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

北海道大学大学院情報科学研究科　○岡崎佑太，湊真一

概要

- 研究背景
- 頻出パタン抽出
- BDD, ZDD
- ZDD演算
- 実験
- 結論

研究背景

- RSSニュース配信やTwitterなどの普及
- 膨大なテキストデータからの知識発見
- 流行・話題性の抽出

頻出パタン抽出

ID	タプル
1	abc
2	abd
3	abc
4	bc
5	c
...	

最小頻度αを与え、DB中にα回以上出現するパタンを列挙

頻出パタン抽出

ID	タプル
1	abc
2	abd
3	abc
4	bc
5	c
...	

$\alpha=5$

パタン	頻度
b	8
c	8
bc	7
a	5

最小頻度αを与え、DB中にα回以上出現するパタンを列挙

頻出パタン抽出

ID	タプル
1	abc
2	abd
3	abc
4	bc
5	c
...	

$\alpha=5$

パタン	頻度
b	8
c	8
bc	7
a	5

最小頻度αを与え、DB中にα回以上出現するパタンを列挙

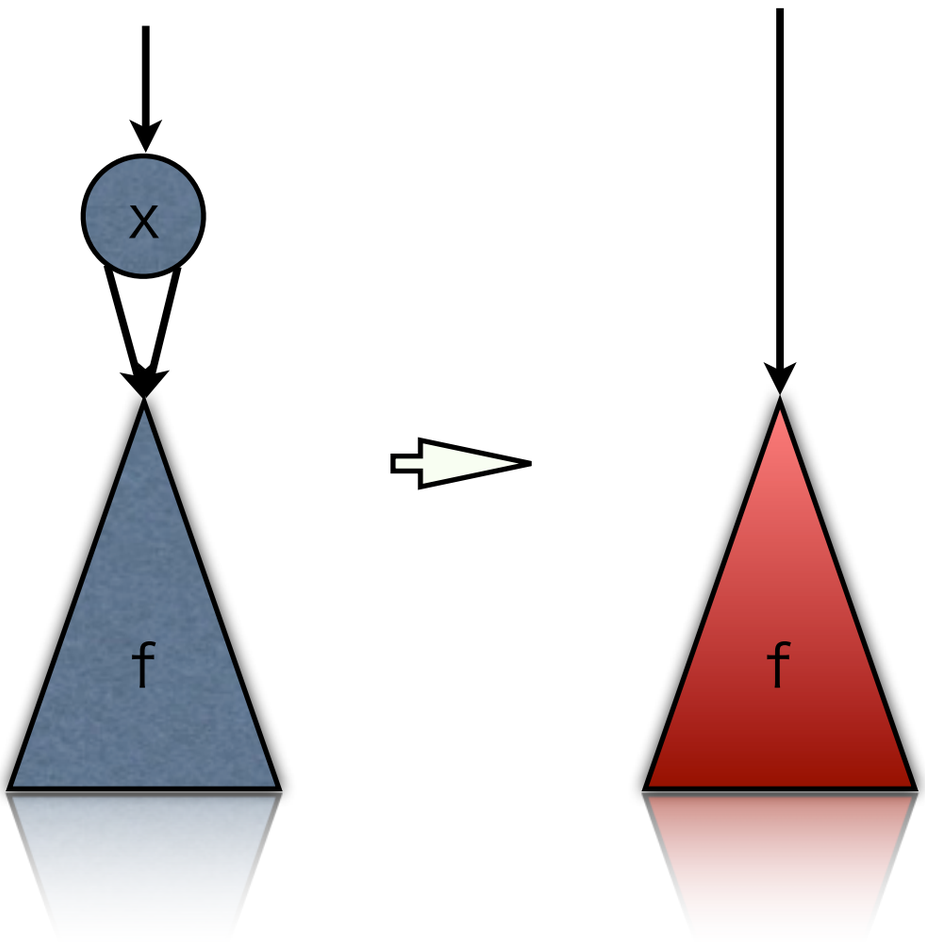
BDD

Binary Decision Diagrams [Bryant1986]

