、類似文書を収集する機能の詳細は省略する。

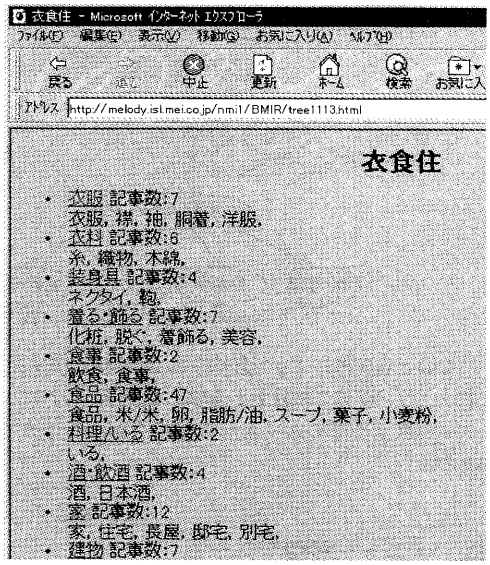


図1: メニュー検索システムの画面

4 検索実験

4.1 データとシソーラス

実験を行なう検索対象データとして、情報検索システム評価用ベンチマークＶｅｒ１．０（ＢＭＩＲ－Ｊ１）[11][1]を用いた。このデータは、日本経済新聞から抽出した新聞記事６００件、検索要求文６０件、検索要求文に対応する正解記事集合で構成されている。検索要求文は、検索のために必要な主機能よって、以下の６グループに分類できる [12]。

グループ	検索に必要な主機能	件数
G1	基本機能	１０
G2	数値・レンジ機能	５
G3	構文解析機能	６
G4	言語知識利用機能	１２
G5	世界知識利用機能	１０
G6	言語知識利用と知識処理機能	１７

また、ＢＭＩＲ－Ｊ１は検索要求文に対する正解記事をＡランクとＢランクで定義している。Ａランクは記事の主題が検索要求文の内容と一致するもの、Ｂランクは記事の主題と一致はしないが、検索要求文の内容を含むものである。

シソーラスには、小学館類語例解辞典 [13] を用いた。小学館類語例解辞典は、約２５０００語を約６０００の類語グループに分類し、さらに、これらの類語グループを１０の大分類と２０の中分類によって２００のグループに分類したものである。

4.2 メニュー中の記事分布評価

被験者による検索実験を行なう前に、検索要求に対する正解記事が、ユーザーが検索可能なメニュー中の適切な位置に配置されるのかを評価した。

評価にはＢＭＩＲ－Ｊ１の検索要求文と正解記事集合を用いた。メニューは文書の主題を抽出しシソーラスを変形して作成していることから、検索要求文の内容と記事の主題が一致するＡランクの正解記事のみを用いた。検索要求文の中にはＡランクの正解記事を持たないものが２件あるため、実際には５８件の検索要求文を用いた。

評価は以下の手順で行なった。

1. 各正解記事から５つの重要語を抽出しておく[2]。
2. 筆者が検索要求文を読み、重要語を１～２個連想する。
3. 検索要求文に対する複数の正解記事から抽出された全重要語と、連想した重要語を照合する。
4. 一致した重要語があれば、検索要求文に対する正解記事をメニューから１件以上検索できると見なす。

上記評価を２つの重要語抽出基準について行なった。表１は結果をまとめたものである。

重要語抽出基準	正解記事を検索できる検索要求文件数（率）
tf	35(60.3%)
$tf \cdot idf$	28(48.3%)

表1: 重要語抽出基準と正解記事分布

検索要求文に対して１つ以上の正解記事を検索できると見なした検索要求文件数は tf の方が多かった。

4.3 メニュー作成

次に、シソーラスの変形を行なってメニューの作成を行なった。メニューの作成に当たって、１文書

[1]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

[2]抽出された重要語はシソーラス中に存在する単語である

から抽出する重要語個数を５、重要語抽出基準を tf 、１画面に表示できる最大項目数 $E_{max} = 20$、とした。シソーラスの各変形操作後のメニューの総画面数、項目選択までの平均階層数について調べた。

