1C5-02

概念共起辞書を用いた文書種類によるクラスタリング

Document Clustering by Using a Concept-Base

友部 博教[*1] 石塚 満[*1]
Hironori Tomobe Mitsuru Ishizuka

[*1]東京大学情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo

In this paper, we propose an approach for document clustering by using a concept-base system. We represent features of documents based on a concept-base dictionary in which terms are described by relations to other terms. Then the features of documents are understood by comparing similarity between sentences by a concept-base dictionary to similarity by using this dictionary and standard vectors.

1. はじめに

我々は日々多くの情報に囲まれている。特に Web 上の情報は、ニュース記事から個人の日記まで、多種多様にわたっている。この膨大な情報の中から、我々は自分の必要な情報を発見しなければならない。

ユーザは Google[*1]や Yahoo![*2] といった検索サイトに検索語を入れれば即座に情報を獲得できる。しかし、ユーザは自分の求める情報に適合した検索語を入力しなければならない。欲しい情報が明確であり検索語がすぐにわかる場合には問題がないが、適当な検索語が見つからないと、ヒットするページ数が膨大になることや、逆に絞りこみによってほとんどページがヒットしないこともある。

特に、得たい情報を的確に見つけることができる検索語を探すことが困難である。自分の欲しい情報が見つかるまで、検索語を試行錯誤し検索を行い、検索結果から探す、という行動を繰り返すのが一般的である。検索語の選択には「コツ」があり、この「コツ」を持っているユーザは効率的に検索を行うことができる。例えば、表記ゆれに注意して検索語を選んだり、固有名詞など具体的な検索語によって絞りこみを行ったり、逆に一般的な語を併用することで絞込みを緩和するなどといった「コツ」がある。

では、なぜこのような「コツ」が必要なのだろうか。一般的に、検索語として人名などの具体的な単語一つで検索を行うと、非常に多くのページがヒットしてしまうため、それ以外にいくつかの言葉を足し合わせて検索することで絞込みを行っている。「コツ」はこのような追加した言葉にある。これらの言葉は、固有名詞と違って同義語や類義語がある場合が多い一般語が多く、正確にマッチしない限り検索結果として提示されないので、選択の仕方に「コツ」が必要となる。最初に入力する具体的な単語が求める文書の主題を表すのならば、追加した単語は主題とは別のものを表現していると考えられる。

そこで、このような一般語に注目する。「日記サイト」や「ニュースサイト」といった文書の種類ごとに、それぞれの一般語の使い方を持っている。これが文書の特徴である。つまり、一般語の使われ方を解析することによって、文書の特徴を発見することができる。

本稿ではこの文書の特徴を用いて、文書の種類によるクラスタリング手法について提案する。人手によって種類ごとに分けた文書集合からそれぞれの文書集合の特徴を抽出し、その特徴を元にクラスタリングを行う。ある主題について書かれている文書集合を文書の種類によって分類することでユーザの検索支援を行えると考えている。

連絡先: 友部博教，東京大学大学院情報理工学系研究科，03-5689-4718, tomobe@miv.t.u-tokyo.ac.jp

∗1 http://www.google.com

∗2 http://www.yahoo.com

2. 文書種類によるクラスタリング

本研究では、文書の主題ではなく、「日記」や「新聞記事」といった文書の種類によってクラスタリングを行う。ここではその手法について説明する。

文書の主題を表すのは、その文書の重要語である。重要語はその文書特有であるものが多く、したがって tf*idf 値が大きくなる。一方で、文書の種類を表すものは、ある種類の文書集合では頻出するが、他の文書ではあまり使用されない単語である。

しかし、ある文書集合でのみ頻度の高い単語の頻度に着目すればよいが、同義語や類語語、表記ゆれ等により、同じ意味で使われている単語であるのに別のものとしてカウントされてしまう。文書の主題のように固有の単語とは異なり、一般的な語であるためこのようなことが多くなる。

そこで、次のような概念共起辞書を用いることで、頻度の高い単語の属性を抽出することにした。

2.1 概念の共起辞書の特徴

この辞書は概念を他の概念との関連度によって表現している [1]。本稿で用いる共起辞書は次のように表現される。

$$word_i = (q_1, q_2, \cdots, q_n) \tag{1}$$

ここで q_m $(m = 1, 2, \cdots, n)$ は $word_i$ と $word_m$ の関連度である。この共起辞書は国語辞典の語義文を参照にして構成している。語義文に含まれる単語をまた引きすることによって得られる単語を組み合わせ、それを属性値として用意している。この概念辞書を用いることによって、検索精度を向上させることも期待できる [2]。