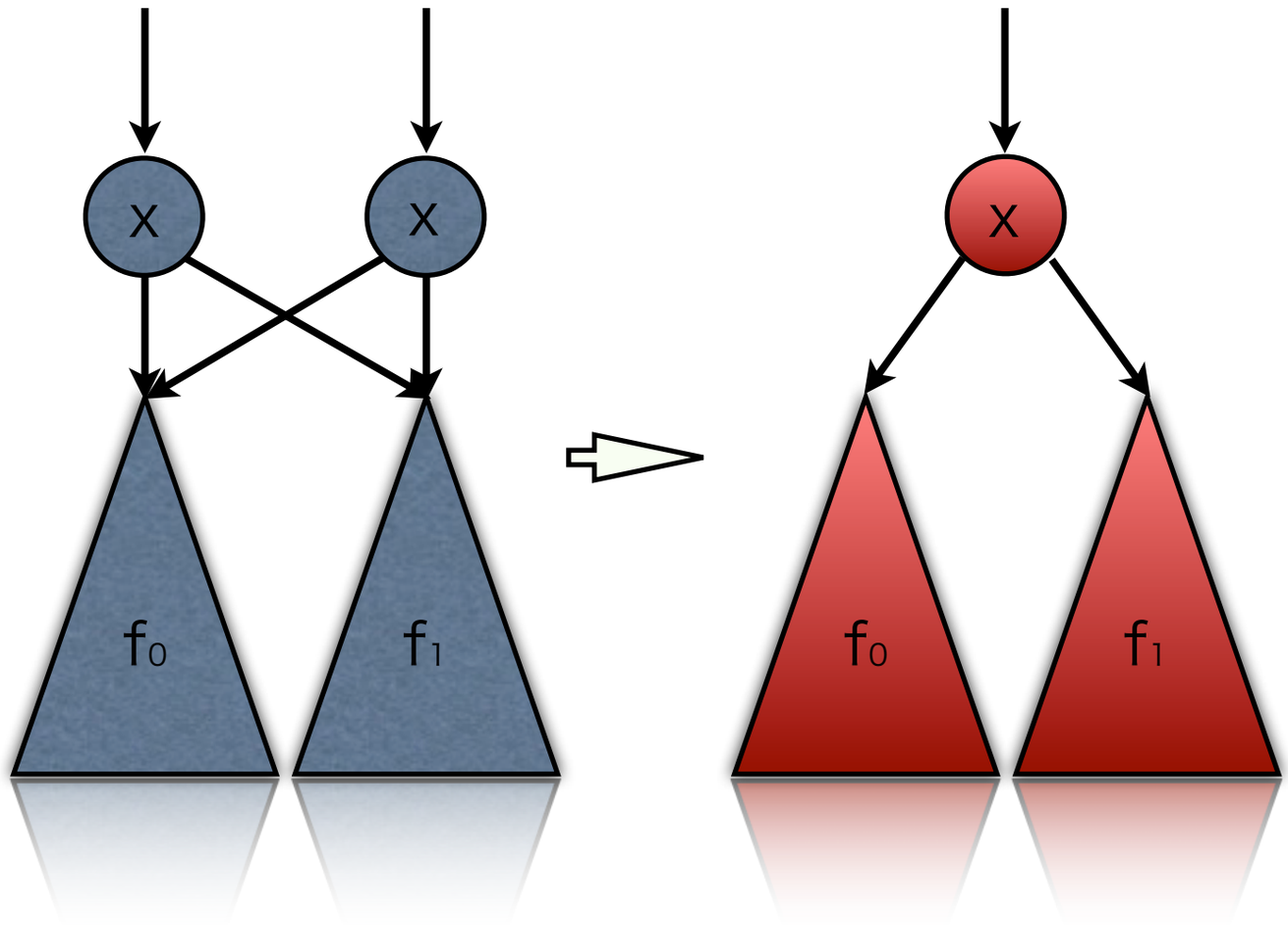
二分決定グラフに以下の簡約化規則を適用

- 冗長な節点の削除
- 等価な節点の共有

既約な形が一意に得られる



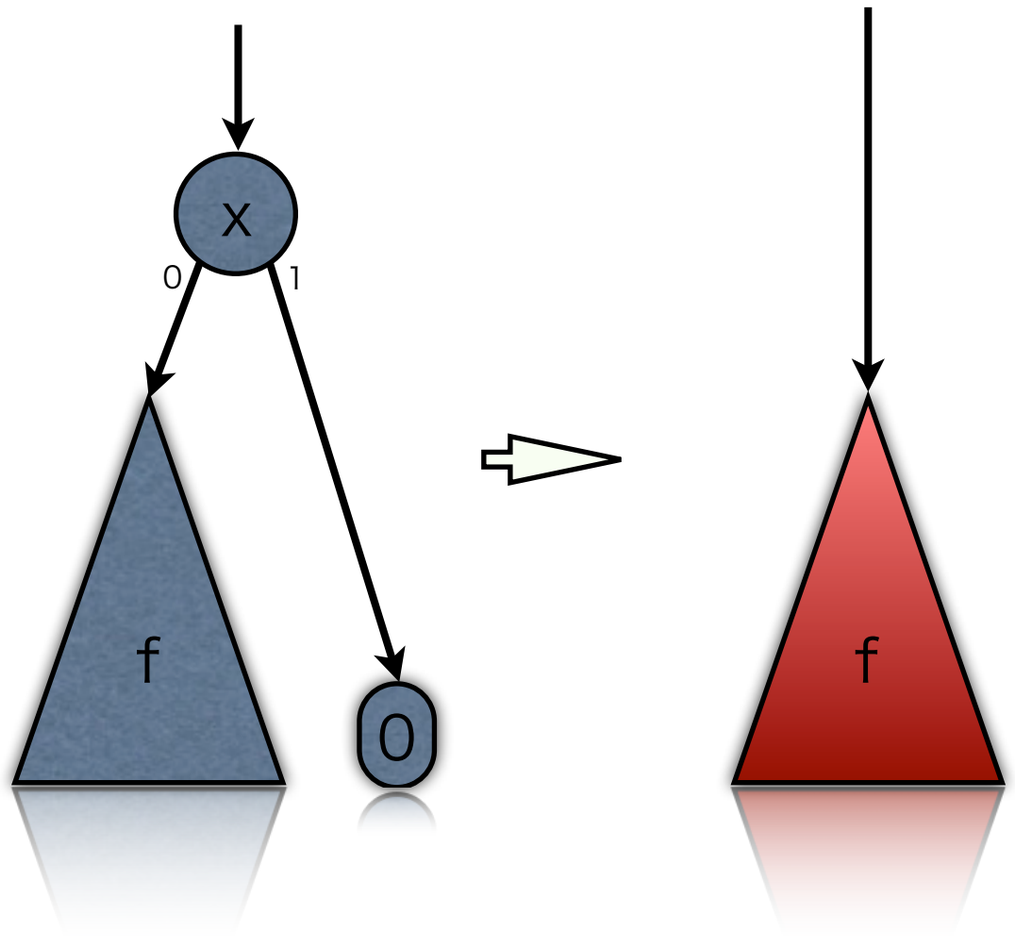
冗長な節点の削除



等価な節点の共有

ZDD

Zero-suppressed BDD [Minato1993]



1-枝が0-終端節点を直接指している節点を削除

LCM

Linear time Closed itemset Miner [Uno2003]

- 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アルゴリズム
- 深さ優先のバックトラック法
- ZDDで出力するLCM over ZDDs [Minato他2008]

