自然言語処理 131－9
音声言語情報処理 26－9
（1999．5．28）

テキストの表層データに基づいた意味カテゴリーインデクシングの支援について

舟生日出男 黒田隆之 高橋毅 内田明宏 長尾仁志 伊丹誠 伊藤紘二

東京理科大学 基礎工学部電子応用工学科
E-mail: {funaoi, kuro, jai, uchida, nagao, itami, itoh}@itlb.te.noda.sut.ac.jp

［あらまし］ 近年では情報化の進展によって、人々が知的作業に際して処理すべき文書の量は増大している。そこで本研究では、文書を的確に分類し、その検索や再利用を容易にするために、文書を意味カテゴリーでインデクシングする作業を支援する手法について提案する。本手法では、文書から属性を抽出し、TfCon 法を用いてそれらの頻度と重みから算出されるベクトルによって、その文書の特徴を表現する。このようにして、各意味カテゴリーに該当する文書集合の特徴を、教師付きの学習によって機械に学習させれば、テキストに含まれる意味カテゴリー群の推定が可能となり、新たな文書のインデクシングの支援に用いることができる。

［キーワード］ インデクシング、TfiDf 法、意味カテゴリー、教師付き学習、情報分類

Supporting Semantic-Categorical Indexing Based on Text Data

FUNAOI, Hideo KURODA, Takayuki TAKAHASHI, Takeshi UCHIDA, Akihiro
NAGAO, Hitoshi ITAMI, Makoto ITOH, Kohji

Department of Applied Electronics, Science University of Tokyo

[Abstract] Owing to progress of informationalization, recently, people should process a quantity of documents on the occasion of intelligent work. In this study, to facilitate classifying, searching and reusing a document, we propose a method of supporting Semantic-Categorical Indexing. In this method, characteristic of a document is expressed with the vector of attribute values, which are consist of those frequency and weight calculated by TfCon method. If a machine learns the characteristic of a document set that corresponds to each meaning category with supervised learning, it can presume semantic categories corresponds to a text and support users to index a new document with those categories.

[Keywords] indexing, TfiDf method, semantic category, supervised learning, information classification

1 はじめに

情報化の進展とともに、近年では、人々の扱うべき情報の量が飛躍的に増大している。そのため、様々な情報を元に知的作業を行うには、多くの情報を的確に分類する必要がある。

この分類の作業が適切に行われなければ、様々な問題が生じることとなる。例えば、ある情報を不適切なカテゴリーで分類し、蓄積してしまうと、後で利用しようとする時に、その情報を取り出すことが困難となる。数年前に、野口悠紀雄氏の『「超」整理法』が大いに注目された理由もこの点にあると言えよう。

さらに、複数の人が協調して作業を進める場合にも、同様のことが言える。例えば、グループ内のあるメンバーと他のあるメンバーが、本来であれば同じカテゴリーに分類されるべき情報をそれぞれ異なるカテゴリーに分類したとする。その結果、グループ内での情報の共有や再利用に支障を来すこととなる。

このように、情報を適切に分類することの重要性は、非常に高いと言えよう。その手法については、金と宮本によれば、キーワードを用いる手法と、全文を用いる手法の、2種類に大別できる[1]。

まず、キーワードを用いた手法であるが、これは、学術論文など、キーワードが付加されている文書を、そのキーワードを用いて分類する。例えば、相澤と峡は、著者キーワードから、「用語グラフ」と「用語間距離」を求めて、学術論文の分類を行っている[2]。ただしこの手法は、当然ながら、キーワードが付加された文書にしか適用できない。

もう一方は、全文を用いる手法である。文書全体を形態素解析し、抽出された単語を用いる。先行研究では、単語の頻度と重みを計算し、*TfiDf* 値を求めて、それに基づいて文書間の関連を計算し、文書を分類する方法が、徳永ら[3]など、数多く報告されている。

河合によれば、*TfiDf* 法は、表層の情報のみを扱っているため、分類精度が低くなるという短所がある[4]。だがその一方で、処理が簡単であり、汎用性が高いという長所も備えている。

また、先行研究では、シソーラスや概念辞書と組み合わせて分類を行う手法が、福本ら[5]などに見られる。しかし本研究では、できるだけ簡便な方法で分類を行うことを目的としており、文書の表層情報のみを用いることとする。

