3D-3

双クラスタリングと概念束を用いた大規模文書検索法の提案

渦尾秀勝 †　　森雅生 ‡　　伊東栄典 *　　廣川佐千男 *

† 九州大学システム情報科学研究院情報理学専攻
‡ 九州大学大学評価情報室　* 九州大学情報基盤センター　* 九州大学情報基盤センター

1　はじめに

近年、インターネットの普及によって web ページが増加の一途を辿っている。web ページが増加すると、利用者は閲覧すべき web ページを検索したり、目的の web ページに辿り付く事が困難になる。検索エンジンの典型例として google の web 検索やじゃらんネットのホテル検索などが挙げられる。google の web 検索は PageRank 方式を使用しており、検索語を含む web ページを順位付けで表示する。またじゃらんネットのホテル検索では、利用者は、地域、部屋タイプ、日付、価格などの条件を指定し、条件に一致したホテルの部屋の一覧が表示される。

現状の検索エンジンでは、利用者に多くの事前知識が期待されている。例えば、利用者は、検索する対象を予め分かっていないと検索できない。検索する対象が明確でない場合、何を検索語とすべきか利用者が分からない。あるいは、利用者が検索語を入力、もしくは条件を選び検索しても、検索結果が空の場合もある。検索結果が空だと、他の可能な検索質問を自分で考え、試行錯誤を繰り返し関連する知識を獲得しなければ、目的の対象まで辿り着けない。

我々は、このような問題点を解決するために概念束による分類を用いた検索法を開発している。しかし、文書や単語の数が増えると実用的な時間での概念束構築が困難となる。そのため概念束を用いた検索も困難となる。そこで本稿では、文書と単語を同時にクラスタリングする双クラスタリングを利用した大規模文書群の検索方法を提案する。双クラスタリングにより、文書と単語のサイズを小さくし、実用的な時間で概念束が構築可能になる。

2　概念束の定義と構築方法

概念束は、Rudolf Wille ら [4] が提案したものである。概念束は、これまで、データマイニング、情報検索、知識管理で用いられてきている。文書を対象、単語を属性とみなし、対象と属性の行列で構築される概念束は、対象-属性間の関係から概念の包含関係、あるいは属性間の含意関係などを導出することができる有効な分類手法である。本稿では、概念束を文書の検索に適用する。以下に概念束の定義と構築方法を述べる。

G を対象の集合、M を属性の集合とするとき、$G\times M$ の部分集合 $I\subseteq G\times M$ を文脈という。$(g,m)\in$I となる $(g,m)\in G\times M$ を gIm と表し、対象 g は属性 m を持つという。

文脈 (G,M,I) が与えられたとき、

$\{(A,B) \mid A\subseteq G, B\subseteq M,\ A'=B, B'=A\ \}$

となる組 (A,B) を集合の包含関係により順序付けした束を概念束という。ここで

$A'=\{\ m\in M \mid$ 任意の $g\in A$ に対し $gIm\ \}$

$B'=\{\ g\in G \mid$ 任意の $m\in B$ に対し $gIm\ \}$

である。

次に概念束の構築方法について述べる。概念束グラフの構築方法は四つの過程から成り立っている。

1. 概念をすべて求める
2. 概念の上下関係を決定する
3. 概念のすぐ上の概念を求める
4. 概念束のグラフを作成

過程 1〜3 において、すべての概念がどのような上下関係を持っているかが得られる。概念をノードとし、上下関係をもとにエッジで繋げば、概念束グラフを構成できる。概念束の構築方法は、他にも様々な方法 [2] があるが、本論文ではこの方法を採用した。

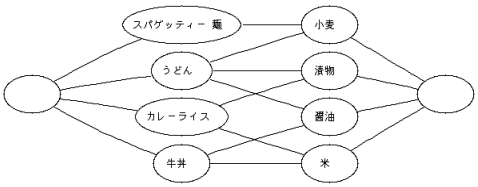


図 1: 概念束の例

3　双クラスタリング

文書検索のために構築される概念束を素朴なアルゴリズムで計算すると、べき時間かかる事が知られている。概念束は、対象となる文書データが追加されたり、削除されたりすると、はじめから再構築しなけらばならないので、頻

繁に更新のあるデータに対しては、構築時間を短縮する必要がある。

そこで我々は、Dhillon ら [3] の提案した、双クラスタリングを用いることにより、文書クラスタと単語クラスタの両面から行列サイズを縮小し概念束構築の性能向上を提案する。