ID	タプル
1	abc
2	ab
3	abc
4	bc
5	ab
6	abc
7	c
8	abc
9	abc
10	ab
11	bc

ID	タプル
1	abc
2	ab
3	abc
4	bc
5	ab
6	abc
7	c
8	abc
9	abc
10	ab
11	bc

LCM

$\alpha = 7$

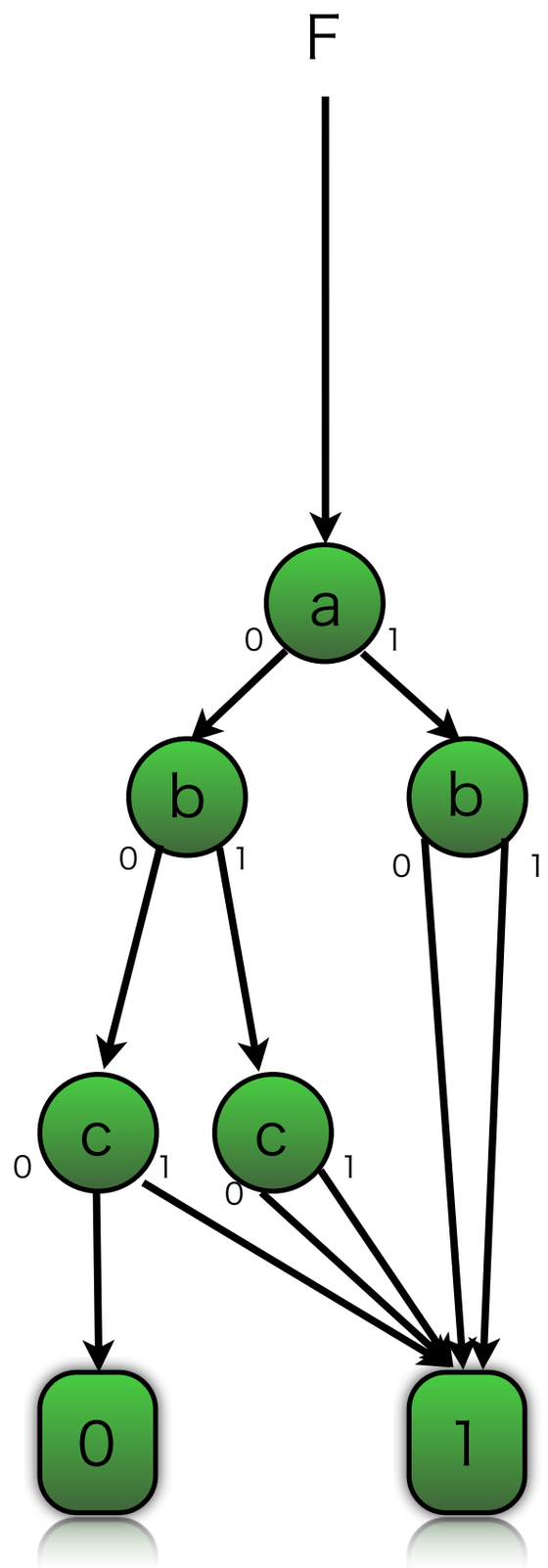
{ab, bc, a, b,c}

ID	タプル
1	abc
2	ab
3	abc