2 本研究の概要

以上の議論をふまえ、本研究では、情報を分類し、検索を容易にするために、文書を意味カテゴリーでインデクシングする方法と、その作業を支援する手法を提案する。

ここで、「意味カテゴリー」とは、“概念の最小単位を表すものであり、文書を処理し管理する各ユーザーまたはグループが、目的に応じて選定すべきもの”と定義する。例えば、地理に関する文書群に対して、「農業」「工業」「商業」などの意味カテゴリーを定義し、インデクシングすることができる。また、より詳細な分類が必要な場合、「促成栽培」「施設園芸農業」などを用いても良い。

このような方法によって、意味カテゴリーを予め定義し、文書をインデクシングしておけば、意味に基づいて文書を分類したり、時間が経ってから求める情報を検索する時に有効である言えよう。

本研究では、このようなインデクシングの作業を支援するために、インデクシングを行おうとする文書の特徴から、適切な意味カテゴリーを選択する手法について提案する。ただし、本研究では、あくまでも支援という立場をとり、意味カテゴリーの選択については、その精度よりも再現率を優先する。

本手法においては、まず、訓練用データを用いて意味カテゴリーに該当すべき文書の特徴を次のように求める。

1) 必要な意味カテゴリーを予め用意する。
2) それぞれの意味カテゴリーに該当すると(グループが合意し、)判断したテキストデータを多数用意する。
3) 後述する属性抽出器を用いて、全体のテキストデータから「属性」を全て抽出し、K 個の属性から成る、「属性リスト」を作成する。
4) 各意味カテゴリーごとに、属性リストに含まれる各属性の頻度を算出し、重み付けを行う。
5) 各意味カテゴリーに該当すべきテキストデータの特徴を、属性リストと各属性の頻度に基づいて、K 次元ベクトルで表現する。

このように教師付きの学習を行い、ある意味カテゴリーが該当する文書の特徴を属性のベクトルで表現して保持しておく。これらのベクトルを利用すれば、新たな文書に対するインデクシングを支援することが可能となる。

インデクシングを行いたい文書に対しては、次のようにしてその文書が該当する意味カテゴリーの推定することができる。

1) 対象の文書を、訓練用データと同様に、K 次元ベクトルで表現する。ただしベクトルの各要素は、各属性がテキストデータに含まれ

ていれば 1 、そうでなければ 0 、とする。

2) 各意味カテゴリーのベクトルとの内積を計算し、類似度を求める。このとき、全意味カテゴリーを基準に正規化した値が 0.22 以上[注1)]であれば、そのテキストデータが該当する意味カテゴリーとして選択する。なお、複数の意味カテゴリーが選択されることもあり得る。

以上の過程を処理するルーチンとインターフェースを備えたシステムを用意すれば、インデクシングを行おうとする文書を、ユーザーがシステムに入力すると、システムは、教師付き学習の結果に基づいて、その文書が該当する意味カテゴリーを推定し、ユーザーに提示できるようになる。

なお本研究では、属性の頻度と重みの計算は、*TfiDf* 法を改良した *TfCon* 法を用いて行う。以降では、属性抽出器と *TfCon* 法について説明する。

3 属性の抽出について

本研究における意味カテゴリーの推定手法では、文書(テキストデータ)を形態素解析し、属性抽出器によって、その結果から文書に含まれる属性を抽出し、その頻度を求め、重み付けを行い、推定に用いる。

そのためにまず、奈良先端科学技術大学院大学自然言語処理学講座が開発した形態素解析器である ChaSen Ver.1.0[6]を用いて文書を形態素解析する。しかし、その結果から得た語群そのままでは、不必要な語も含まれ、処理すべき語の総数が大きくなるため、頻度の計算には適さない。

そこで本研究では、不要な語を削除したり、必要に応じて複数の語を結合する「属性抽出器」を用いて、属性を抽出する。その規則を次に示す。

1) 基本的に、品詞が「名詞」「動詞」「形容詞」である形態素の辞書形を、属性とする。

2) 「未定義語」や「助詞」などの形態素や、改行を表す「EOS」、「ある」「する」など頻出するものを取り除く。

3) 出力された形態素のうち、辞書形に着目して同じものが複数存在する場合、それらを1個にまとめて重複を解消する。

4) 名詞や数詞の形態素が連続していれば、それらの辞書形を結合して1個にする。さらに、品詞細分類に注目し、それが名詞性接頭辞でその直後にある形態素の品詞が名詞であれば、それらの辞書形どうしを結合する。また、名詞性接尾辞についても同様に、その直前の形態素に着目して、結合する。

