

書誌情報抽出および統合のための テキストマイニング

2015年11月13日

インタラクティブ情報アクセスと可視化マイニング研究会 第11回研究会

国立情報学研究所

高須 淳宏

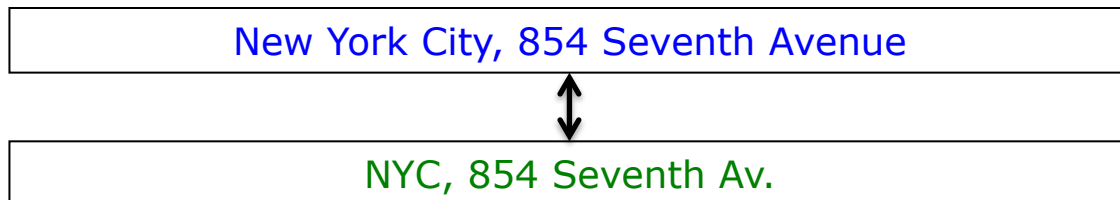
共同研究： 太田学(岡山大学)、正田備也(長崎大学)

発表の概要

- **学術情報抽出**
 - 参考文献文字列の解析と抽出
 - タイトルページの解析と抽出
- **書誌マッチング**
 - 木マッチングを用いた書誌統合
 - 著者名の曖昧性解消
- **情報分析**
 - 論文のキーワードの推薦
 - 学術研究のトピック分析

書誌情報統合(1/2)

要素レベルのマッチング



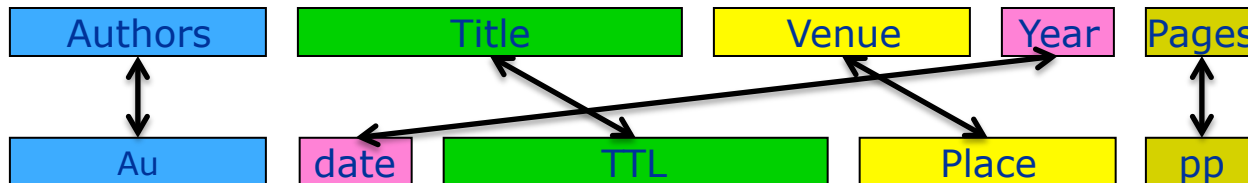
近似文字列マッチング

レコードレベルのマッチング (e.g., Duplicate Detection)



近似文字列マッチング
+
分類

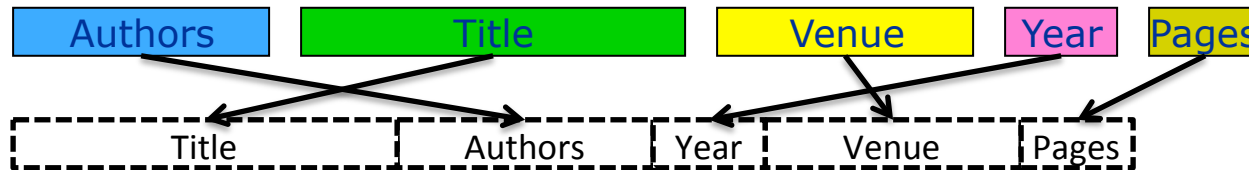
スキーマレベルのマッチング (e.g., DB Integration)



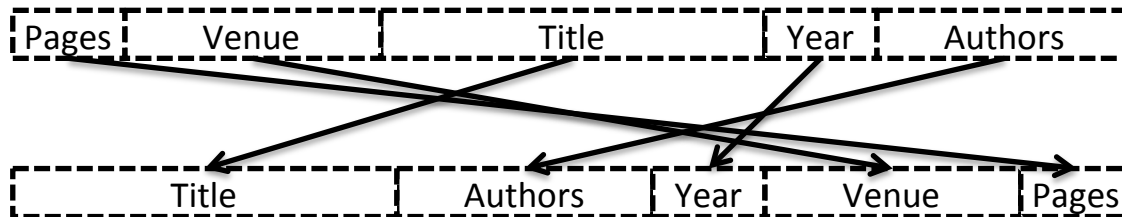
近似文字列マッチング
+
構造マッチング

書誌情報統合 (2/2)

レコードと半構造データのマッチング



半構造データ同士のマッチング



字句・構文解析
＋
近似文字列マッチング
＋
構造マッチング

参考文献書誌情報抽出

- 参考文献文字列：有用な書誌情報の集積
 - 雑誌ごとに書式が異なる

□ 電子情報通信学会論文誌 (IEICE)

W. Diffie and M. Hellman, "New Directions in Cryptography,"
IEEE Trans. Inf. Theory, vol.22, no.6, pp.644-654, 1976.

□ 情報処理学会論文誌 (IPSJ)

Diffie, W. and Hellman, M.: New Directions in Cryptography,
IEEE Trans. Inf. Theory, Vol.22, No.6, pp.644-654 (1976).

- 機械学習 (CRF) による参考文献書誌情報の自動抽出
 - 書式ごとに大量のラベル付き参考文献文字列が必要
- 少量学習データによる参考文献書誌情報抽出

2段階参考文献書誌情報抽出

Cheswick, W.R. and Bellovin, S.M.: Firewalls and Internet Security, Addison-Wesley (1995).

書誌要素ラベル一覧 左列

- | | |
|----------------|------------|
| ▶ Author | ▶ Volume |
| ▶ Editor | ▶ Number |
| ▶ Translator | ▶ Page |
| ▶ Other author | ▶ Location |
| ▶ Book title | ▶ URL |
| ▶ Title | ▶ Day |
| ▶ Conference | ▶ Month |
| ▶ Journal | ▶ Year |
| ▶ Publisher | ▶ Other |

<Author> Cheswick, W.R.
<Delimiter> and
<Author> Bellovin, S.M.
<Delimiter> :
<Delimiter>
<Book title> Firewalls and Internet Security
<Delimiter> ,
<Publisher> Addison-Wesley
<Delimiter>
<Delimiter> (
<Year> 1995
<Delimiter>)
<Delimiter> .

条件付き確率場

Conditional Random Field

入力トークン系列: $\mathbf{x} = x_1, \dots, x_n$

出力ラベル系列: $\mathbf{y} = y_1, \dots, y_n$

$$P(\mathbf{y} | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i=1}^n \sum_k \lambda_k f_k(y_{i-1}, y_i, \mathbf{x})\right)$$

Diagram annotations:

- λ_k : f_k の重み (Weight of f_k)
- f_k : 素性関数 (Feature function)
- $Z(\mathbf{x})$: 正規化項 (Normalization term)

$$\mathbf{y}^* = \operatorname{argmax}_{\mathbf{y} \in Y(\mathbf{x})} P(\mathbf{y} | \mathbf{x})$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y}' \in Y(\mathbf{x})} \exp\left(\sum_{i=1}^n \sum_k \lambda_k f_k(y'_{i-1}, y'_i, \mathbf{x})\right)$$

- feature

- Unigram (47)

- トークン位置, 文字種, 文字数, キーワードの有無, 辞書的素性など

- Bigram (1)

- ラベル遷移

人名, 論文誌名, 会議名, 出版社名, 地名, 月名

学習データの削減

- 能動サンプリング
 - 書誌情報抽出が困難な参考文献文字列が学習に有効
 - 確信度の低い参考文献文字列を優先的に学習
 - 3種類の確信度を評価
- 擬似学習データ
 - 正解書誌要素ラベル付き参考文献文字列の書誌要素を書誌情報DB中のエンティティと無作為に置換（自動生成）
- 転移学習
 - 他雑誌用書誌情報抽出器による書誌要素推定
 - 推定ラベルを対象雑誌の書誌情報抽出器の素性に追加

川上、太田、高須、安達 “少量学習データによる参考文献書誌情報抽出精度の向上,”
情報処理学会論文誌 データベース, Vol. 8, No.2, pp.18 – 29, 2015

