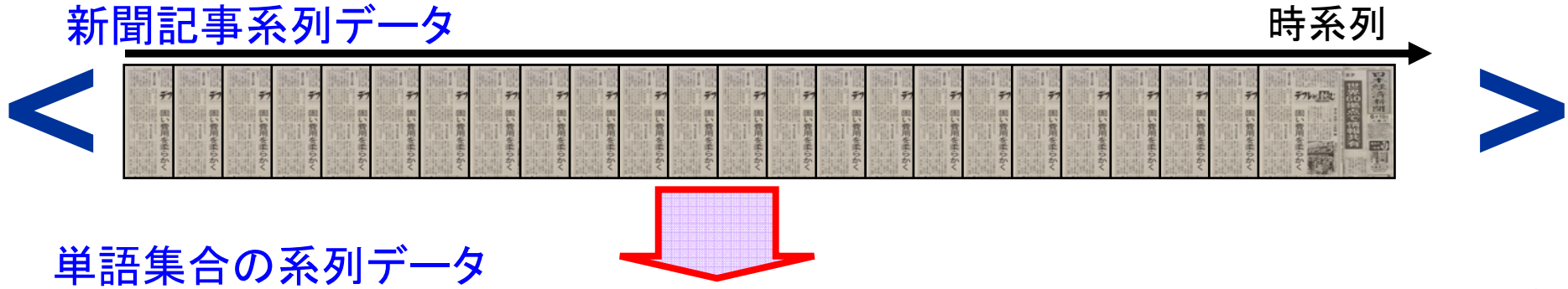
No.81 候補系列抑制による高速系列データマイニング

水上 紘悠*, 岩沼 宏治**, 鍋島英知**

研究背景：系列データマイニング

新聞記事系列データ　　時系列



単語集合の系列データ

〈(台風　送検…) (洪水　初公判…) (判決　大雪…)…(台風…) (洪水…)…〉

〈(台風) (洪水)〉 **有用頻出パターン**として抽出

研究目的

パターンの抽出をアプリオリアルゴリズムに基づき行うと、段階的に**頻出な系列**を抽出する事ができる.
しかし，膨大な数の**候補系列**（頻出の可能性を持つ系列）を生成してしまうという問題がある.

候補系列数を抑制する事で高速化を図る

* 山梨大学大学院医学工学総合教育部修士課程コンピュータ・メディア工学専攻
** 山梨大学大学院医学工学総合研究部

候補系列の抑制

主な手段として，ハッシュを用いて候補系列の抑制を行う.

概要:
step1. 頻出系列L_kを求めると同時に，長さk+1の系列を枚挙し，ハッシュテーブルに格納する
step2. 長さk+1の候補系列生成時にテーブルの値を元に生成の判定を行う

実験結果

系列長	抑制無し	table_size				頻出系列
		256	4096	16384	262144	
2	38416	38416	13166	2234	565	483
3	10781	10781	2984	696	334	310
4	1810	1792	126	48	42	42
5	74	9	3	3	3	3
6	3	2	2	2	2	2
time [s]	4.14	3.83	2.29	1.57	1.42	

図:候補系列数と実行時間

実験パラメータ
系列長5000, アイテム種類数1127, 1日1アイテム,
最小サポート0.1%

十分なサイズのハッシュテーブルを用意することで,
頻出系列数と同等な値まで候補系列数を抑える事ができる.
また，候補系列の抑制に伴い，実行時間に関しても向上がみてとれる.

文献 Jong Soo Park,Ming-Syan Chen and Philip S. Yu:
An Effective Hash-Based Algorithm for Mining Association Rules.
In *Proc*. ACM SIGMOD *Int. Conf. Management of Data*,pp.175-186,1995.