5) 「名詞」+「助詞」+「動詞・形容詞」のように、あるいは、「動詞・形容詞」+「名詞」のように連続している場合は、主格を明確にするために、「動詞・形容詞」+「名詞」のように結合する。

例えば、「大阪は、16世紀の終わりごろ大阪城がつくられ、城下町として発展した。」というデータを形態素解析すると表1のようになり、この結果から表2に示した属性が抽出される。

表1 形態素解析の結果

辞書形	品詞	品詞細分類
大阪	名詞	固有名詞
は	助詞	副助詞
、	特殊	読点
1	名詞	数詞
6	名詞	数詞
世紀	名詞	普通名詞
の	助詞	名詞接続助詞
終わる	動詞	動詞
ごろ	接尾辞	名詞性名詞接尾辞
大阪城	名詞	固有名詞
が	助詞	格助詞
つくる	動詞	動詞
れる	接尾辞	動詞性接尾辞
、	特殊	読点
城下町	名詞	普通名詞
と	助詞	格助詞
する	動詞	動詞
発展	名詞	サ変名詞
する	動詞	動詞
。	特殊	句点
EOS		

注1) ここで閾値を 0.22 としている理由は、訓練用データをそのまま評価用データとして用いた場合、閾値を 0.22 に設定した時に、精度と再現率が最大(*TfCon* 法と *TfiDf* 法で、ともにそれぞれ、98.4%、100.0%)となったことによる。

表2 抽出される属性

属　性	品　詞
大阪	名詞
16世紀	名詞
終わる-16世紀	動詞-名詞
大阪城	名詞
つくる-大阪城	動詞-名詞
城下町	名詞
発展	名詞

4 *TfiCon* 法－改良型 *TfiDf* 法－について

属性抽出器を用いて全テキストデータから抽出した各属性をそれぞれ、第 i 属性（i = 1,2,3,…,K）とする。次に、第 j 意味カテゴリー（j = 1,2,3,…,N）に対応するテキストデータの集合（以下、第 j カテゴリー集合）に含まれる文字数を M_j とし、第 i 属性の第 j カテゴリー集合での出現回数を f_{ij} とする。

ここで、第 i 属性の第 j カテゴリー集合での正規化された *tf* 値（頻度）を

$$w \cdot tf_{ij} = \frac{f_{ij}}{M_j} \cdot w \qquad (1)$$

とする。ただし、w は属性あたりの平均文字数であり、w を定数と考えれば、比較のためには、tf_{ij} を扱えば十分である。

また、第 i 属性の出現する意味カテゴリー数を L_i とし、従来の *TfiDf* 法をここに当てはめれば、

$$idf_i = \log \frac{N}{L_i}$$

のように *idf* 値を求め、*tf* 値に掛け合わせて重み付けを行い、*tfidf* 値を算出することになる。しかし情報理論の見地に立てば、この評価値には、各カテゴリー集合の大きさも、各属性の各カテゴリー集合での頻度も考慮されていないと言える。

そこで本研究では、それらを考慮し、以降で説明する「集中度」を重み付けに用いる。

第 i 属性の集中度は、

$$con_i = \log N - H(q_{i1}, q_{i2}, \ldots, q_{iN}) \qquad (2)$$

で、求めることができる。ここで、

$$q_{ij} = \log \frac{\frac{f_{ij}}{M_j} \cdot w}{\sum_{k=1}^{N} \frac{f_{ik}}{M_k} \cdot w}$$

$$H(q_{i1}, q_{i2}, \ldots, q_{iN}) = -\sum_{j=1}^{N} q_{ij} \cdot \log q_{ij}$$

（ただし、$q_{ij} = 0$ のとき、$q_{ij} \cdot \log q_{ij} = 0$ とする）とする。

また、q_{ij} は、各文書が全て同じ文字数になるように換算した場合に、全カテゴリー集合に出現する第 i 属性の総数のうち、第 j カテゴリー集合に現れる第 i 属性の割合を表す。

したがって、もし、q_{ij} が j に依らずほぼ一様（つまり集中度は最低）であるとすると、

$$q_{ij} \fallingdotseq \frac{1}{N}$$

であるから、

$$H(q_{i1}, q_{i2}, \ldots, q_{iN}) = -\sum_{j=1}^{N} \frac{1}{N} \log_2 \frac{1}{N} = \log N$$