書誌情報抽出の確信度

- NLH: Normalized Likelihood

- 入力系列の条件付き確率が最大のラベル系列に注目

$$C_{NLH}(\mathbf{x}) = \frac{\log(P(\mathbf{y}^*|\mathbf{x}))}{|\mathbf{x}|}$$

CRFの出力する
条件付き確率

$\mathbf{x}$: 入力トークン系列
 $\mathbf{y}^*$: 出力ラベル系列

- MP: Minimum Probability of Token Assignment

- 各トークンに付与された確率が最大のラベルに注目

$$C_{MP}(\mathbf{x}) = \min_{i \leq |\mathbf{x}|} \max_{l \in L} P(Y_i = l)$$

Y_i : i 番目のトークンに付与された
ラベルの確率変数
 L : ラベル集合

- ATE: Average Token Entropy

- 各トークンの全ラベル候補の確率に基づくエントロピー

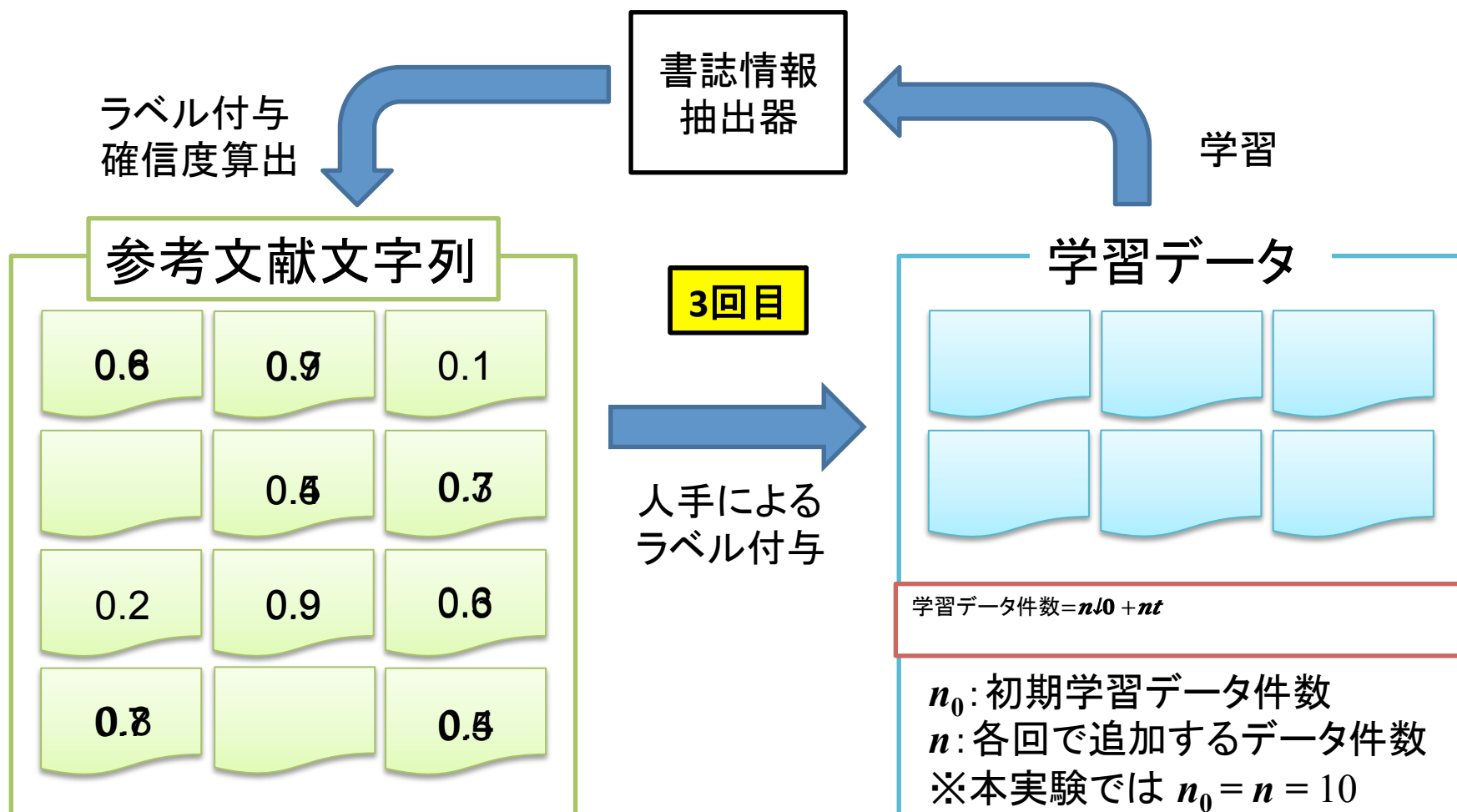
$$C_{ATE}(\mathbf{x}) = - \frac{\max_{x_i \in \mathbf{x}} \sum_{l \in L} -P(Y_i = l) \log P(Y_i = l)}{|\mathbf{x}|}$$

書誌ラベル l が i 番目のトークン
に割り当てられる周辺確率

i 番目のトークンに対する
ラベル付けのエントロピー

能動サンプリング

Active Sampling



擬似学習データ

Pseudo Training Data

- 擬似学習データの生成方法

オリジナル

S.U. Pillai: Array Signal Processing, Springer-Verlag, New York (1989).

S.U. Pillai: Array Signal Processing, Springer-Verlag, New York (1989).

<Author>

<Book title>

<Publisher>

<Location><Year>

書誌情報データベース

擬似学習
データ

H. Ogawa: Distance in Graphs, Academic Press, London (2003).

m

学習データ件数 $= n_0 + n(1+m)t$

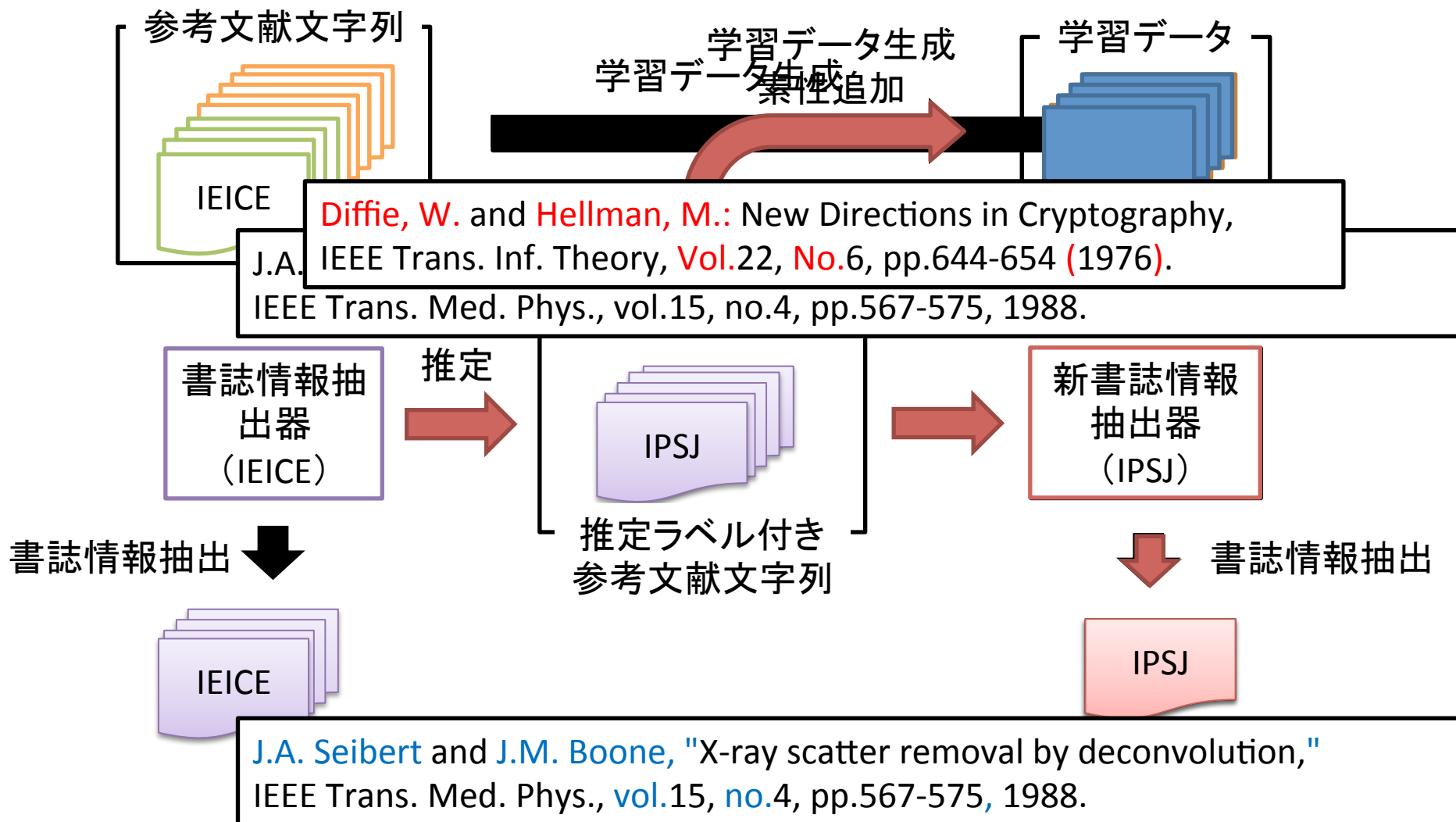
n_0 : 初期学習データ件数

n : 能動サンプリングの各回で追加するデータ件数

m : 参考文献文字列一件につき生成する擬似学習データ件数

転移学習

Transfer Learning



参考文献書誌情報抽出実験

- 実験データ：国内学術論文誌

雑誌	言語	参考文献文字列数(英/日)
IEICE-J	英語, 日本語	4,787(2,594/2,193)
IEICE-E	英語	4,497(4,497/0)
IPSJ	英語, 日本語	4,574(3,037/1,537)

- 人手でトークン列に変換

- 評価指標：書誌情報抽出精度

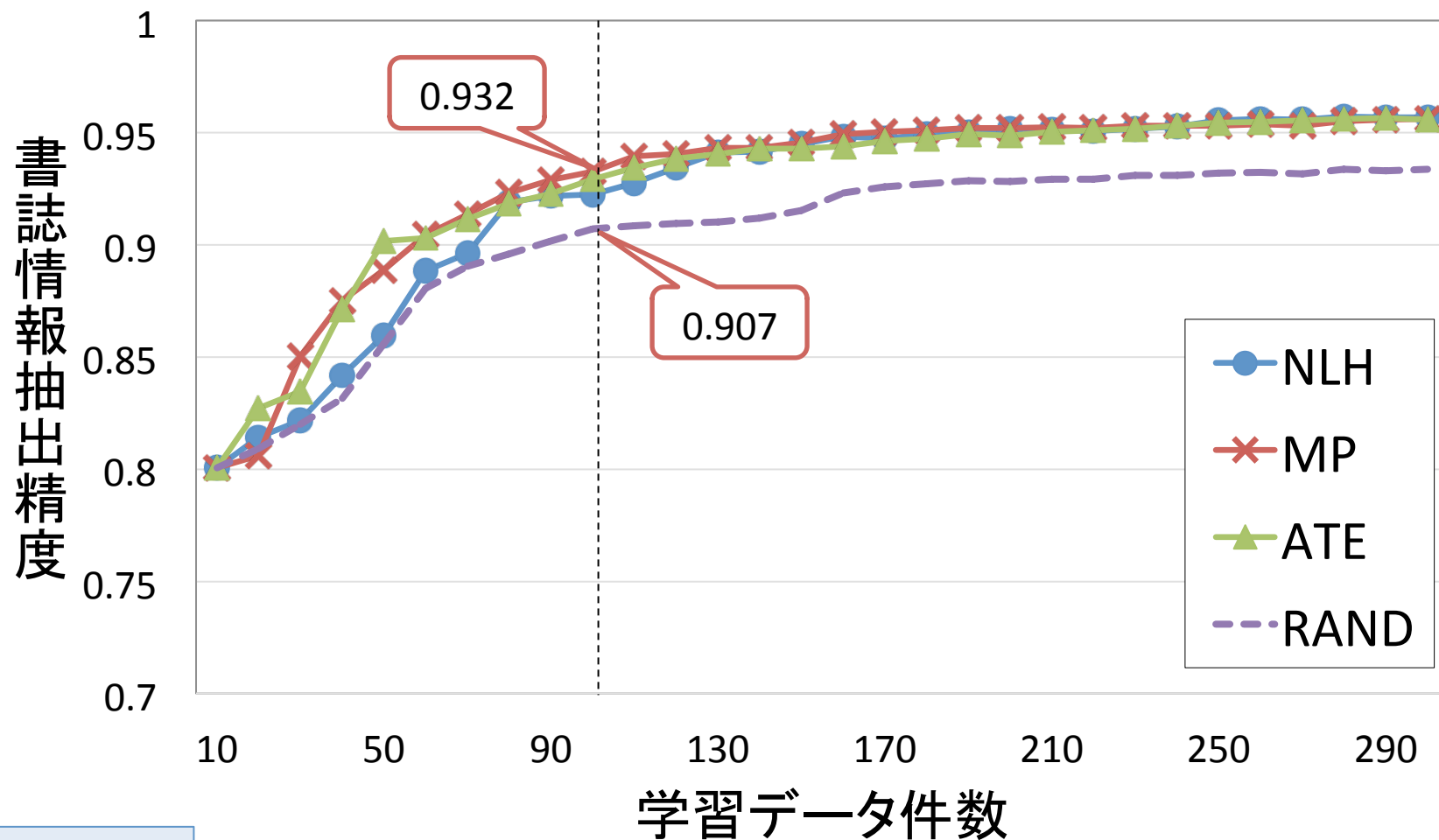
$$\frac{\text{全ての書誌情報を過不足なく抽出できた参考文献文字列数}}{\text{全参考文献文字列数}}$$

