【VLDB2010勉強会】

No. 41: Data Mining, Copy Detection and Data Publication

担当: 鈴木 優(名古屋大学)

タイトル一覧

- Interesting-Phrase Mining for Ad-Hoc Text Analysis
  - 共起するテキストフレーズを大量の文書群から高速に発見
- Global Detection of Complex Copying Relationships Between Sources
  - 複数の情報源からどこからどこへコピーされたのかを発見
- Fragments and Loose Associations: Respecting Privacy in Data Publishing

Interesting-Phrase Mining for Ad-Hoc Text Analysis (Srikanta et al, mpi, Germany)

- Interesting-Phrase = 興味のあるフレーズを発見
  - “Steve Jobs” を入力 → “Mac OS X” “The Computer Maker”
  - クエリを入力すると，クエリに関連する単語を結果として返す
- Interestingness (I) の高い単語を抽出

$$\text{単語 p の Interestingness} = \frac{\text{単語 p とクエリ q が両方含まれている文書数}}{\text{単語 p が含まれている文書数}}$$

計算コストが高い

Pre-processing で Forward Indexing を構築

Interesting-Phrase Mining for Ad-Hoc Text Analysis (Srikanta et al, mpi, Germany)

- 高いコストとなる理由

- キーとなる考え方
  - 部分文字列をできるだけ考えない
    - abcde が一回出たら abcd も bcde も必ず1回出現する
    - abcd や bcde は数えずに，abcdeだけをカウントする
      - abeabef -> abca, abef, bcab, bef, cabe, ef, f

Interesting-Phrase Mining for Ad-Hoc Text Analysis (Srikanta et al, mpi, Germany)

- Forward Index
  - Inverted index の拡張

Inverted Index

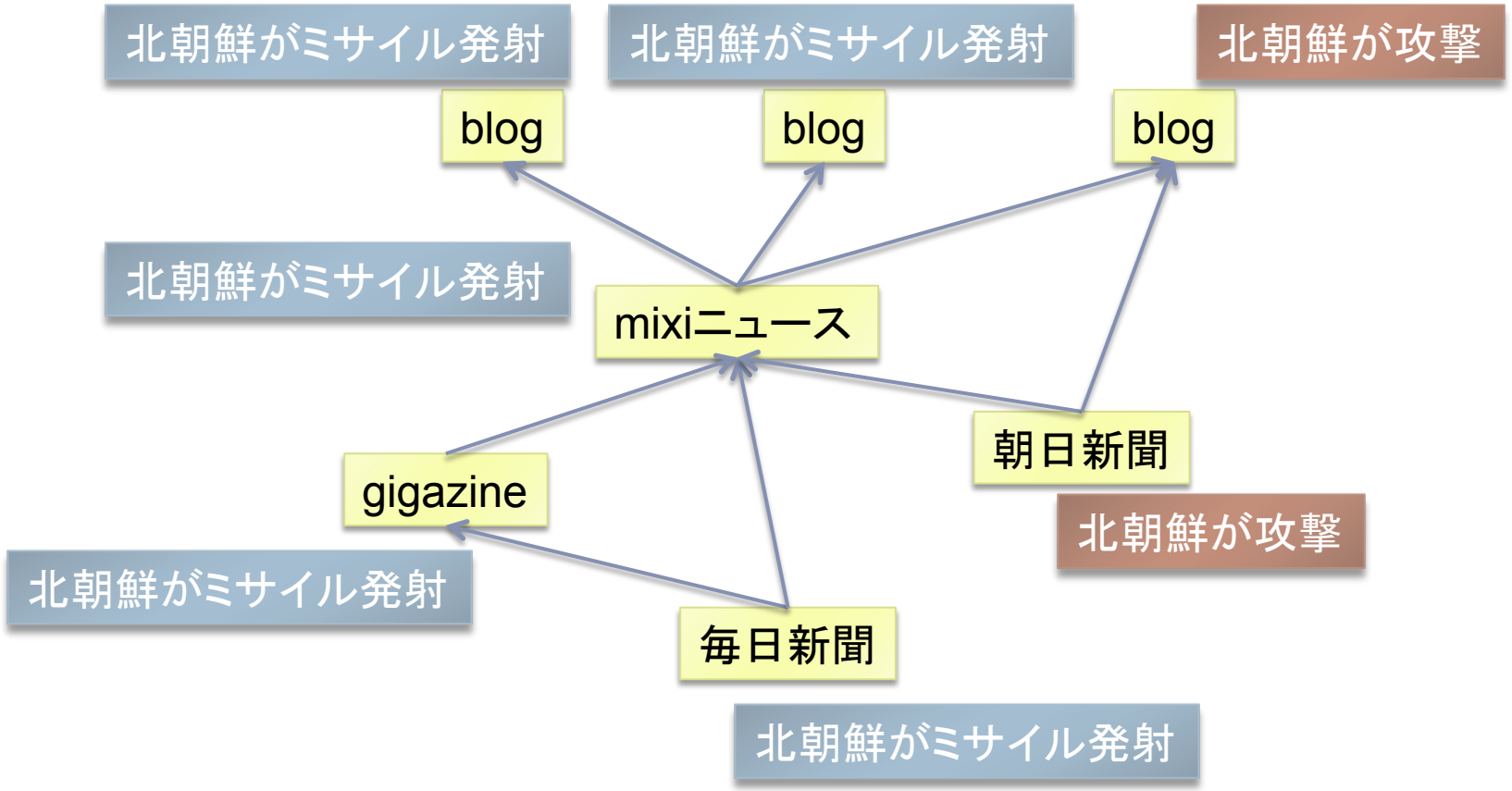
フレーズ	コンテンツ	I
p1(4)	{d2, d3, d9, d13}	1/4
p2(6)	{d3, d4, d5, d9, d12, d18}	5/6
p3(8)	{d1,d2,d4,d9,d12,d14,d17,d18,d20}	6/8
...	...	...

