D-009

ZDD を用いた頻出パタン演算による Web テキストデータからの知識発見とその評価

ZDD-based Processing of Frequent Patterns for Knowledge Discovery in Web Text Data

岡崎 佑太 † Yuta Okazaki
湊 真一 † Shin-ichi Minato

1. はじめに

本研究では Web 上で配信されるテキスト情報から話題性を取り出すために，テキスト中に現れる頻出パタンに着目した．ある時点で話題になっている事柄はその前後で多く言及されていると考えることができる．しかし単純に頻出パタンを抜き出すと，もともと広く使われる一般的な語のパタンが選ばれてしまい，本来求めたかった情報が取り出せない．

この問題に対し Web 検索の分野では，tf-idf モデルという頻度に基づいた特徴量による解析 [4, 5] を用いることが多いが，本研究では頻出パタンをゼロサプレス型二分決定グラフ (ZDD: *Zero-suppressed Binary Decision Diagrams*) と呼ばれるグラフ構造で表現し，ZDD 同士の演算によってこれに似た機能を実現する．また同じく ZDD の演算処理を用い，頻出パタンの時間的変化という観点からも話題性抽出を試みた．実験では，Web ニュースサイトの記事見出しから頻出パタンを抽出し，ZDD のパタン演算による話題抽出について調査した．

2. 頻出パタン抽出

テキストデータ	今年初戦へ…上村愛子，早くも緊張感
	⇓
語へ分解	今年/初戦/へ/…/上村/愛子/，/早く/も/緊張/感
	⇓
名詞のみ抽出	今年/初戦/上村/愛子/緊張/感
	⇓
アイテム集合	a b c d e f

表 1　テキストデータからアイテム集合への変換

Web ニュースサイトの記事見出しに対し表 1 のように変換を施し，一つの語をアイテムとするアイテム組合せ集合を作る．ある期間中に配信された記事見出しは，複数のアイテム集合をリスト化したデータベースと考えられる．

本研究で扱う頻出パタンとは，データベース中に最小頻度 α 回以上出現する，アイテムの部分集合である．いくつかのデータベースから図 1 のように頻出パタンを抽出し，常に頻度の大きな一般的な語のパタンや，突発的にある期間だけ頻出するパタンを抽出することができる．しかし一般にアイテムが k 個あるとするとパタンは 2^k 通りにも及ぶので，大規模なデータベースから頻出パタンを抽出することは容易ではない．

ID	レコード
1	a b
2	a
3	a c
4	b
5	a b c
6	a b
7	b
8	b c

頻出パタン集合

α=6 ⟶ {b}
α=5 ⟶ {a, b}
α=3 ⟶ {a, b, c, ab}

図 1　データベースから頻出パタンを抽出

3. 二分決定グラフを用いた頻出パタン表現

3.1 BDD

BDD(*Binary Decision Diagrams*)[1] とは図 2 のように，論理関数をグラフによって表現したものである．論理関数におかえる各々の変数に 0,1 の値を代入した結果は，二分木のそれぞれの枝に対応する．論理関数の解は，二分木の葉，2 値の定数接点で表す．

BDD は変数の順序を固定し，冗長な接点の削除と等価な接点を共有することで，既約な形が一意に決まることが知られている．この簡約化規則によって，論理関数を比較的コンパクトなデータ構造として扱うことができる．

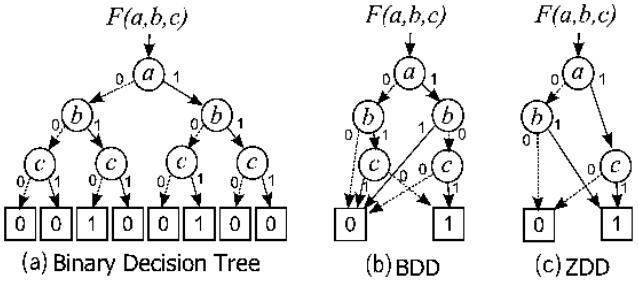

図 2　二分決定木，BDD，ZDD

† 北海道大学大学院情報科学研究科

3.2 ZDD

一方，組合せ集合を表現するには ZDD と呼ばれるグラフ表現を用いると効率良く扱うことができる．ZDD での簡約化規則では 1-枝が 0-終端接点を直接指している接点を削除する．これによって，組合せ集合に選ばれることのないアイテムに関する接点が自動的に削除されることになり，BDD よりもコンパクトなデータ構造が得られることが期待できる．