- RAND：ランダムサンプリング

- CRF++0.58* *<http://crfpp.googlecode.com/svn/trunk/doc/index.html>

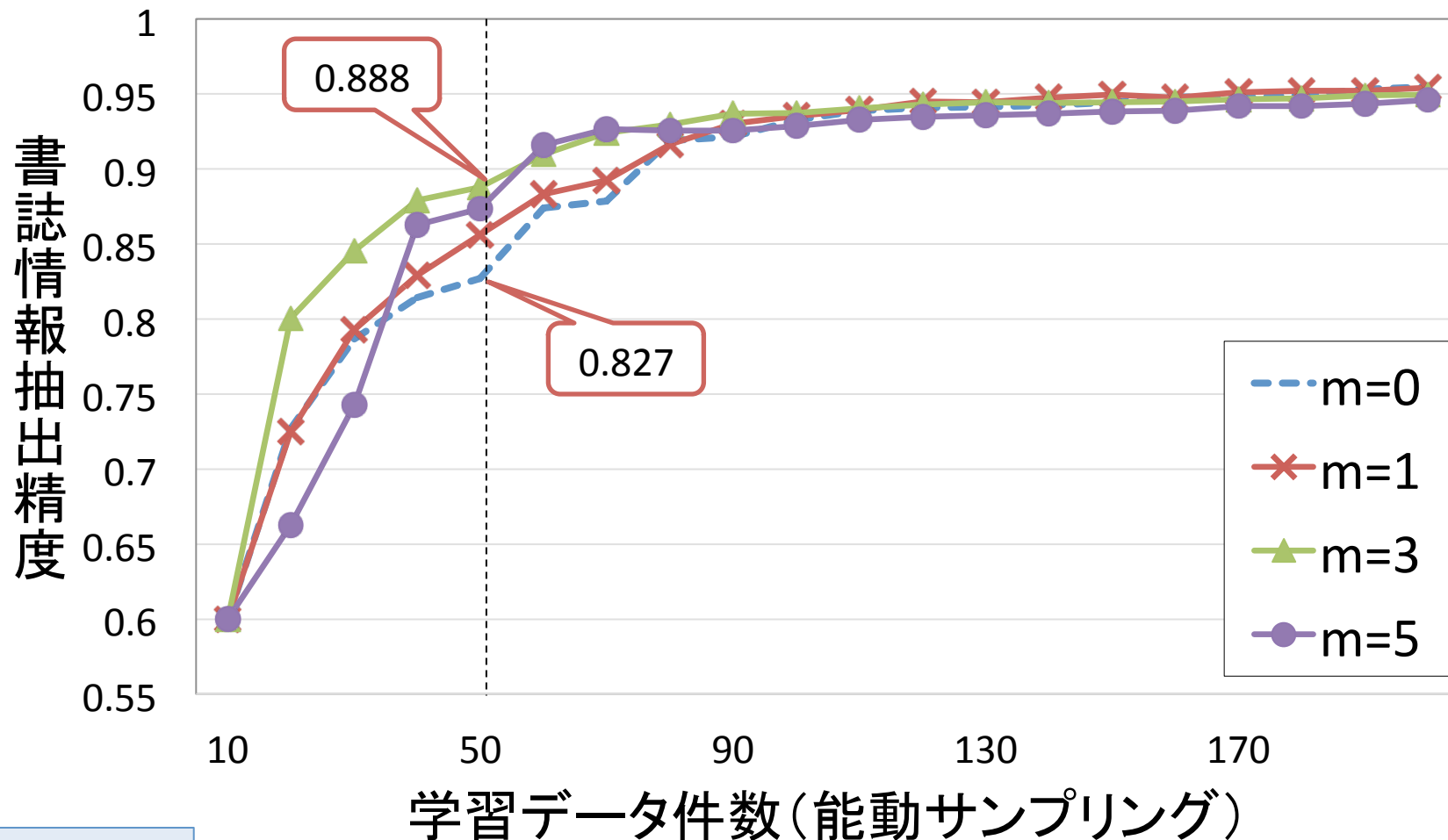
- 5分割交差検定

能動サンプリングの効果



IEICE-J

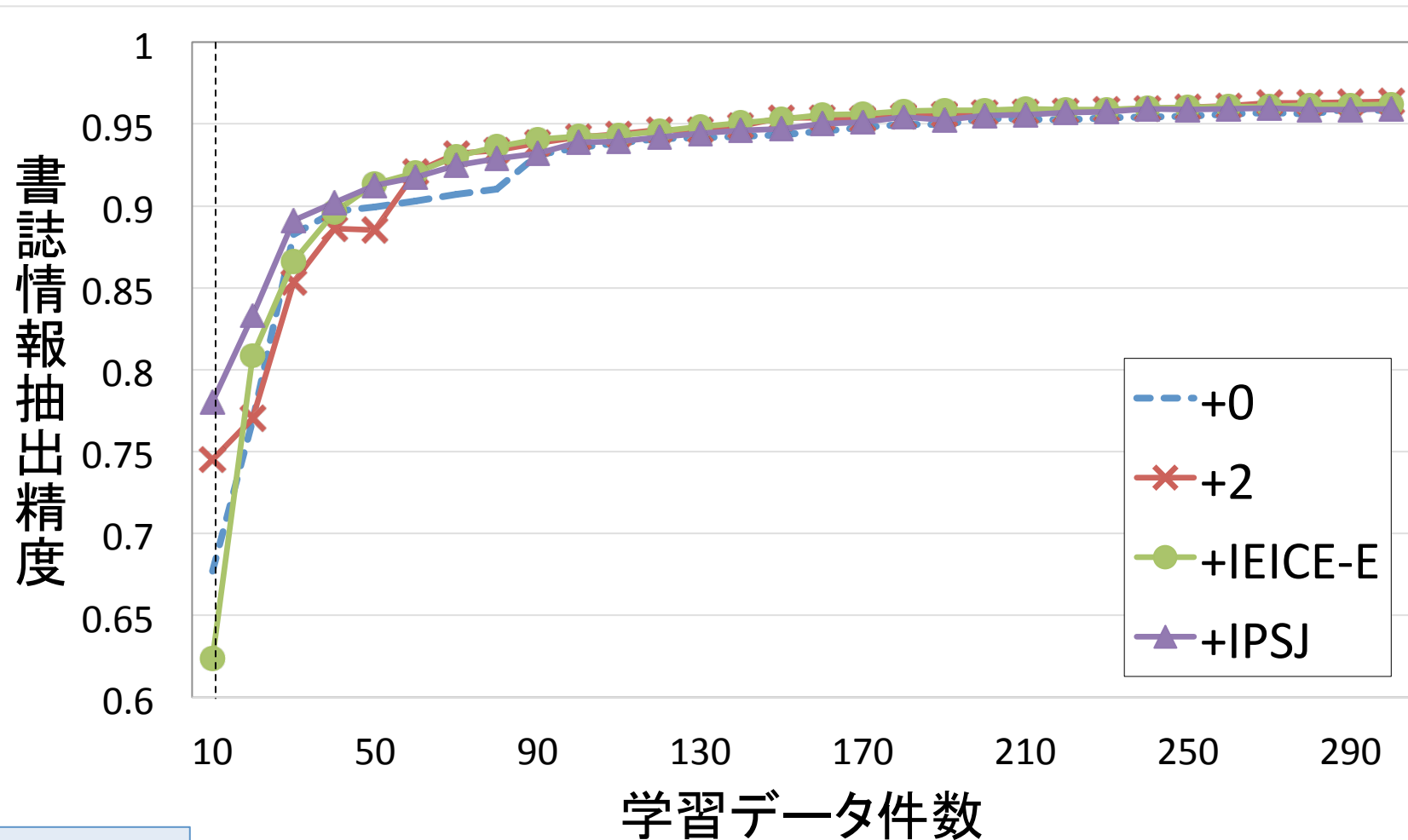
擬似学習データの効果（確信度 = NLH）



IEICE-J

$$\text{学習データ件数} = n_0 + n(1+m)t$$

転移学習の効果（確信度 = NLH）



IEICE-J

転移素性の比較

- 少量（初期）データ（10件）学習時の書誌情報抽出精度

		追加素性の雑誌			
		IEICE-J	IEICE-E	IPSJ	+2
抽出 対象 雑誌	IEICE-J	0.6770	0.6238	0.7802	0.7451
	IEICE-E	0.7594	0.7432	0.7581	0.7721
	IPSJ	0.5466	0.5354	0.5313	0.5490