となる。この結果、$con_i = 0$ であるため、第 i 属性の集中度は最小値の 0 となる。

（1）、（2）より、第 i 属性の第 j カテゴリー集合における *tfcon* 値は、

$$tfcon_{ij} = tf_{ij} \cdot con_i$$

となる。

以上の結果から、第 j 意味カテゴリーの特徴ベクトルは、

$$C_j = (tfcon_{1j}, tfcon_{2j}, \ldots, tfcon_{Nj})$$

のように表すことができる。

5 評価実験

5．1 概要

本研究で提案する *TfCon* 法の有用性を検証するために、*TfiDf* 法との比較実験を行った。

まず、教師付きの学習を行い、各意味カテゴリー（表3）が該当する文書の特徴ベクトルを算出した。そのために、意味カテゴリーによってインデクシングされた訓練用文書データを用意し、特徴ベクトルを *TfiDf* 法、*TfCon* 法、それぞれの手法で 12 個ずつ計算した。

次に、インデクシングされていない評価用文書データの特徴ベクトルを、両手法で算出した。そして、12 個の意味カテゴリーの特徴ベクトルとの内積をそれぞれ計算して、各文書が該当すると思

われる意味カテゴリーを推定し選択した。

推定に際して、閾値は先述した 0.22 に加え、0.15 から 0.26 まで 0.01 刻みで設定して行った。

各文書ごとの意味カテゴリーの選択数については、文書によっては、1つのもの、複数のもの、また、全く選択されないものも見られた。

5. 2 訓練用文書データ

中学校社会科の教科書(東京書籍[平成6、10年発行版]、教育出版[平成10年発行版]の計3冊)のうち、日本地理の領域を用いた。

教科書に記載されている各トピックをさらに内容に基づいて分割して「ブロック」とし、そのうち、54 ブロックに、意味カテゴリー(表3)を付加した。

表3 実験で用いた意味カテゴリー

意味カテゴリー名	該当するブロック数
農業がさかんである	1
稲作が 〃	6
畑作が 〃	6
果物の生産が 〃	3
漁業が 〃	2
工業が 〃	18
製糸業が 〃	4
商業が 〃	8
観光が 〃	2
海上交通が 〃	2
水運が 〃	4
政治が 〃	5
合 計	61

注)複数の意味カテゴリーに該当するブロックもある。

5. 3 評価用文書データ

中学校社会科の教科書(帝国出版[平成10年発行版])の九州、中国、四国地方に関するトピックを、内容に基づいて分割し、47 ブロックを抽出し、評価用データとした。このうち、10 ブロックは複数の意味カテゴリーに該当し、9 ブロックはいずれの意味カテゴリーにも該当しないものとなっている。

5. 4 結果

まず、47 のブロックについて、該当する意味カテゴリーが正しく選択されているか否かを、ブロックごとに判断して集計した。表4には、各閾値ごとでどのように選択されたかを、ブロック数とその割合で示した。各項目の意味は次の通りである。

「全て正しく選択」…選択すべき意味カテゴリーを全て正しく選択した、もしくは、どの意味カテゴリーも選択すべきではないときには全く選択していない場合。

「一部正しく選択」…複数選択した中で、どれか1つでも正しく選択されている場合。

「誤って選択」…選択した意味カテゴリーが、全て誤っている、もしくは、選択すべき意味カテゴリーが1つ以上あるのに全く選択していない場合。

ここで、「全て正しく選択」を優先して、これと「一部正しく選択」との合計に着目すると、*TfCon* 法は閾値 0.22 で 25(53.2%)、*TfiDf* 法は 0.21 で 22(46.8%)となっており、*TfCon* 法の方が僅かな