3.3 LCM over ZDDs

LCM(*Linear time Closed item set Miner*) は宇野ら [3] によって開発された，出力サイズに対して線形時間で頻出アイテム集合を列挙する高速なアルゴリズムである．LCM は深さ優先探索に基づくアルゴリズムの 1 つで，既に求められた頻出パタンにアイテムを 1 つずつ追加しながら再帰的な処理を繰り返すことにより全ての頻出パタンを列挙する．

最近ではこの LCM で求められた計算結果の頻出パタンを ZDD として出力する，LCM over ZDDs[2] というアルゴリズムが提案されている．LCM で求めた膨大なアイテム集合は，ZDD によってコンパクトに圧縮して表現できる．

4. ZDD を用いた頻出パタン演算

4.1 差集合

頻出パタンを抽出して話題となっている語を取り出す前に，常に頻度の大きい一般的な語のパタンを除外しなければならない．これは ZDD の演算を用いて，頻出パタン集合の差分を求めれば良い．例えば最小頻度 x 回以上出現するものを一般的な語のパタンとする場合，まず LCM over ZDDs により頻出パタン集合 F_x を ZDD として抽出する．次に，話題となっている頻出パタンの最小頻度を $y(< x)$ とし，同様に F_y を求める．$F_x \subseteq F_y$ なのでこれら 2 つの ZDD より，一般的な語のパタンを除いた話題パタンは $F_y \setminus F_x$ で表される頻出パタン集合から導かれる．

4.2 包含関係

また ZDD は複数の頻出パタン集合との共通集合を求めることができる．ある期間配信されたテキスト情報の頻出パタン集合を F_m，また別の期間のものを F_n とすると，両期間とも頻出なパタン集合は $F_m \cap F_n$ で表される．F_m, F_n さえ求めてしまえば，ZDD の高速な演算によって容易に計算できる．本研究では時間的変化を伴なう話題を発見するため，いくつもの ZDD の間でこれらの演算を行う．

5. 実験

5.1 準備

本実験では Web 上のニュースサイトが配信するニュースの記事見出しをテキスト情報とした．記事見出しを解析することで，話題となっているニュースと関連するパタンが取り出せるか確かめる．まず国内大手と思われる Yahoo!JAPAN ニュース [7]，スポーツニッポン新聞社 [8]，サンケイスポーツ [9] が配信するスポーツニュースを 2 ヶ月分を 1 日ごとに取得し，それぞれ 1 つのデータベースとして保存する．記事見出しには記号や助詞など，話題性とは直接関係しないと考えられる語も含まれるので，MeCab[6] で形態素解析を行い名詞だけを抽出した．

次にこの名詞の列に対しアイテム番号をつける．同じ名詞には一意に番号を付けなければならないので，ハッシュテーブルによって名詞とアイテム番号の対応表を実装した．記事見出しを取得し続けるとアイテムの総数は膨大になるが，多くなり過ぎる場合は何らかの方法でアイテム集合を絞り込む必要がある．今回収集した 2 ヵ月分のデータベースでは 7,870 個の名詞が出現したが，我々の ZDD 処理系では約 60,000 個までのアイテムを扱えるので，特に絞り込みは行っていない．

5.2 一般語のフィルタリング

まず従来の tf-idf 手法により，記事見出しに現れる特徴的な単語を調べてみると，表 2 のようになった．特に，ある時期に話題となった人名が特徴的と計算されることが多かった．

また一般語の指標として，DF(*Document Frequency*) だけを用いその上位を取り出すと表 3 のような結果になった．本実験における DF とは，多くの記事見出しに出現する語となる．スポーツニュースではよく見る順位を表す語や「五輪」,「日本」といった語が抽出された．

語	tf-idf
貴乃花	0.00155607565580147
国母	0.00140000293185235
真央	0.00134526917369851
一門	0.00131717224152507
ヨナ	0.00128976394784498
娘	0.00127514880966186
藍	0.00123846044074029
安治川	0.00118798311138787
理事	0.00115884278366767
池田	0.00114600424100314

表 2　tf-idf モデルによる特徴的な語（上位 10 語）

次に表 4 に示すように，各日のデータベースに対し，ZDD の演算を用いて最小頻度 10 回以上現れるパタンを抽出した．DF を用いて抽出した一般語と同様,「位」や「五輪」といった語が多くの日で頻出していることがわかる．各日の頻出語をもとに，ZDD の集合演算を用いて一般語のフィルタリングが行えると考えられる．また ZDD を用いると，一単語だけでなく一般的な語の組合せに関しても抽出することができる．