- 全データ（4,000件弱）学習時の他雑誌用抽出器による書誌情報抽出精度

		学習データ		
		IEICE-J	IEICE-E	IPSJ
テスト データ	IEICE-J	0.9637	0.5567	0.8901
	IEICE-E	0.9457	0.9706	0.9199
	IPSJ	0.8266	0.7298	0.9652

参考文献書誌情報抽出のまとめ

- **参考文献書誌情報抽出における学習データの削減**
 - CRFの確信度を利用した能動サンプリング
 - 擬似学習データの利用
 - 転移学習
- **削減効果**
 - 能動サンプリング：学習データを百件程度にまで削減可能
 - 擬似学習データ：有効、ただし雑誌により効果に差がある
 - 転移学習：書式等が類似した雑誌の推定結果が有効
- **今後の主な課題**
 - 多様な雑誌論文からの書誌情報抽出
 - 書式の異なる参考文献文字列の検出

発表の概要

- **学術情報抽出**
 - 参考文献文字列の解析と抽出
 - タイトルページの解析と抽出
- **書誌マッチング**
 - 木マッチングを用いた書誌統合
 - 著者名の曖昧性解消
- **情報分析**
 - 論文のキーワードの推薦
 - 学術研究のトピック分析

学術論文OCRテキスト

```
<?xml version="1.0" encoding="euc-jp" ?>
<paper>
  <page number="1">
    <block x="666" y="479" width="1814" height="87" sec="">
      <line x="666" y="479" width="1814" height="87">
        <word x="666" y="479" width="1814" height="87">
          <char x=" 666" y=" 486" width="80" height="80">学</char>
          <char x=" 756" y=" 484" width="84" height="82">術</char>
          <char x=" 863" y=" 493" width="52" height="64">論</char>
          <char x=" 944" y=" 493" width="70" height="65">文</char>
          :
          <char x="2325" y=" 514" width="61" height="41">抽</char>
          <char x="2416" y=" 489" width="61" height="62">出</char>
        </word>
      </line>
    </block>
    :
  </page>
</paper>
```

レイアウトタグ

各要素の位置や大きさ

テキスト

OCRテキストへの書誌要素ラベリング

Bibliography Labeling for OCR Text

```
<block>
  <j-title>
    <line> 学術論文文書画像からの自動書誌要素抽出 </line>
  </j-title>
</block>

<block>
  <j-authors>
    <line> 井上諒平†太田学††高須淳宏††† </line>
  </j-authors>
</block>
```

※各line(行)に対して書誌ラベル付けを行う

素性テンプレート (1/2)

レイアウト素性

種類	素性	内容
unigram	<i(0)>	lineの識別番号
	<x(0)>	lineのX座標
	<y(0)>	lineのY座標
	<w(0)>	lineの幅
	<h(0)>	lineの高さ
	<g(0)>	前のlineとの間隔
	<cw(0)>	line内の文字の幅の中央値
	<ch(0)>	line内の文字の高さの中央値
	<#c(0)>	line内の文字数

素性テンプレート (2/2)

言語的素性

種類	素性	内容
unigram	<ec(0)>	line内の英数字の割合
	<kc(0)>	line内の漢字の割合
	<jc(0)>	line内の平仮名・片仮名の割合
	<s(0)>	line内の記号の割合
	<kw(0)>	line内の特徴的な語の有無

Bigram素性

種類	素性	内容
bigram	<y(-1),y(0)>	ラベルの遷移

タイトルページからの書誌情報抽出実験

- CRF++ 0.53
 - <http://crfpp.sourceforge.net/>
- OCR認識率99% (アブストラクトの部分)
- 評価データ
 - 情報処理学会論文誌(IPSJ)
2003年 479件
 - 電子情報通信学会論文誌-英文(IEICE-E)
2003年 473件
 - 電子情報通信学会論文誌-和文(IEICE-J)
2003-04年 174件
- 5分割交差検定

誤り検出による後処理コストの削減

書誌推定を誤った論文件数(誤推定率)

論文誌	学習データ20件	学習データ100件	学習データ300件
IPSJ	15.88件(16.58%)	7.81件(8.15%)	5.97件(6.23%)
IEICE-E	28.83件(30.48%)	9.76件(19.32%)	3.92件(4.14%)
IEICE-J	11.95件(34.34%)	7.04件(20.23%)	—

人手による後処理で誤りを1%未満にするために確認すべき論文数
(全論文に占める割合)

論文誌	学習データ20件	学習データ100件	学習データ300件
IPSJ	43件(44.9%)	17件(17.7%)	10件(10.4%)
IEICE-E	76件(80.3%)	49件(51.8%)	10件(10.6%)
IEICE-J	34件(97.7%)	18件(51.7%)	—

- 誤りを1%未満にするための低確信度の論文の後処理

→ 全体の**10%程度**を処理すれば良い (IPSJ, 学習データ300件, NLH)

発表の概要

- **学術情報抽出**
 - 参考文献文字列の解析と抽出
 - タイトルページの解析と抽出
- **書誌マッチング**
 - 木マッチングを用いた書誌統合
 - 著者名の曖昧性解消
- **情報分析**
 - 論文のキーワードの推薦
 - 学術研究のトピック分析

スキーマの異なる書誌情報の統合

- **順序木の編集距離**

- 相対的な計算効率が良い
- フィールドの記述順序が異なるデータの統合 ×

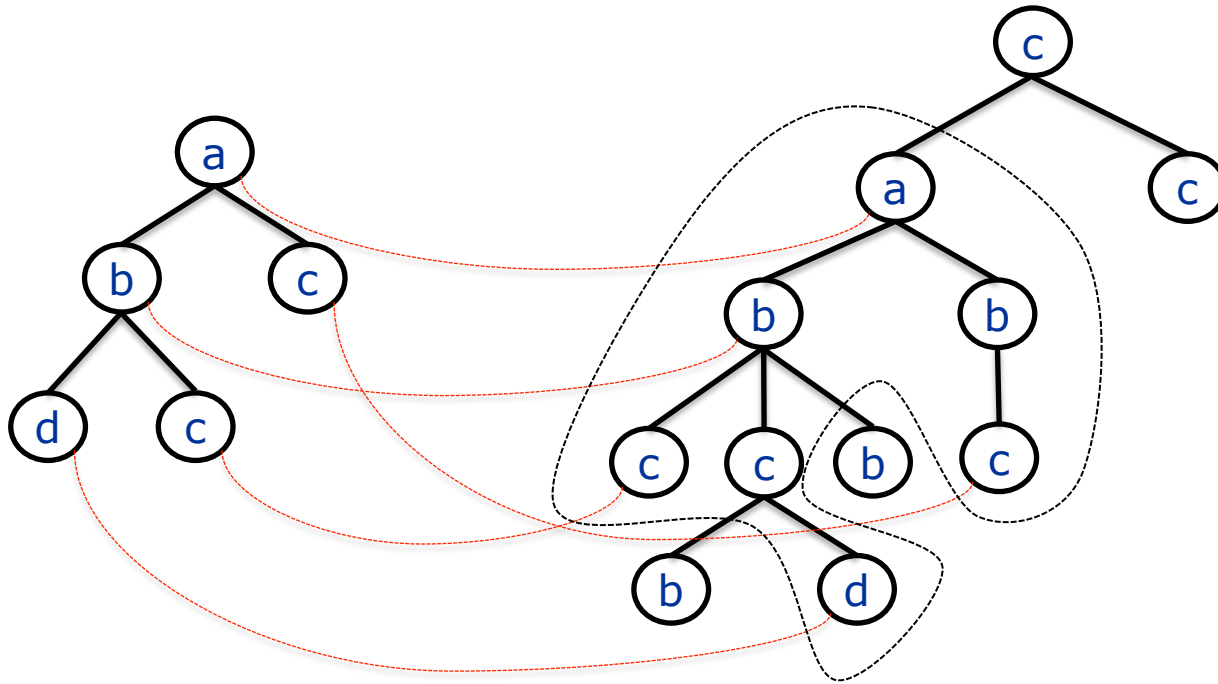
- **無順序木の編集距離**

- フィールドの順序が異なるデータ ○
- 大きなデータでは現実的な計算が困難

- **無順序木の包含**

- 効率的な計算可
- フィールドの順序に非依存
- 一方のデータが他に含まれる必要がある

木包含(tree inclusion) 問題



Pagazoglou, M. P.