表4 ブロックごとの意味カテゴリーの推定結果

閾値	*TfCon* 全て正しく選択		*TfCon* 一部正しく選択		*TfCon* 誤って選択		*TfiDf* 全て正しく選択		*TfiDf* 一部正しく選択		*TfiDf* 誤って選択	
0.15	5	10.6%	27	57.4%	15	31.9%	5	10.6%	27	57.4%	15	31.9%
0.16	6	12.8%	26	55.3%	15	31.9%	9	19.1%	21	44.7%	17	36.2%
0.17	7	14.9%	22	46.8%	18	38.3%	10	21.3%	19	40.4%	18	38.3%
0.18	10	21.3%	17	36.2%	20	42.6%	13	27.7%	10	21.3%	24	51.1%
0.19	13	27.7%	11	23.4%	23	48.9%	14	29.8%	7	14.9%	26	55.3%
0.20	16	34.0%	9	19.1%	22	46.8%	17	36.2%	6	12.8%	24	51.1%
0.21	16	34.0%	9	19.1%	22	46.8%	18	38.3%	4	8.5%	25	53.2%
0.22	18	38.3%	7	14.9%	22	46.8%	17	36.2%	2	4.3%	28	59.6%
0.23	18	38.3%	5	10.6%	24	51.1%	16	34.0%	2	4.3%	29	61.7%
0.24	17	36.2%	4	8.5%	26	55.3%	16	34.0%	0	0.0%	31	66.0%
0.25	17	36.2%	3	6.4%	27	57.4%	16	34.0%	1	2.1%	30	63.8%
0.26	18	38.3%	2	4.3%	27	57.4%	16	34.0%	1	2.1%	30	63.8%

表5 意味カテゴリーの選択の精度と再現率

閾値	TfCon 正しく選択 (D)	TfCon 誤って選択 (E)	TfCon 選択せず (F)	TfCon 精度 D/(D+E)	TfCon 再現率 D/(D+F)	TfiDf 正しく選択 (D)	TfiDf 誤って選択 (E)	TfiDf 選択せず (F)	TfiDf 精度 D/(D+E)	TfiDf 再現率 D/(D+F)
0.15	36	78	15	31.6%	70.6%	36	79	15	31.3%	70.6%
0.16	35	67	16	34.3%	68.6%	33	67	18	33.0%	64.7%
0.17	31	57	20	35.2%	60.8%	32	55	19	36.8%	62.7%
0.18	28	46	23	37.8%	54.9%	25	43	26	36.8%	49.0%
0.19	25	38	26	39.7%	49.0%	23	35	28	39.7%	45.1%
0.20	24	31	27	43.6%	47.1%	21	27	30	43.8%	41.2%
0.21	23	28	28	45.1%	45.1%	19	19	32	50.0%	37.3%
0.22	23	23	28	50.0%	45.1%	16	14	35	53.3%	31.4%
0.23	20	20	31	50.0%	39.2%	14	13	37	51.9%	27.5%
0.24	18	18	33	50.0%	35.3%	12	12	39	50.0%	23.5%
0.25	17	15	34	53.1%	33.3%	11	10	40	52.4%	21.6%
0.26	13	10	38	56.5%	25.5%	10	6	41	62.5%	19.6%

がら良好な結果となっている。

次に、意味カテゴリーそれぞれの推定の成否に着目し、両手法の精度と再現率について集計した(表5)。各項目の意味を次に示す。

「正しく選択」…選択すべき意味カテゴリーを正しく選択した場合。

「誤って選択」…選択した意味カテゴリーが誤っている場合。

「選択せず」…選択すべき意味カテゴリーを選択できていない場合。

「精度」…意味カテゴリーを正しく選択した割合。

「再現率」…選択すべき意味カテゴリーを選択した割合。

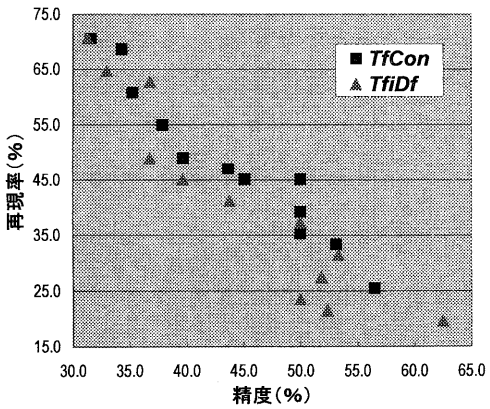

図1 意味カテゴリー推定の精度と再現率

ここで、「正しく選択」の個数と、精度、再現率を総合的に考慮すると、*TfCon* 法は、*TfiDf* 法と比較して、より高い結果を得ていると言えよう。例えば、精度が 50%となった箇所に着目すると、*TfCon* 法(閾値 0.22、0.23、0.24)は、*TfiDf* 法(同 0.21、0.24)よりも「正しく選択」の個数も多く、再現率も高い。