表２は、結果をまとめたものである。

平均階層数は１．１削減できた。操作別では、操作３までが有効であることもわかった。

操作	1	2	3	4	5	6
総画面数	2067	680	457	451	445	432
平均階層数	5.0	4.0	3.9	3.9	3.9	3.9

表 2: 操作とメニュー構造

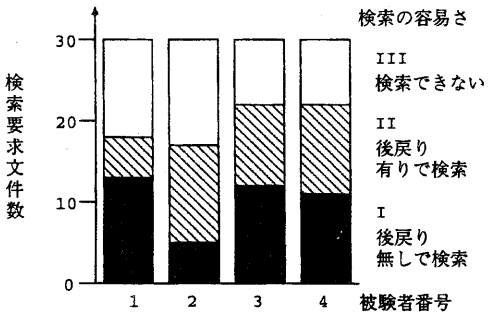


図 2: 被験者検索実験結果（被験者別）

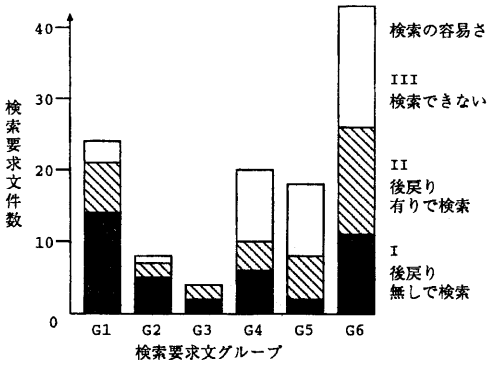


図 3: 被験者検索実験結果（検索要求文グループ別）

4.4 被験者実験

最後に、作成したメニューの有効性を調べるため、被験者による検索実験を行なった。

ＢＭＩＲ－Ｊ１の検索要求文の中から、事前評価を行なった結果、正解記事が検索できると見なした３５件から、類似した４件の質問を除いたの３１件を用いた。実験の前に被験者に対して、検索要求文１件を用いて検索方法を具体的に説明した。実験は、残りの検索要求文３０件に対して被験者に検索して貰いメニュー検索システムで正解記事を１件以上発見できるかどうかを評価した。後戻りできる上限を３回とまでどし、

I 後戻りすることなく発見できた

II 後戻りありで発見できた

III 発見できない

の３段階で検索の容易さを評価した。

被験者は小学館類語例解辞典を見たことがない２名（被験者１、２）と見たことがある２名（被験者３、４）の合計４人にお願いした。

図 2は、実験の結果を被験者別にまとめたものである。小学館類語例解辞典を見たことがある被験者３、４の方がより多くの検索要求文に対する回答記事に辿り着いていることが分かる。

図 3は、実験の結果を検索要求文のグループ別にまとめたものである。図中の検索要求文件数は４人被験者の延べ数である。検索要求文が必要とする検索主機能が少いものほど回答記事に辿着ける率が高いことがわかる。

また、検索者が検索要求文に対して回答した記事を、ＢＭＩＲ－Ｊ１の規定する正解記事と照合してみた。全回答７９件のうち、６５件がＡランクあるいはＢランクの正解記事であった。不正解記事は検索要求文グループの難易度に対応しG4,G5,G6 に多く分布していた。

5 考察

メニュー中の記事分布について　作成したメニューでは、ＢＭＩＲ－Ｊ１の検索要求文５８件の約６０％しか、対応する正解記事が抽出できるようになっていない。この原因を調べたところ、正解記事からの適切な重要語の抽出率が tf の場合でも２５％程度と非常に低いことが分かった。原因を調べたところ、

- 適切な重要語が抽出されても、その単語がシソーラス中にない。
- 検索要求文から連想する重要語を１～２語として評価した。
- 採用した重要語抽出基準では適切な重要語候補が抽出されない。
- 形態素解析処理の失敗（主に未知語／固有名詞の扱い）。

であった。シソーラスの変更、形態素解析の改良を行なうことで適切な重要語の抽出率を５０％程度

に引き上げることができるので、検索要求に対する正解記事を1件以上抽出できる率は60%から、さらに向上が期待できる。今後、実用的なシステムを構築できるのではないかと考えた。