異なる標記の
マッチングが必要

Mike P. Pagazoglou

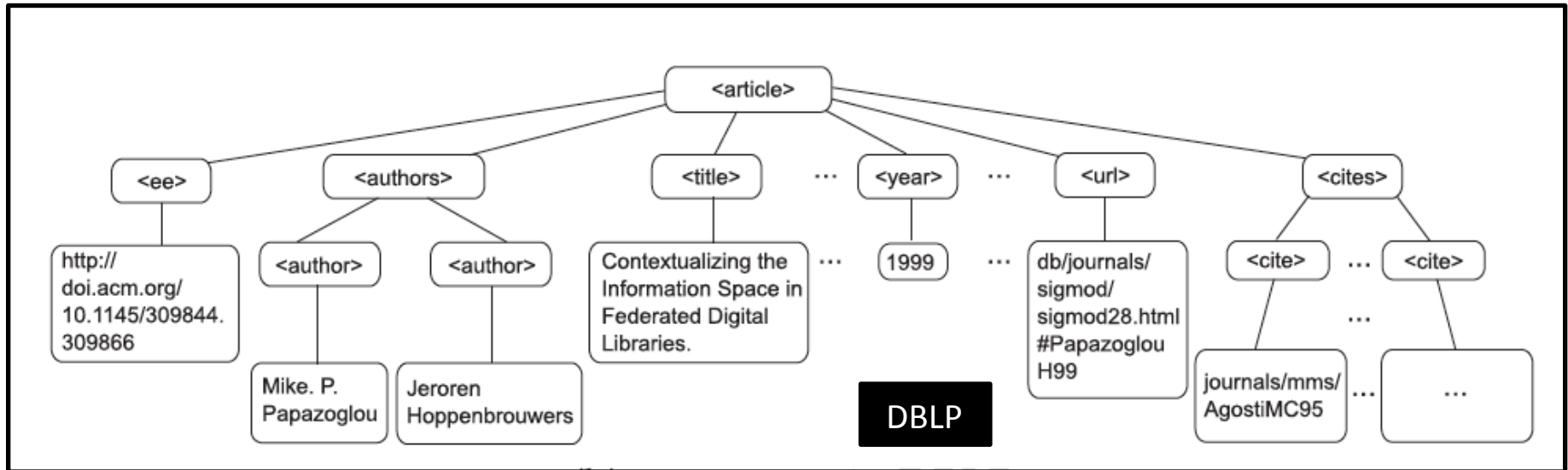
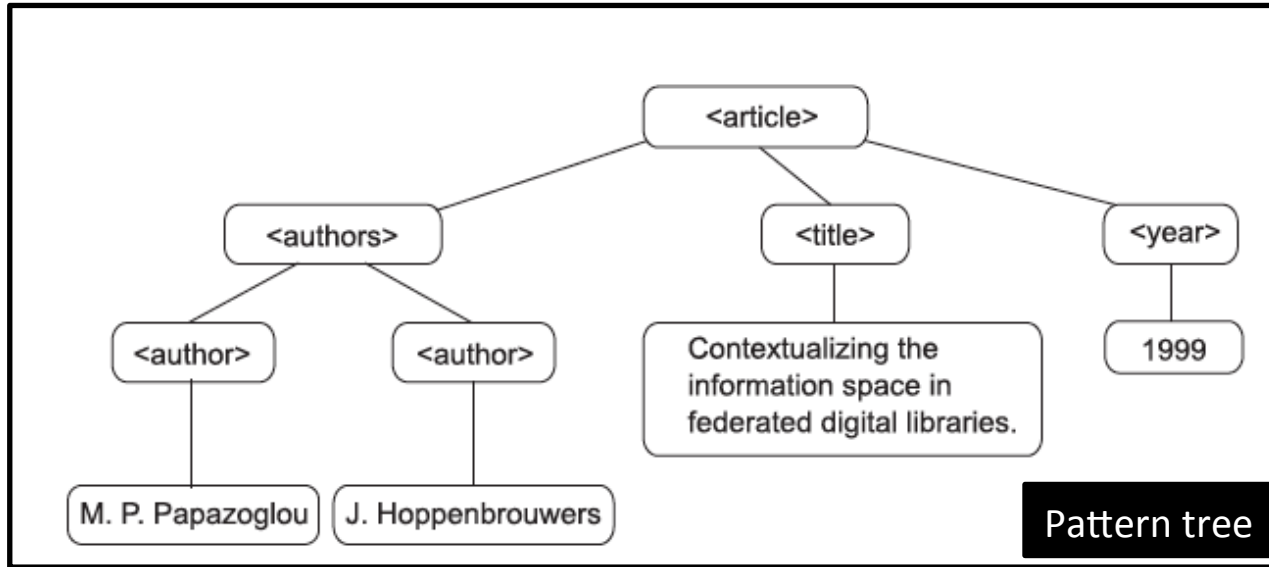
構造マッチング



構造＋文字列
マッチング

Mori, Takasu, Jansson, Hwang, Tamura, Akutsu,
"Similar Subtree Search using Extended Tree Inclusion,"
IEEE Trans. Knowledge and Data Engineering, Vol. 27, No.12, pp.3360 – 3373, 2015

無順序木包含によるタグマッチング



タグマッピング精度

• 評価データ

- ライプチヒ大学
Entity Resolution
Benchmark Datasets

• 手順

- Jaccard係数に基づき
similarity join
- Joinの結果得られたレ
コード対を木包含
マッチング
- ノードの対応よりtag
の対応付け

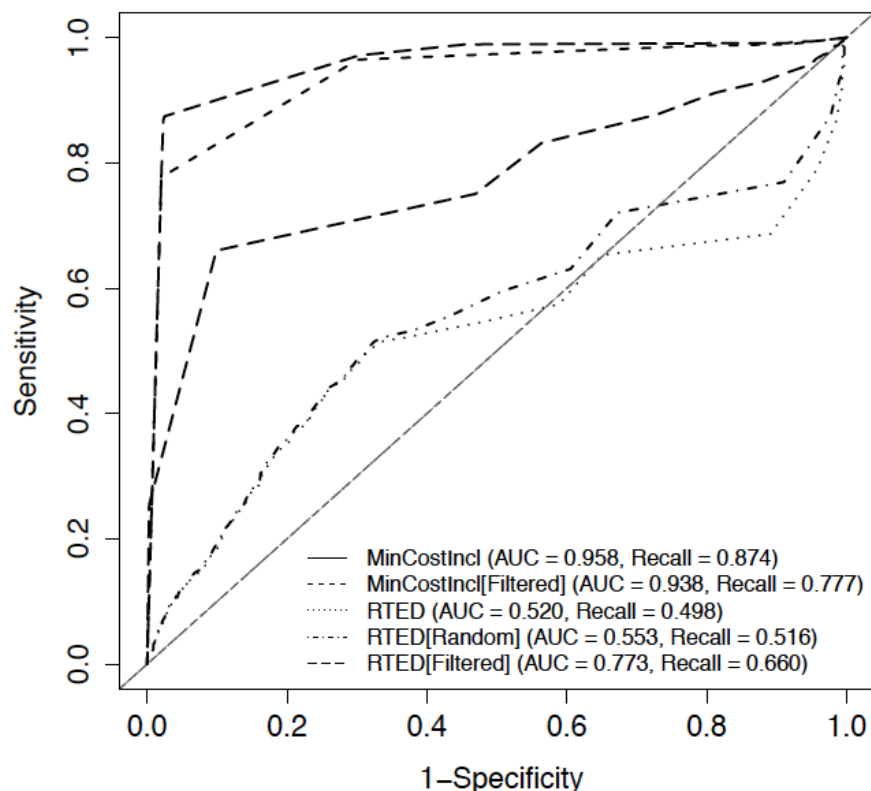
	ACM-DBLP	Scholar-DBLP
pattern articles	2,616	2,616
DBLP articles	2,294	64,263
common articles	2,224	5,347
blocking result	4,000	20,000

(a) ACM-DBLP					
	0 (article)	1 (authors)	2 (author)	3 (title)	4 (year)
article	3893	0	0	0	0
authors	0	3672	0	0	0
author	0	0	7962	0	0
title	0	0	0	3378	0
year	0	0	0	0	2987
url	0	10	0	283	529
cites	0	95	0	44	338
ee	0	71	0	148	102
cite	0	0	199	0	0
others	0	45	126	40	37
(b) Scholar-DBLP					
	0	1	2	3	4
article	18475	0	0	0	0
authors	0	13519	0	0	0
author	0	0	27217	0	0
title	0	0	0	15680	0
year	0	0	0	0	5082
url	0	0	0	1167	2772
cite	0	0	5768	0	0
editors	0	1798	0	0	0
editor	0	0	3365	0	0
others	0	3158	0	1261	2334

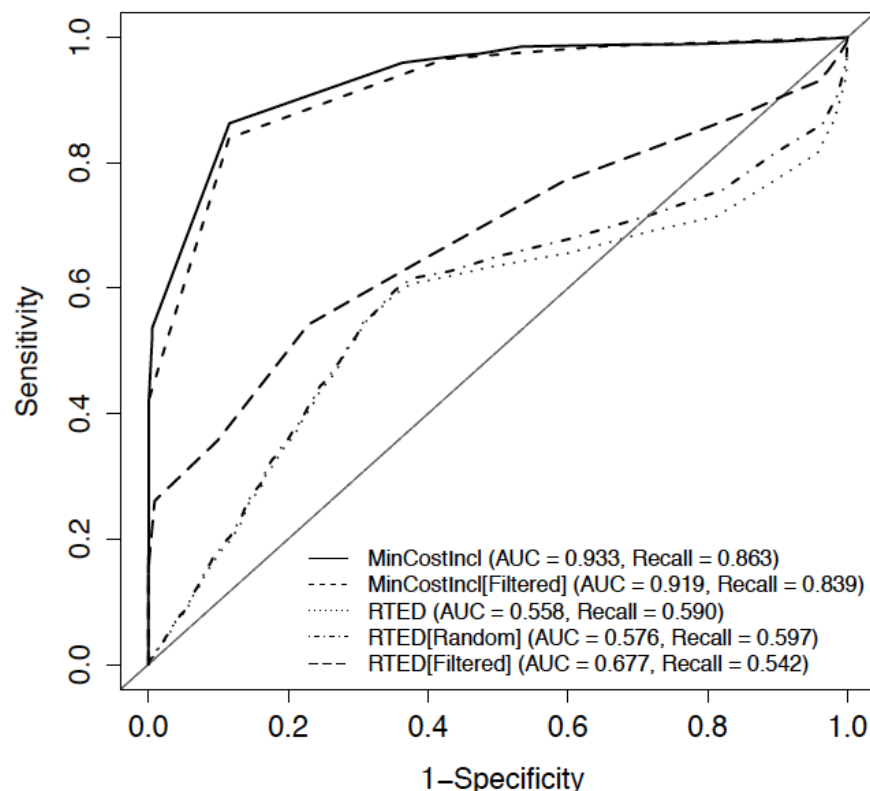
マッチング精度

順序木のマッチングアルゴリズム(木編集距離)と比較

ACM DBLP



Scholar DBLP



発表の概要

- **学術情報抽出**
 - 参考文献文字列の解析と抽出
 - タイトルページの解析と抽出
- **書誌マッチング**
 - 木マッチングを用いた書誌統合
 - 著者名の曖昧性解消
- **情報分析**
 - 論文のキーワードの推薦
 - 学術研究のトピック分析