4	bc
5	ab
6	abc
7	c
8	abc
9	abc
10	ab
11	bc

LCM over ZDDs

$\alpha = 7$

{ab, bc, a, b,c}

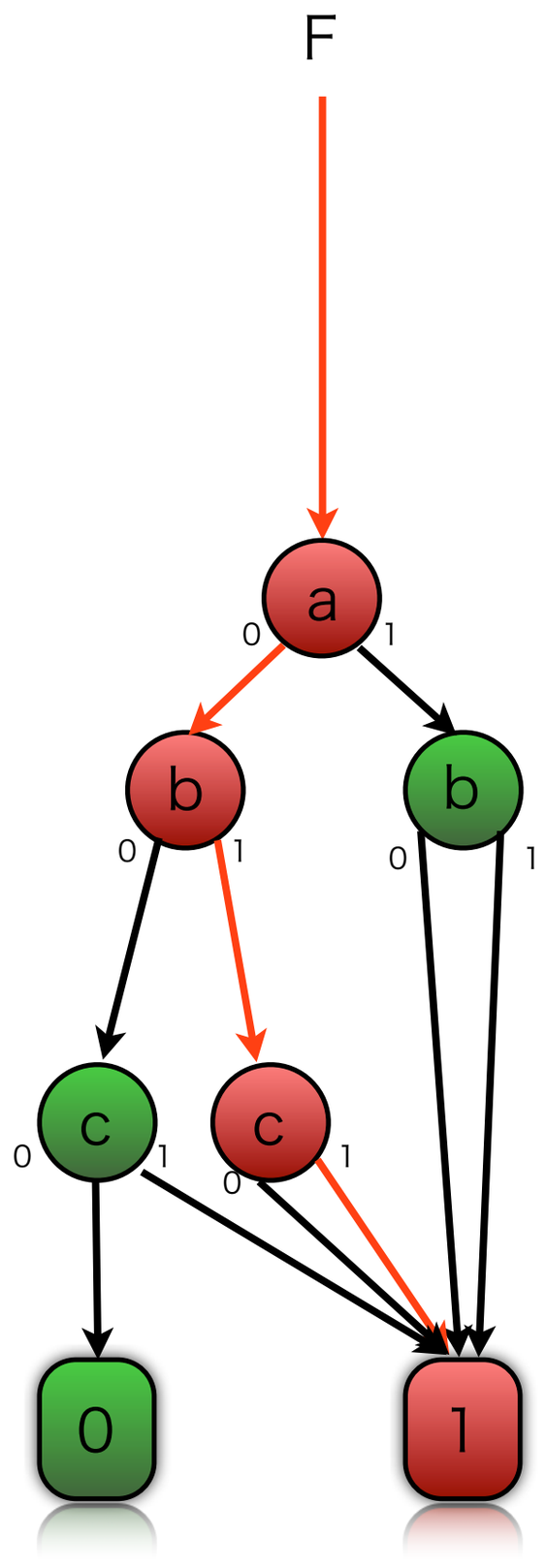


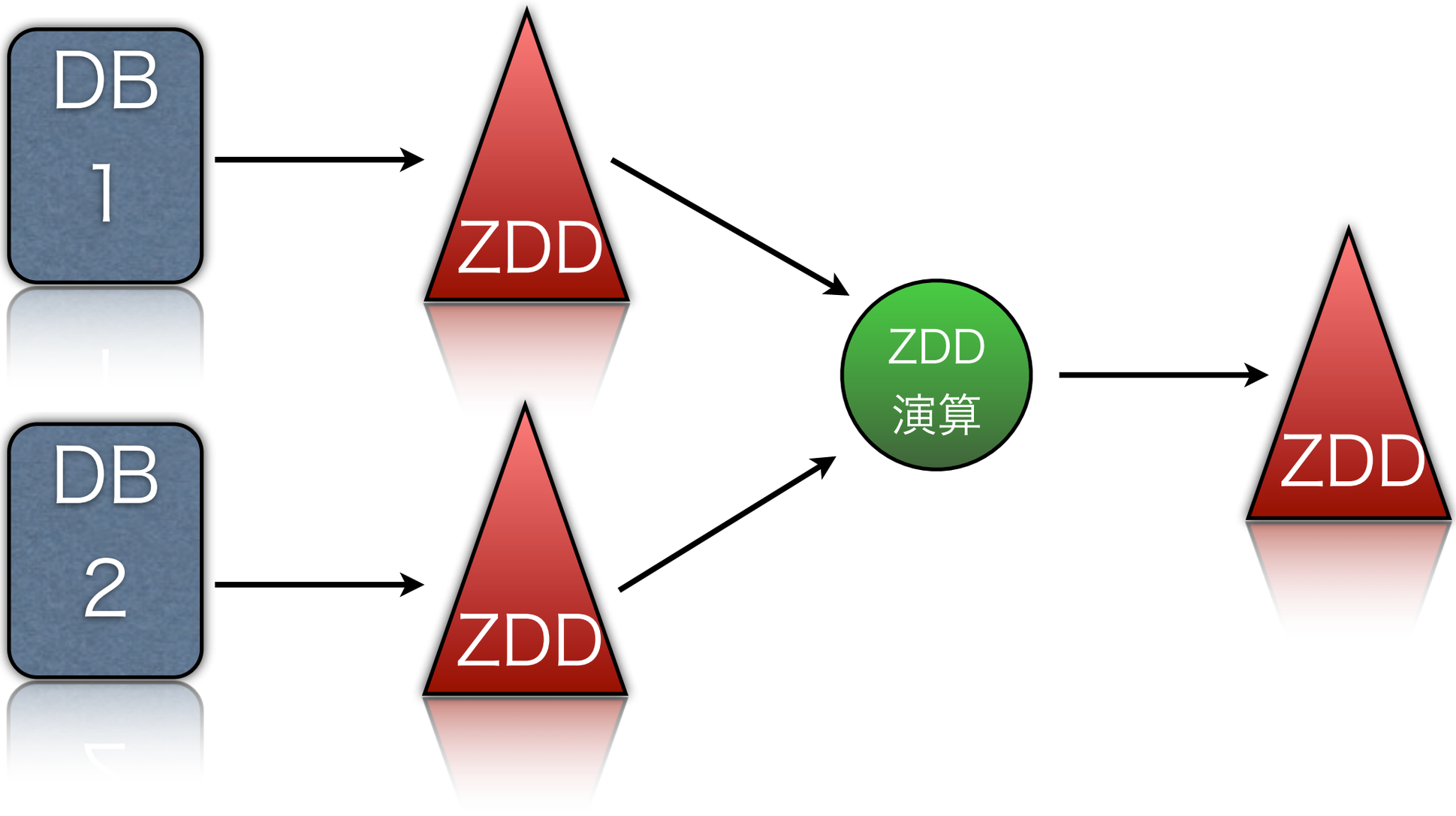
ID	タプル
1	abc
2	ab
3	abc
4	bc
5	ab
6	abc
7	c
8	abc
9	abc
10	ab
11	bc