被験者実験について　メニュー内容に不慣れなユーザー（被験者1、2）は、やはり検索精度が悪い。しかし、メニュー検索の特性である慣れによる検索精度向上が期待できるので、被験者1、2も被験者3、4と同等の検索精度を出せるようになると思われる。実際に、実験後の被験者のインタビューで「実験後半はメニューに慣れて、検索しやすくなった」との感想を頂いている。

最も結果の良い被験者は、検索要求文の73%に対して記事を抽出している。この数字は高いとはいえない。しかし、回答できなかった検索要求文は高レベルの検索要求文が多く、本手法を用いたメニュー検索システムは簡単な記事の検索では有効であると判断できる。

実験後のインタビューで多く聞かれたのが、「思いつく単語がどこにあるのか探しにくい」との感想であった。シソーラスの表示方法や、よりデータに適したシソーラスを用いることで、より高い検索精度を実現できると考える。

また、実際にどのような経路を辿ってデータを探しているのかを調べると、いわゆる一般語で探していることが多いようである。メニュー検索システムでは、機械が苦手な世界知識や言語知識を人間が補って検索することで、より高い精度の検索結果を出すことも期待していたが、実際には検索要求文中の表層的な単語を利用している場合が多いようである。このユーザーの特性を考慮し、重要語抽出基準に見直しを行なうなどして、よりよいシステムが構築できる可能性がある。

6　まとめ

以上、本稿ではシソーラスを用いた文書データの自動分類法とこれを用いたメニュー検索システムについて報告した。今後、重要語抽出基準やメニュー表示方法の改良を行ない、よりよい分類提示ができるようにしていく予定である。

謝辞

ＢＭＩＲ－Ｊ１を提供して頂いた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、特に対応して頂きましたＮＴＴデータ通信 木谷様に感謝致します。

小学館類語例解辞典の電子化データの使用を許可して下さった株式会社小学館に感謝致します。

参考文献

[1] 上田隆也，大谷紀子，伊藤史郎，柴田昇吾，池田裕治：" フロー情報収集・活用のための知的検索システムＦｉｔ（1）（2）（3），" 第53回情報処理学会全国大会講演論文集，3-183～188, 1996.

[2] 森本由紀子，間瀬久雄，辻洋，絹川博之：" 新聞記事自動分類システム構築の検討と評価，" 第53回情報処理学会全国大会講演論文集，3-205～206, 1996.

[3] 西野文人：" テキスト分類のためのカテゴリ割り付け戦略，" 情報処理学会研究会報告，Vol.NL 106-3, pp. 13-18, 1995.

[4] 宮崎哲夫，田中栄治，古城則道：" 文書の意味空間へのマッピング，" 第53回情報処理学会全国大会講演論文集，3-167～168, 1996.

[5] 有田英一，安井照昌，津高新一郎：" 単語集合の自動構造化機能を持つ「情報散策」方式，" 電子情報通信学会・信学技法，NLC95-17, pp. 69-74 1995.

[6] 国立国語研究所（編）：" 分類語彙表，" 秀英出版，1964

[7] 千田恭子，篠原靖志，坂内広蔵：" 汎用シソーラスを利用した検索用の索引メニュー構成法，" 情報処理学会研究会報告，Vol.NL 111-4, pp. 21-26, 1996.

[8] Ben Shneiderman（東、井関訳）：" ユーザー・インターフェースの設計，" 日経ＢＰ, pp.93-98, 1988.

[9] G.Salton: "Automatic Text Processing," Addison Wesley, 1989.

[10] 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，" 第52回情報処理学会全国大会講演論文集，4-202～203, 1996.

[11] 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.DB 106-19, pp. 139-145, 1996.

[12] 木谷強，高木徹，木原誠，関根道隆：" フルテキストと抽出キーワードを利用した情報検索，" 情報処理学会研究会報告，Vol.FI 43-10, pp. 71-76, 1996.

[13] 小学館辞典編集部：" 類語例解辞典，" 小学館，1994.