著者名の曖昧性解消



dblp
computer science bibliography

home | browse | search | about

search dblp

[+] A. Suzuki

> Home > Persons

by type Trier 1

[-] Journal Articles

Asato Suzuki
運動生理学

refine list

2013

[j3] Motoshi Masuda, Takahiro Emoto, **A. Suzuki**, Masatake Akutagawa, Tomoki Kitawaki, Kazuyoshi Kitaoka, Hiroyuki Tanaka, Shigeru Obara, Kazuo Yoshizaki, Shinsuke Konaka, Yohsuke Kinouchi:
Evaluation of blood flow velocity wa
carotid artery using multi-branched
of human arteries. Biomed. Signal Pro... 509-519 (2013)

Atsuyuki Suzuki
計算機科学

2004

[j2] K. Shima, M. Todoriki, **A. Suzuki**:
SVM-based feature selection of latent s
Pattern Recognition Letters 25(9): 1051-1057 (2004)

Akihiro Suzuki
原子力工学

2001

[j1] N. Shaaban, S. Hasegawa, **A. Suzuki**, H. Takahashi:
The Use of Genetic Algorithms for the Improvement of Energy
Characteristics of CdZnTe Semiconductor Detectors. Genetic Programming and Evolvable Machines 2(3): 289-299 (2001)

クラスタリング

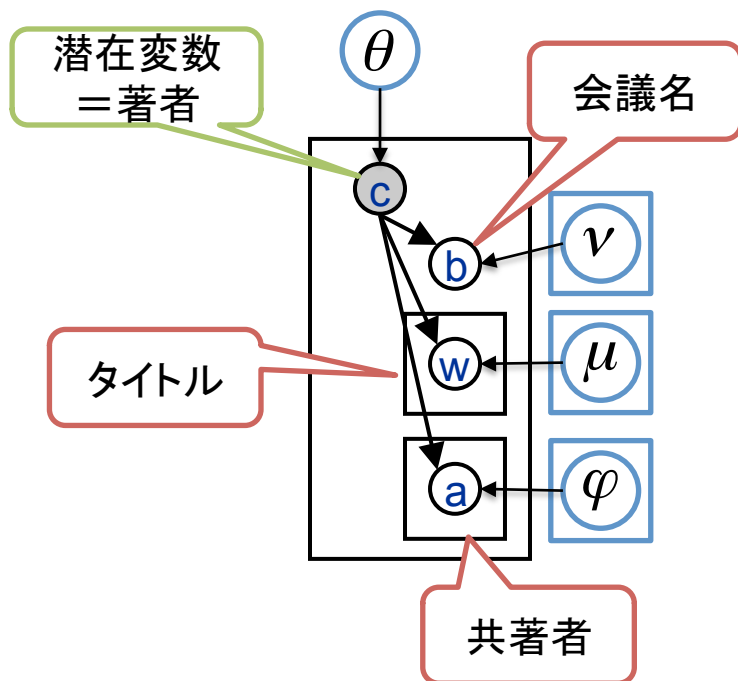
分類

Authors	Title	Venue	Year	Pages
Authors	Title	Venue	Year	Pages
Authors	Title	Venue	Year	Pages
Authors	Title	Venue	Year	Pages

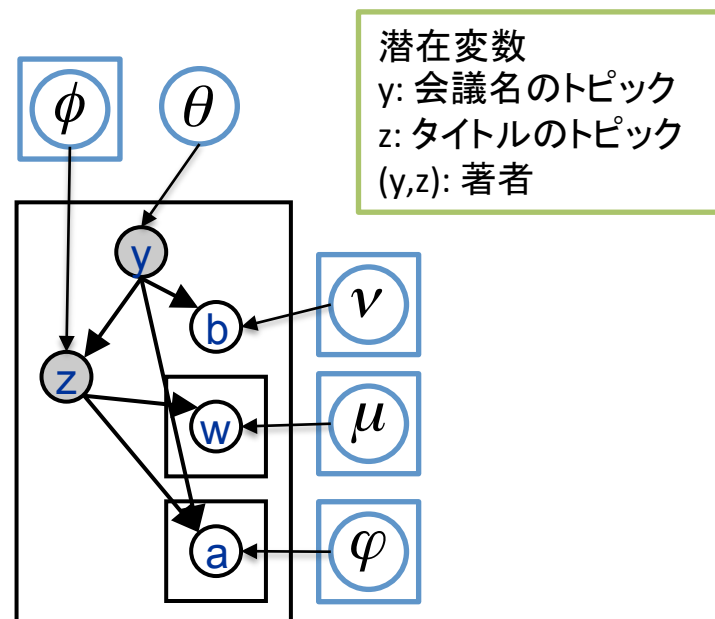
Authors	Title	Venue	Year	Pages
---------	-------	-------	------	-------

著者名の曖昧性解消

- 潜在変数を用いて研究者の執筆した論文をモデル化
- 共著者、タイトル、会議名の共起に基づいて研究者の特性を表す



Naïve Bayes Mixture Model (NBM)



Two-variable Mixture Model (TVM)

Masada, Takasu, Adachi, "Citation Data Clustering for Author Name Disambiguation," Proc. Intl. Conf. on Scalable Information Systems (INFOSCALE 2007), p.8, 2007.

データセット

DBLP ~50 full names

abbreviated name	# of fullnames	# of data	abbreviated name	# of fullnames	# of data	abbreviated name	# of fullnames	# of data
s.lee	161	971	c.lee	75	468	jzhang	60	308
j.lee	134	892	y.kim	74	373	s.li	59	242
jkim	129	769	h.chen	74	419	z.li	56	210
jwang	112	575	x.zhang	72	287	jwu	56	320
s.kim	108	598	y.lee	71	385	jlin	55	196
y.wang	101	533	k.kim	71	330	z.zhang	54	255
h.kim	100	506	x.li	69	315	s.liu	54	122
h.lee	99	346	s.park	69	379	h.liu	54	197
x.wang	86	322	jpark	68	376	d.kim	53	304
jchen	86	487	y.liu	67	313	y.yang	51	250
s.wang	84	274	c.wang	65	360	x.liu	51	187
y.zhang	83	391	s.chen	64	297	c.chen	51	462
y.chen	81	525	z.wang	63	142	m.lee	50	309
k.lee	81	380	jliu	63	406	lwang	50	253
h.wang	79	380	h.li	63	220	jyang	50	301
y.li	76	261	jli	61	314			

評価結果

K=256

Table 2: Evaluation results under the assumption that the true number of clusters is unknown.

method	P_{mic}	P_{mac}	R_{mic}	R_{mac}	F_{mic}	F_{mac}
NBMa	0.7034	0.9026	0.1314	0.3586	<u>0.2140</u>	<u>0.5085</u>
NBM	0.8501	0.8890	0.0741	0.2488	0.1343	0.3859
TVM	0.7807	0.8653	0.0985	0.2995	0.1710	0.4408
k -means	0.7686	0.8055	0.0612	0.2133	0.1115	0.3340

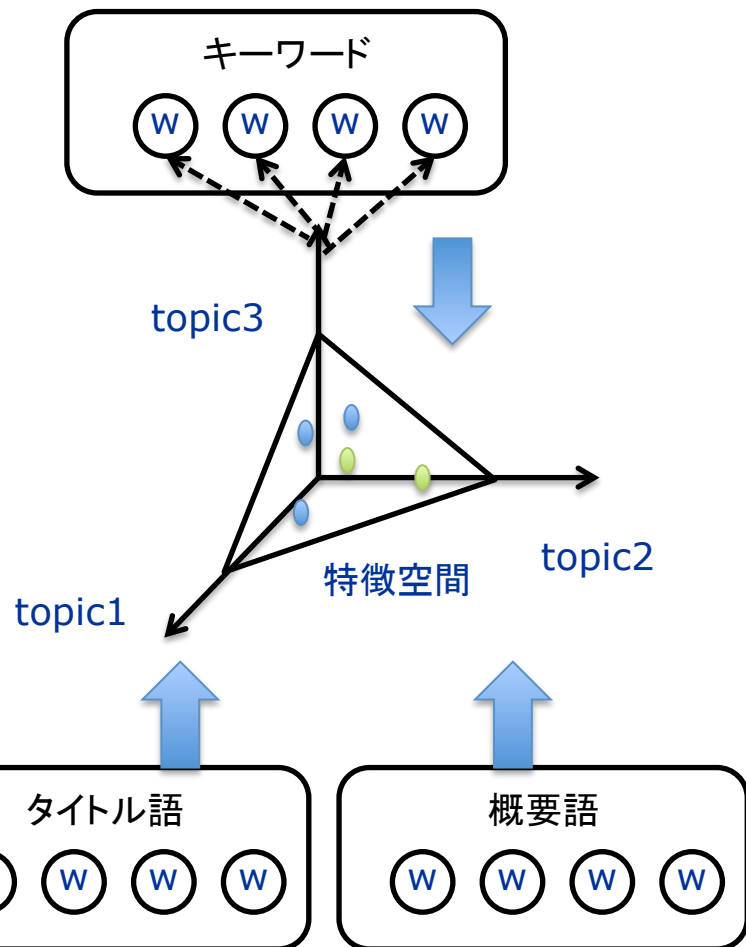
Table 3: Evaluation results under the assumption that the true number of clusters is known.

method	P_{mic}	P_{mac}	R_{mic}	R_{mac}	F_{mic}	F_{mac}
NBMa	0.6002	0.7729	0.1701	0.4051	<u>0.2574</u>	<u>0.5282</u>
NBM	0.5208	0.5277	0.1163	0.2804	0.1856	0.3632
TVM	0.5469	0.6358	0.1367	0.3278	0.2129	0.4294
k -means	0.3555	0.3517	0.0731	0.1916	0.1174	0.2424

発表の概要

- **学術情報抽出**
 - 参考文献文字列の解析と抽出
 - タイトルページの解析と抽出
- **書誌マッチング**
 - 木マッチングを用いた書誌統合
 - 著者名の曖昧性解消
- **情報分析**
 - 論文のキーワードの推薦
 - 学術研究のトピック分析