LCM over ZDDs

$\alpha = 7$

{ab, bc, a, b,c}





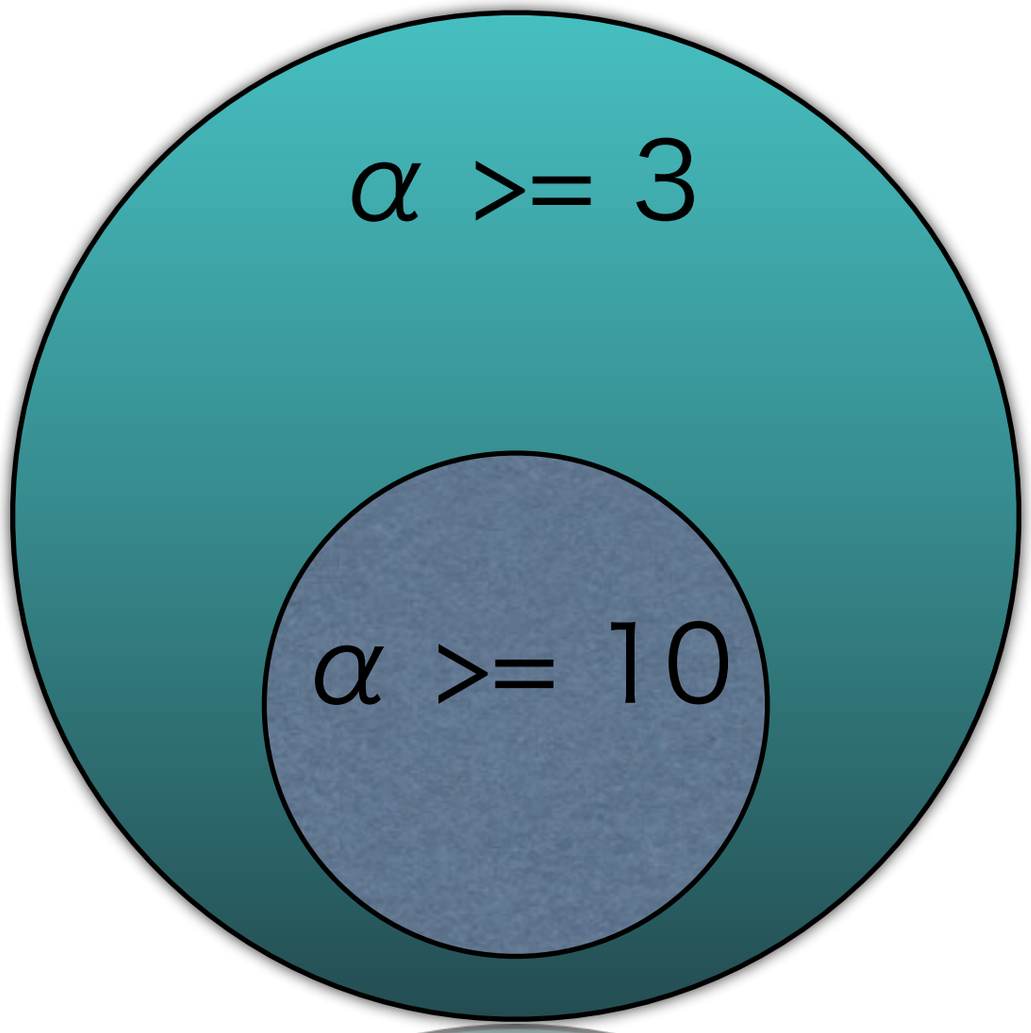
ZDD演算

ZDD演算 - 差集合

例...

10回以上出現するパタンを一般的な語と決め、それを除く3回以上出現するパタンを抽出

ZDD演算 - 差集合



例...