本研究では文書に出現する単語をこの辞書を用いて属性レベルにすることで、文書集合ごとによく出現する属性を発見することができる。

2.2 基準ベクトルの作成

まず、クラスタリングを行う上で基準となる基準ベクトルを作成する。それぞれの文書種類による基準ベクトルと文書を比較することによって、文書を分類することができる。

基準ベクトル作成には、文書集合を人手でクラスタリングした。文書の種類は「日記」「ポータルサイト」「用語解説」「新聞記事」の 4 つとした。

まず、概念共起辞書を用いて、単語を属性と属性値の集合で表現する。次に文書中の単語のすべての属性値を合成、正規化し、文書ベクトルを作成する。

$$Doc_i = word_{i1} + ... + word_{in} \tag{2}$$

そして文書集合 Cj 中の文書ベクトルを線形結合する。

$$V_{C_j} = Doc_1 + Doc_2 + ... + Doc_n \tag{3}$$

$$\tag{4}$$

こうして得られたベクトルをその文書集合における基準ベクトルとする。また、文書全体の基準ベクトル V_{all} も作成する。

2.3 基準ベクトルによる判別

対象となる文書ベクトルと基準ベクトルの類似度により文書種類を判別する。

まず、対象となる文書の文書ベクトルを作成する。次に文書ベクトルおよび基準ベクトル中に V_{all} と同じ属性が含まれている場合、その属性値を変調する。

$$\begin{aligned} Word_i^v &= (q_{i1}^v, q_{i2}^v, ..., q_{ij}^v, ..., q_{ik}^v) \\ q_{ij}^v &= \widehat{q}_{ij} \cdot M(\widehat{q}_{vj}) \end{aligned} \tag{5}$$

M は変調関数であるが、ここでは観点の属性の重みがある閾値を超えたときに重みを軽減する。

$$M(\widehat{q}_{vj}) = \begin{cases} r & for \widehat{q}_{vj} \geq 0.1 \\ 1 & for \widehat{q}_{vj} \leq 0.1 \end{cases} \tag{6}$$

S は V_all によって変調した基準ベクトルと文書ベクトルの類似度とする。ここでは類似度に基準ベクトルと文書ベクトルのなす角の余弦で計算する

$$\begin{aligned} S &= \cos\theta = V_{C_n} \cdot Doc_i \\ &= \sum_{j=1}^{k} q_{V_{cn_j}} q_{Doc_j} \end{aligned} \tag{7}$$

3. 実験と評価

基準ベクトル作成には、日記サイトから 40 件、ポータルサイト 15 件、ニュースサイト (記事) 40 件、用語説明 40 件を用いた。それぞれの文書集合から基準ベクトルを作成し、またこれらすべての記事から全文書の基準ベクトルを作成した。

文書判定には、4 つの基準ベクトルと変調類似度を算出し、類似度が一番高かったものを選ぶ。4 つの類似度がある閾値を越えない場合には判別不可能とする。表 1 は基準ベクトル作成に用いた文書、表 2 は基準ベクトル作成に用いていない文書の判別結果である。

また、「日記」と「新聞記事」の基準ベクトルの内容を表 3 に示す。「日記」サイトでは、「心」や「気持」といった属性を持つ単語が比較的多く使われることがわかる。一方「新聞記事」のサイトでは、「事柄」や「行動」といった属性を持つ単語が現れる。これらの属性が文書種類の特徴を表すことがわかる。

今回は基準ベクトルとして 4 つの文書種類の集合を用いたが、これを応用することによって、ある人の文書の特徴 (表現にどのような属性が多く用いられるか) がわかる。ある人の書いた文書を集めてきてそこから基準ベクトルを作成することにより、その人の表現の仕方の特徴がわかると考えられる。

表 1: 基準ベクトルに用いる文書の適合率と再現率

	日記	新聞記事	ポータル	用語解説
適合率 (%)	73.9	87.2	55.2	85.7
再現率 (%)	71.9	82.0	56.3	63.1

表 2: 基準ベクトルに用いていない文書の適合率と再現率

	日記	新聞記事	ポータル	用語解説
適合率 (%)	56.8	72.8	45.3	72.9
再現率 (%)	55.4	69.0	46.3	53.2

4. おわりに

本稿では概念の共起辞書を用いた文書種類によるクラスタリング手法について述べた。文書のキーワードによる文書分類とは異なり、文書の種類によってクラスタリングを行うことで、ユーザの検索支援を行うことができるだろう。

参考文献

[1] 笠原要, 松澤和光, 石川勉. 国語辞書を利用した日常語の類似性判別. 情報処理学会論文誌, Vol. 38, No. 7, pp. 1272–1283, 1997.

[2] 熊本睦, 島田茂夫, 加藤恒昭. 概念ベースの情報検索への応用. 情報処理学会・知識と複雑系研究会 (ICS-119-10), pp. 9–16, 1999.

[3] 友部博教, 石塚満. 概念の共起辞書を用いた文書特徴の抽出. 情報科学技術フォーラム (FIT)2002, pp. 153–154, 2002.

表 3: 基準ベクトルの主な属性

日記		新聞記事	
心	0.1899	所	0.1191
時間	0.1084	集まる	0.1032
表す	0.9608	行動	0.1026
様子	0.8776	一般	0.0916
考える	0.0821	全体	0.0895
他人	0.0734	事柄	0.0816
気持ち	0.0721		