キーワード推薦



Mozilla Firefox
http://www.idea.nii.ac.jp/takasu/tmp/crosslingual/4.html

10026 / AA11546431

Reliable Total Dominant Pruning : MANETのための部分的ユニキャスト : 再送による高信頼フラッディング方式(アクティブネットワーク, IP-VPN, ネットワークセキュリティ, 超高速ネットワーク, P2P通信, ネットワークソフトウェア, 一般)

Japanese title

モバイルアドホックネットワーク(MANET)は、4-スステーションなどを必要とせず自己組織的に形成されるモバイル無線ネットワークである。各ホストはルータとして動作し、他のホスト宛のバケットを中継する。MANETにおいて必要不可欠なオペレーションであるフラッディング、すなわち一つのホストがネットワーク内の他の全てのホスト宛にバケットを送る処理は、しばしばブロードキャストストーム問題を発生させる。さらにトポロジーの変化が頻りに発生、フラッディングの信頼性が低下し、コストが高くなる。本稿では、分散型ツリーに基づいた信頼性のあるフラッディング手法を提案し、その設計について詳述する。

Japanese abstract

モバイルアドホックネットワーク / 信頼性のあるフラッディング / 部分的ユニキャスト再送 / 無線マルチホップ
モバイルアドホックネットワーク (dfrank=651, lrank=2)

high frequent ATM (dfrank=1), QoS (dfrank=10), TCP (dfrank=17), IP (dfrank=36), マルチキャスト (dfrank=24)

mid frequent P2P (dfrank=339), ルータ (dfrank=310), インテリジェントネットワーク (dfrank=319), B-ISDN (dfrank=319)

low frequent リアルタイム通信 (dfrank=631), モバイルアドホックネットワーク (dfrank=651), IP電話 (dfrank=698), FEC (dfrank=644), OSMA (dfrank=671)

recommended keywords

Reliable Total Dominant Pruning : A Reliable Flooding Scheme with Partial Unicast Retransmissions for MANET

English title

Mobile Ad Hoc Networks (MANETs) are self-organizing mobile wireless networks where each host also acts as a router and forwards packets toward their destination. Flooding in MANET often incurs the broadcast storm problem. This, combined with topology changes inherent to MANET, hampers the performance of these protocols. In this paper, we discuss the design of a reliable, distributed tree based, flooding protocol for MANET.

English abstract

Mobile Ad Hoc Networks / Reliable Flooding / Partial Unicast Retransmission / Wireless Multi-hop / RTDP
Mobile (dfrank=432, lrank=1)

high frequent network (dfrank=75), IP (dfrank=20), ATM (dfrank=1), QoS (dfrank=4), Wireless LAN (dfrank=17)

mid frequent delay (dfrank=222), Protocol (dfrank=389), Streaming (dfrank=253), evaluation (dfrank=369), delay (dfrank=281)

low frequent traffic (dfrank=437), mobile (dfrank=618), model (dfrank=575), IP Multicast (dfrank=569), distributed processing (dfrank=617)

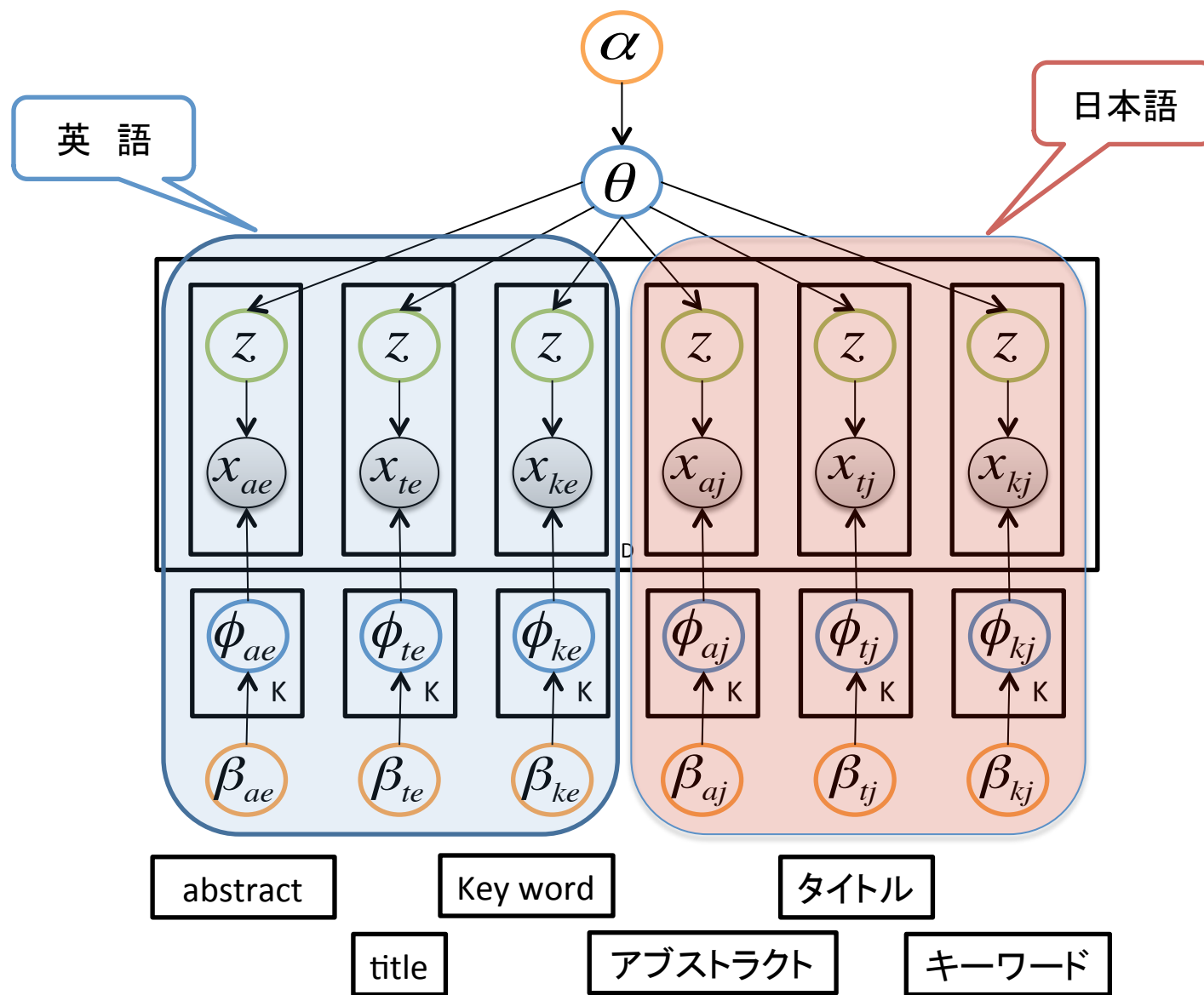
recommended keywords

10027 / AA11546431

マルチエージェントに基づく大規模災害時における知人安否確認システム(アクティブネットワーク, IP-VPN, ネットワークセキュリティ, 超高速ネットワーク, P2P通信, ネットワークソフトウェア, 一般)

災害時において、携帯電話やインターネットを用いて知人の安否確認を行うシステムが提案されている。しかし、大規模災害時には有線インフラを用いた通信が断絶する。

多言語キーワード推薦モデル



キーワード

高頻度キーワード:

アルゴリズム, 並列処理, 計算機シミュレーション, 画像処理, プロトコル, 移動通信, 光通信, 神経回路網, 計算の複雑さ, 性能評価, データベース, ユーザインタフェース, 音声認識, 自然言語処理, コンピュータグラフィックス, 特徴抽出, インターネット, 遺伝的アルゴリズム, インターネット, ヒューマンインタフェース, 分散処理, 電磁場解析, 三次元, 符号化, ATM網, 符号誤り, 計算機アーキテクチャ, 回路設計, CDMA, 通信制御, 光ファイバ, 計算機網, ...

中頻度キーワード:

コンパイラ, 音声合成, 電磁干渉, 制約条件, スペクトル拡散通信, ウェーブレット変換, アミノ酸配列, 公開鍵暗号, 遅延特性, マルチメディア, 仕様記述, 待ち行列, グラフ理論, メンバシップ関数, エッジ検出, 専用プロセッサ, データ通信, コンピュータセキュリティ, 認証, ヒト, 電磁干渉, 立体視, データベース管理システム, ソフトウェア設計, 音楽, ネットワークアーキテクチャ, 液晶ディスプレイ, ネットワーク管理, 文書処理, ...

程頻度キーワード:

計算量理論, 酸化膜, 計算機シミュレーション, 人工知能推論, ひ化ガリウム, エピタクシー, 信号処理, 木【グラフ】, 位置制御, CMOS, モーメント法, フィルタリング, 雑音低減, 電磁波散乱, 表面粗さ, テレビジョン放送, 速度制御, デジタル画像, 回折格子, 復号化, インターネット, 画像信号, 強化学習, 知能ロボット, アンテナ設計, 基質特異性, X線回折, ...