10回以上出現するパタンを一般的な語と決め、それを除く3回以上出現するパタンを抽出

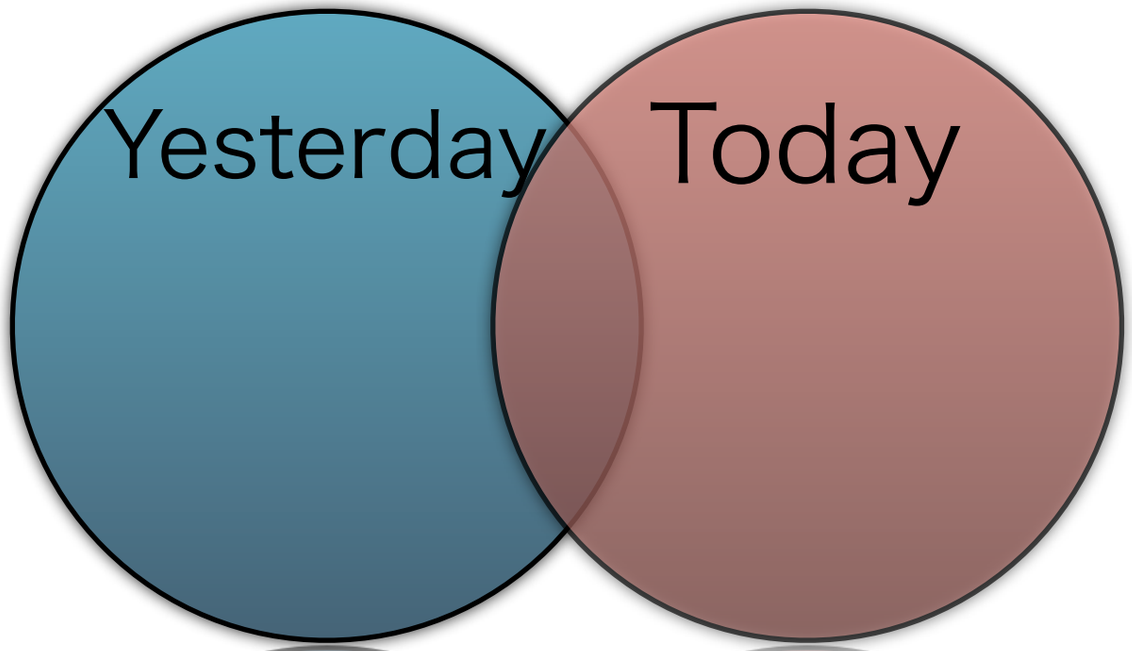
ZDD演算 - 共通集合

例...

昨日、今日と続けて頻出なパタンを抽出

ZDD演算 - 共通集合

例...



昨日、今日と続けて頻出なパタンを抽出

ZDD演算

差集合・共通集合を求めることにより

「x日間だけ頻出で、それ以外の期間は出現しない」

パタンの抽出が可能

実験

実験

- 一般語フィルタリングの可能性

実験

- 一般語フィルタリングの可能性

- x日間だけ頻出なパタンが話題性を示すか

実験データ

以下3サイトが配信するスポーツニュースの記事見出しを1日ごとに2ヶ月分取得

1.Yahoo! JAPAN ニュース headlines.yahoo.co.jp
2.スポーツニッポン www.sponichi.co.jp
3.サンスポ www.sanspo.com

1日分のデータベースを複数用いて実験を行う

欧州女王復活 真央ちゃんにも脅威!?

欧州女王復活 真央ちゃんにも脅威!?

形態素解析(MeCab)

欧州 / 女王 / 復活 / 真央 / ちゃん / に / も / 脅威 / ! / ?

欧州女王復活 真央ちゃんにも脅威!?

形態素解析(MeCab)

欧州 / 女王 / 復活 / 真央 / ちゃん / に / も / 脅威 / ! / ?

名詞抽出(MeCab)

欧州, 女王, 復活, 真央, ちゃん, 脅威

欧州女王復活 真央ちゃんにも脅威!?

形態素解析(MeCab)

欧州 / 女王 / 復活 / 真央 / ちゃん / に / も / 脅威 / ! / ?