また、表4の数値をプロットした図1を見ると、精度と再現率の双方を、ともにできるだけ高くなるように閾値を設定する場合、*TfiDf* 法よりも、*TfCon* 法が適していることが分かる。

5.5 考察

以上を総合すると、*TfiDf* 法と比較して、*TfCon* 法は、良好に意味カテゴリーを選択していると言える。しかし、精度も再現率も高い数値であるとは言えず、今後の改善の必要性が認められる。

今回は、訓練に用いた文書の数が比較的少ないものであった。そこで今後、その数を増やして、推定の精度や再現率を改めて検証する必要があろう。また、当然ながら、属性の抽出や計算の方法にも、改善の必要があると言える。

6 まとめと今後の課題

今回、文書の分類や検索を容易にするために、文書を意味カテゴリーでインデクシングする方法

と、その作業を支援する手法について提案した。

そのために、各意味カテゴリーに該当する文書集合を形態素解析して、その結果から属性を抽出し、属性の頻度を求め、重み付けをして、その特徴を属性のベクトルで表現した。

新たに意味カテゴリーをインデクシングしようとする文書も、同様にベクトルで表現し、各意味カテゴリーのベクトルとの内積を計算して類似度を求め、選択すべき意味カテゴリーを推定した。

属性の頻度と重み付けの計算には、*TfiDf* 法を改良した *TfCon* 法を用いた。

TfCon 法の推定の精度や再現率を検証するため、*TfiDf* 法との比較を行ったところ、僅ではあるが良好な結果を得た。しかし、精度、再現率、ともにそれほど高くはなく、一層の改善の必要性が認められた。

この点については、まず、訓練に用いるデータ数を増やして、もう一度検討する必要があろう。また、データ数の増加とともに、推定の精度や再現率がどのように向上するかを確認することも重要である。それによって、学習に最低限必要なデータ数を同定することができるであろう。

属性の抽出についても、規則を見直す必要がある。現状では、比較的簡単な規則に従って属性を抽出している。そこで、多様な事例について検討して規則を精緻化し、より有用な属性を抽出できるようにすべきであろう。

TfCon 法についても同様に、頻度や集中度の計算について、様々な可能性を探る必要がある。

また、本研究では、できる限り簡便な手法を目指しているため、意味辞書や概念辞書の利用は避けているが、将来的には、それらの利用も検討する必要があるかもしれない。

今回提案した手法を用いることで、現段階では、比較的サイズの小さなテキストデータ(のブロック)が該当する意味カテゴリーを推定することが可能である。しかし本手法は、サイズの大きい文書の処理には向いていないと言える。

これは、本手法をそのままサイズの大きな文書に対して適用すると、該当する意味カテゴリーが文書全体から選択されしまい、それらが文書のどの部分に対応しているのかについては判別できないためである。

そこで、サイズの大きい文書に対するインデクシングの支援を行うためには、ある意味カテゴリーに分類されるべきテキストデータを同定する手法について、検討する必要があろう。これは、フレキシブルな範囲でテキストデータをサーチしていき、いずれかの意味カテゴリへの適合度が最大になるような範囲を求めて判定を行う手法を確立することで可能になると思われる。

7 参考文献

[1]金明哲、宮本加奈子(1999)、ラフな意味情報に基づいた文章の自動分類、言語処理学会第5回年次大会発表論文集、pp.235-238

[2]相澤彰子、影浦峡(1999)、グラフ的類似度尺度による学術文献の自動分類に関する検討、言語処理学会　第5回年次大会発表論文集、pp.217-220

[3]徳永健伸、岩山真(1994)、重み付き IDF を用いた文書の自動分類について、情報処理学会自然言語処理研究会、Vol.100、No.5、pp.33-40

[4]河合敦夫(1992)、意味属性の学習結果にもとづく文書自動分類方式、情報処理学会論文誌、Vol.33、No.9、pp.1114-1122

[5]福本文代、鈴木良弥、福本淳一(1996)、辞書の語義文を用いた文書の自動分類、情報処理学会論文誌、Vol.37、No.10、pp.1789-1799

[6]松本裕治、北内啓、山下達雄、平野善隆、今一修、今村友明(1997)、日本語形態素解析システム「茶筌」Version 1.5 使用説明書、NAIST Technical Report, NAIST-IS-TR97007