双クラスタリングでは、文書クラスタと単語クラスタを同時に行う。青野ら [1] によると、高次元の関係の高い部分行列成分が、双クラスタリングすることによって、低次元の部分行列成分にクラスタリングされる。双クラスタリングによる影響は局所的であり大域的ではなく、行列の要素は、疎性を保ち、かつ、負の値にならないという特徴をもつ。元来、概念束を構築するための行列の要素は 1 か 0 であり、負の値をもたないので双クラスタリングの上記の性質は良い特徴であるといえる。

4 提案する検索アーキテクチャ

4.1 対象となるデータについて

我々が提案する手法は、文書とその文書が持つ様々な単語を定義できるデータの概念束構築に適用できる。大規模文書の実例として携帯電話のカタログデータを扱っているが、文書が 174 個、単語が 751 個であり、174×751 行列になっている。他にも、宿泊施設予約サイトのじゃらん net で閲覧することができるホテルのデータにも適用可能である。文書をホテル名とし、単語をそのホテルが持つ機能とすればホテルに関する概念束を構築できる。他にも、ネットショップのカタログデータにも適用できると予想される。

このような大規模文書をもとに概念束を構築するのは膨大な時間がかかってしまう。そこで大規模文書の場合における検索法を提案する。

4.2 双クラスタリングと概念束

n 個の文書 $D_i(i=1,\ldots,n)$ と、m 個の単語 t_{j} ($\mathrm{j}=1,\ldots,m$) が与えられたとする。このとき、$n\times m$ の行列 $\{c_{ij}\}$ を考える。D_i に t_i が含まれているとき、$c_{ij}=1$ とし、そうでなければ、、$c_{ij}=0$ とする。今 $n\times m$ の行列を考える。行は文書、列は単語である。文書をとし、単語を t_j とする。このとき、文書 D_i に単語を t_j が含まれているならその行列の ij 成分を 1 とし、含まれていないならば 0 とすることにより、$n\times m$ の行列を定義できる。

30×30 の次元の低い行列ならば、この行列をもとに概念束を構築し検索できる。前述したように、本稿で行列は次元が高いものを扱う事を前提としているために概念束構築が困難である。そこで、双クラスタリングを利用することで、行列サイズを縮小する。サイズが高々30×30 に縮小されれば、概念束構築が可能となり、検索が行える。高次元行列に双クラスタリングを利用する。

双クラスタを行うことにより、複数の行、列を一つの行に変換するという操作が行われ、双クラスタリングする前の行列よりも小さい行列が作成される。一つの行、列に変換される際には、変換される前の行、列の要素である 1 の個数をもとに我々の定義に従ってクラスタ後の要素が 1 であるか 0 であるかを決定する。クラスタ後の文書と単語を次のように定義する。$\hat{d}=\{h_1,h_2,h_3,h_4,\ldots,h_i\}$ を i 個の文書クラスタとし、$\hat{t}=\{g_1,g_2,g_3,g_4,\ldots,g_j\}$ を j 個の単語クラスタとする。これにより、$i\times j$ の行列が定義される。h は複数の文書、g は複数の単語を含むことになる。この双クラスタ後の $i\times j$ 行列をもとに概念束を構築する。

双クラスタリング後の一つの文書には、複数のクラスタする前の文書が含まれるため、概念束をもとに検索を行うと、複数の文書群に辿り着く。検索を行い検索対象に辿りつけない場合には、再びたどり着いた文書群と単語群をもとに行列を作成し、概念束を構築し検索を行う。概念束構築困難な場合には、再び双クラスタリングを利用する。この操作の繰り返しにより、利用者は目的の検索対象に辿りつけるであろう。

5 おわりに

本稿では、概念束の理論と双クラスタリングを用いた文書検索法を提案した。文書と単語から作られる行列のサイズが小さければ、行列から概念束を構築し、検索できる。文書数が大きい場合、行列のサイズも大きくなり概念束構築が困難になる。本稿では、双クラスタリングすることにより、行列のサイズを縮小することで、概念束構築し、検索する手法を提案した。

参考文献

[1] 青野雅樹, 土肥広典, 文書ー単語双クラスタリングを用いた特許データの概念検索性能向上手法について, 豊橋技術科学大学情報工学系.

[2] Cloudio Carpineto, Giobanni Romano, Concept Data Analysis Theory and Application, ISBN-13: 978-0470850558.

[3] Iderjit S.Dhillon and Y.Guan, Information Theoretic Clustering of Sparse Co-Occurrence Data, Proceedings of The Third IEEE International Conference on Data Mining, pages 517-520, November 2003.

[4] Bernhard Ganter, Rudolf Wille,C .Franzke, Fomal Concept Analysis:Mathematical Foundation, ISBN-13: 978-3540627715.