5.3 特定期間の頻出パタン

次に，ある期間中だけ記事見出しに頻出するパタンを抽出した．表 5 では，3 日間だけ連続して現れるがそれ以外の日には一度も現れないパタンの例を示す．なおパタンの部分集合は省略している．記事見出しに現れる語の

語	DF
位	1319
五輪	610
日本	509
朝	481
青龍	447
戦	415
遼	391
人	362
年	348
くん	318

表 3　DF を指標に抽出した一般語（上位 10 語）

日	パタン
1	{W 杯 }, { 戦 }, { 位 }
2	{ 代表 }, { 五輪 }, { 位 }
3	{ 魁皇 }, { 勝 }, { 五輪 }, { 位 }
4	{ 首位 }, { ウッズ }, { 日本 }, { 位 }
5	{ 千代, 大海 }, { 千代 }, { 大海 }, { 引退 }, { 五輪 }
6	{ 連勝 }, { 五輪 }, { 位 }
7	{ 五輪 }, { 位 }

表 4　ZDD 演算により抽出した頻出パタン例（最小頻度 10）

中でも特徴的なものが抽出できた．3 日間以外には現れないものを抽出しているので，先の一般語のフィルタリングも効いている．

例えば，1〜3 日目だけ頻出で 4〜60 日目には一度も現れないパタンを抽出する ZDD 演算処理を行った実験では，途中に生成した 60 日分の頻出パタンの総数は 315159 個であったが，それらを表現する ZDD のサイズ（ノード数）の総和は 37333 であり，この例では 10 倍近く圧縮されている．これらの演算処理に要した計算時間は，CPU に Intel Core2 Duo E8400，主記憶 4GByte，OS Ubuntu 8.04 LTS という環境で 0.0264 秒であった．

日	パタン
1〜3	{ ジェームズ, 活躍 }, { 古閑 }, { 化, 貴乃花 }, { 化, 親方 }, { 魁皇, 連勝 }, { 女王, 初 }, { 池田, くん, 遼, 位 }, { 高木, 五輪 }, { 開幕, 位 }
2〜4	{ アジア, 大会 }
3〜5	ø
4〜6	{ 最多, 葛西, 連続, 大会, 代表 }

表 5　3 日間だけ頻出するパタン例

6. 結論

本研究では Web 上のニュース記事見出しから頻出パタンを ZDD としてコンパクトに抽出し，ZDD の特徴でもある高速な演算による話題発見の可能性を調査した．複数のデータベースから ZDD の演算を用いて，話題となっているニュースに関連する語のパタンが取り出せた．また複数のデータベースに共通して出現する語を取り除くことで，一般語のフィルタリングもできることがわかった．特に本研究で扱ったスポーツニュースは，大会などのイベント開催前後にニュースが多く配信されたり，選手の活躍や大会後の結果を報じるニュースが集中するため，本手法でうまく話題性を抽出できたと考えられる．

しかし語の解析にはまだ工夫の余地が残っている．特にスポーツニュースでは人名が多く，現状 MeCab による解析精度は良くない．また同義語を同一のアイテムとみなすことで，より広く話題性を取り出すことができるかもしれない．

参考文献

[1] Randal E. Bryant, ”Graph-Based Algorithms for Boolean Function Manipulation” IEEE Transactions on Computers, Vol.C-35, No.8, pp.677-691, 1986.

[2] S. Minato, T. Uno and H. Arimura, ”LCM over ZBDDs: Fast Generation of Very Large-Scale Frequent Itemsets Using a Compact Graph-Based Representation,” in Advances in Knowledge Discovery and Data Mining, pp.234-246, 2008.

[3] T. Uno, M. Kiyomi and H. Arimura, ”LCM ver.2: ecient mining algorithms for frequent/closed/maximal itemsets,” In Proc. of IEEE ICDM ’04 Workshop FIMI ’04, 2004.

[4] 相澤 彰子，”低頻度語の利用によるテキスト分類性能の改善と評価”，情報処理学会論文誌，44(7), pp.1720-1730, 2003.

[5] 和多 太樹，関 隆宏，田中 省作，廣川 佐千男，”単語の出現頻度に着目した病院評判情報の分析”，情報処理学会研究報告，pp.15-20, 2005.

[6] ”MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, http://mecab.sourceforge.net/

[7] ”Yahoo!JAPAN ニュース”, http://headlines.yahoo.co.jp/

[8] ”スポニチ Sponichi Annex”, http://www.sponichi.co.jp/

[9] ”SANSPO.COM”, http://www.sanspo.com/