名詞抽出(MeCab)

欧州, 女王, 復活, 真央, ちゃん, 脅威

変数付け

a　b　c　d　e　f

tf-idf

比較としてtf-idf法を用い特徴量を求める

- tf(Term Frequency):
  単語xの全レコード中の出現頻度

- df(Document Frequency):
  単語xが出現する日数の割合

- tf-idf = tf * idf

tf-idf

比較としてtf-idf法を用い特徴量を求める

- tf(Term Frequency):
  単語xの全レコード中の出現頻度

- df(Document Frequency): → 一般的な語
  単語xが出現する日数の割合

- tf-idf = tf * idf → 特徴的な語

語	頻度
位	1,319
五輪	610
日本	509
朝	481
青龍	447
戦	415
遼	391
人	362
年	348
くん	318

出現頻度

語	頻度
位	1,319
五輪	610
日本	509
朝	481
青龍	447
戦	415
遼	391
人	362
年	348
くん	318

語	出現割合（%）
ら	100.00
五輪	100.00
代表	100.00
位	100.00
出場	100.00
日本	100.00
男子	100.00
年	98.43
戦	98.43
日	98.43

出現頻度　　DF(出現割合)

語	頻度
位	1,319
五輪	610
日本	509
朝	481
青龍	447
戦	415
遼	391
人	362
年	348
くん	318

出現頻度

語	出現割合（%）
ら	100.00
五輪	100.00
代表	100.00
位	100.00
出場	100.00
日本	100.00
男子	100.00
年	98.43
戦	98.43
日	98.43

DF(出現割合)

語	tf-idf
貴乃花	0.0016
国母	0.0014
真央	0.0013
一門	0.0013
ヨナ	0.0013
娘	0.0013
藍	0.0012
安治川	0.0012
理事	0.0012
池田	0.0011

tf-idf

語	頻度
位	1,319
五輪	610
日本	509
朝	481
青龍	447
戦	415
遼	391
人	362
年	348
くん	318

出現頻度

語	出現割合（%）
ら	100.00
五輪	100.00
代表	100.00
位	100.00
出場	100.00
日本	100.00
男子	100.00
年	98.43
戦	98.43
日	98.43

DF(出現割合)

語	tf-idf
貴乃花	0.0016
国母	0.0014
真央	0.0013
一門	0.0013
ヨナ	0.0013
娘	0.0013
藍	0.0012
安治川	0.0012
理事	0.0012
池田	0.0011

tf-idf

ZDD演算による頻出パタン抽出

各日のDBからZDD演算を用いて頻出パタンを求める

- 語単位ではなく、大量のパタンの頻度を求められる
- 単純に高頻度のパタンを抽出するので一般語が多く含まれると考えられる

日	パタン
2010/1/11	{W杯}, {戦}, {位}
2010/1/12	{代表}, {五輪}, {位}
2010/1/13	{魁皇}, {勝}, {五輪}, {位}
2010/1/14	{首位}, {ウッズ}, {日本}, {位}
2010/1/15	{千代, 大海}, {引退}, {五輪}
2010/1/16	{連勝}, {五輪}, {位}
2010/1/17	{五輪}, {位}

ZDD演算による頻出パタン抽出結果(α=10)

日	パタン
2010/1/11	{W杯}, {戦}, {位}
2010/1/12	{代表}, {五輪}, {位}
2010/1/13	{魁皇}, {勝}, {五輪}, {位}
2010/1/14	{首位}, {ウッズ}, {日本}, {位}
2010/1/15	{千代, 大海}, {引退}, {五輪}
2010/1/16	{連勝}, {五輪}, {位}
2010/1/17	{五輪}, {位}

ZDD演算による頻出パタン抽出結果(α=10)

話題性の発見

2ヶ月間のうち、3日間だけ出現するパタンを抽出

話題性の発見

2ヶ月間のうち、3日間だけ出現するパタンを抽出

- その期間中話題だったパタンを抽出できる
- “3日間だけ”という制約上、一般語も除去できる

日	パタン
1/11～13	{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}
1/12～14	{アジア, 大会}
1/13～15	ø
1/14～16	{最多, 葛西, 連続, 大会, 代表}

3日間だけ出現するパタンの抽出結果例

実験結果(補足)

総パタン数	315,159
ZDDノード数	37,333

計算時間	0.0264 s

Intel Core2 Duo E8400 3GHz
4GB Memory
Ubuntu 8.04 LTS

結論

- 話題性を示すパタンを抽出できた
- 高速にパタンを抽出しコンパクトに出力できるZDDと親和性が高い
- 特にスポーツニュースでは、特定期間中の頻出パタンを取り出すことで流行を読み取れる可能性
- 今後は他のテキストデータ、形態素解析の改善など試したい