Forward Index

リスト	コンテンツ
Fd1	(9)p8, (10)p9, p10, (12)p13
Fd4	(4)p2, (5)p5, (6)p6,...
Fd5	(4)p2,p3, (5)p5, (6)p6,...

Global Detection of Complex Copying Relationships Between Sources (Xin et al. AT&T)

- データの複製の系統を明らかにする
  - どこのサイトがどのサイトの情報をコピーしたか?



Global Detection of Complex Copying Relationships Between Sources (Xin et al. AT&T)

- どのように特定するのか?
  - 完全性
    - 元の情報と先の情報にどれだけ差があるのか
  - Formatting style
    - 形式がどれだけ似ているのか
  - Accuracy
    - 元の情報がどれだけ正しいか? [6]参照
    - データの整合性を検証

- 大規模なデータで検証しても実用的な時間で実行できる
  - 天気データ，本のデータで検証

Fragments and Loose Associations: Respecting Privacy in Data Publishing

- 秘匿情報を公開する方法の提案
  - 既存の手法：k-anonymity, l-diversity....etc.
    - データの一部を隠すことによって join の候補を k 個以上にする
    - 対応関係が分からなくなるので，秘匿性が保たれる
  - 既存の手法の問題
    - visibility が不十分
      - “所属：??大学大?院?報?研?科” などという情報は不要
- この研究での方針
  - データを？で隠したりしない
  - データを属性ごとに分割し，一部だけを見せる

k-anonymity の問題設定を変えただけ???

Fragments and Loose Associations: Respecting Privacy in Data Publishing

SSN	患者の名前	誕生日	郵便番号	病名	医者の名前
123-456-7	鈴木	1976/1/3	123-4556	風邪	田中
567-138-2	佐藤	1992/3/1	765-4321	結核	吉田
728-283-4	加藤	2001/8/8	222-1243	鼻炎	中本

誕生日	郵便番号	G
1976/1/3	123-4556	bz1
2001/8/8	222-1243	bz2
1992/3/1	765-4321	bz1

G1	G2
bz1	id2
bz1	id1
bz2	id1

病気	医者名	G
鼻炎	中本	id1
結核	吉田	id2
風邪	田中	id2

Fragments and Loose Associations: Respecting Privacy in Data Publishing

- (kl,kr)−grouping
  - bz1がいくつのカラムに振られているか(kl)
  - id1がいくつのカラムに振られているか(kr)
- k−looseness
  - 何通りの組み合わせが考えられるか
  - k−anonymity の k と同じ

G1	G2
bz1	id2
bz1	id1
bz2	id1

- この論文はモデルの提案
  - k, kl, kr をいくつにすると，どれくらいの確率で正しい表が復元されるかについては，APPENDIX で議論されている