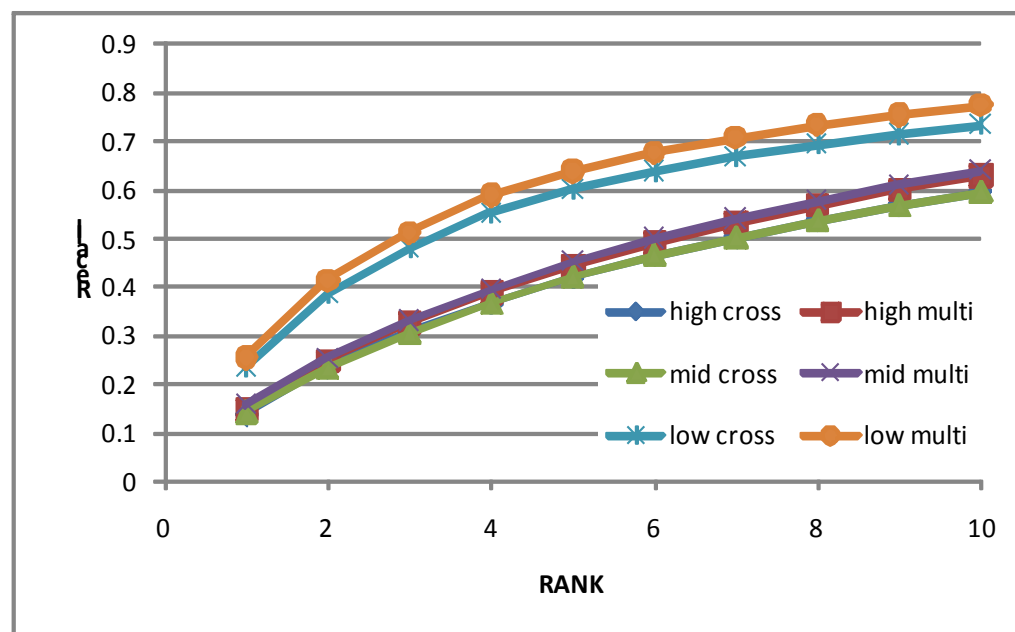
キーワード推薦精度

- データセット

- 計算機科学分野の論文: 245,840 論文

- 語数

- 日本語タイトル: 4,807
- 日本語概要: 6,815
- 日本語キーワード: 5,604
- 英語タイトル: 7,759
- 英語概要: 13,368
- 英語キーワード: 5,742
- 著者: 15,543



Cross-lingual keyword recommendation

発表の概要

- **学術情報抽出**
 - 参考文献文字列の解析と抽出
 - タイトルページの解析と抽出
- **書誌マッチング**
 - 木マッチングを用いた書誌統合
 - 著者名の曖昧性解消
- **情報分析**
 - 論文のキーワードの推薦
 - 学術研究のトピック分析

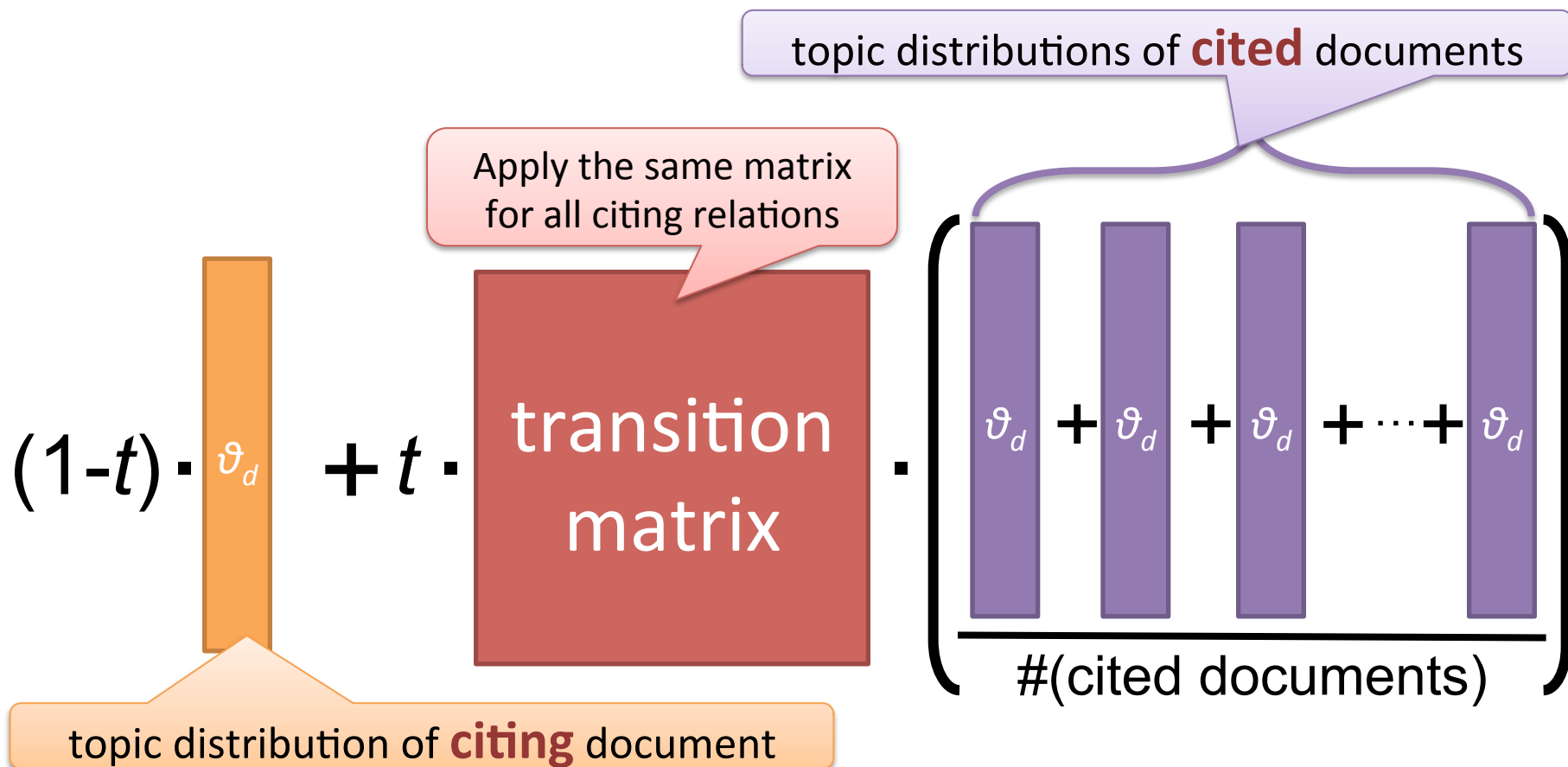
引用関係をトピック確率分布の設定に利用

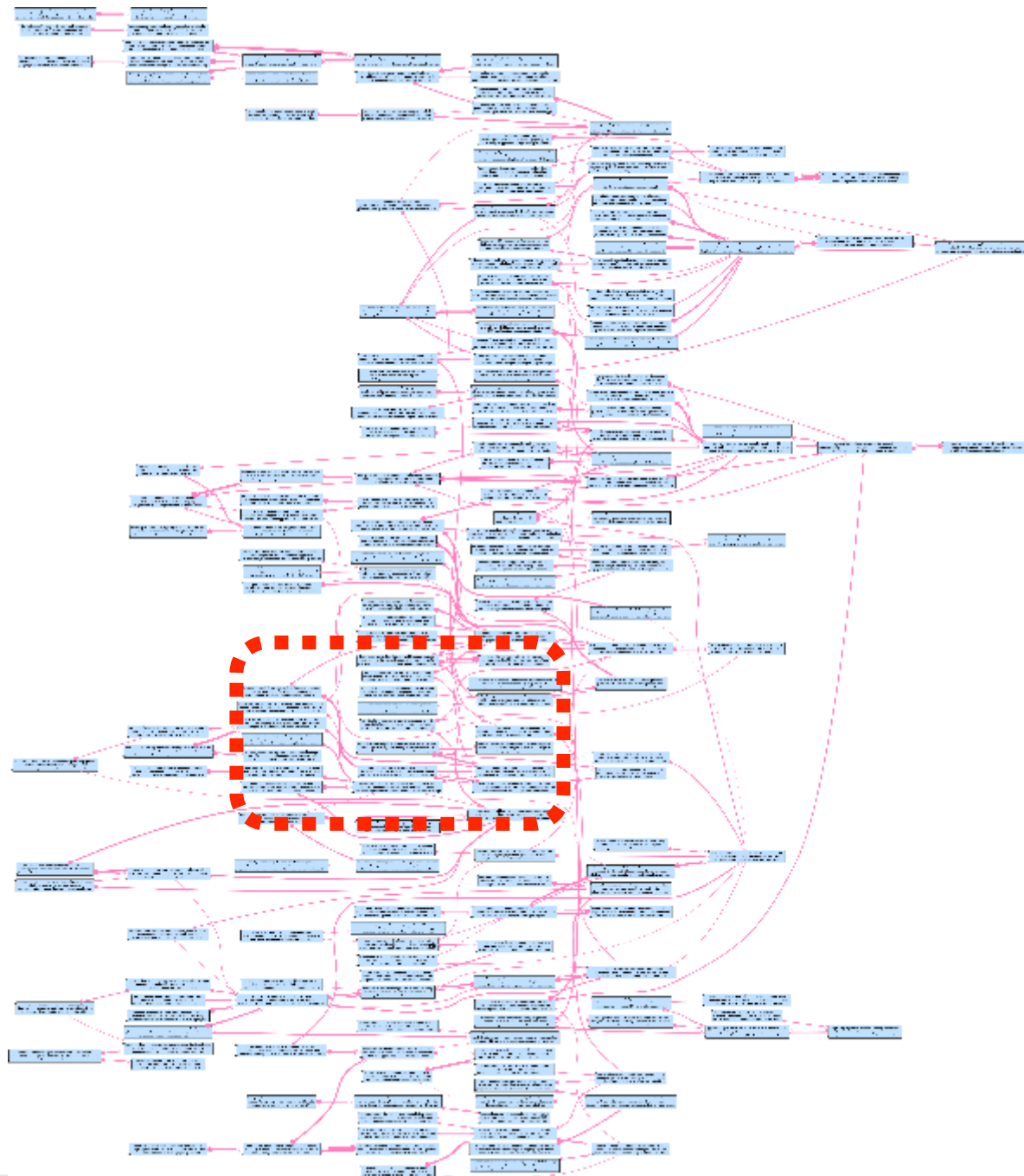
- TERESA (Topic Evolution from REferences in Scientific Articles)
 - 引用されている文献のトピック確率分布が、引用している文献のトピック確率分布に影響を与える
 - 研究トピックの変化（進化）を分析

Masada, Takasu, "Extraction of Topic Evolutions from Reference in Scientific Articles and Its GPU Acceleration," Proc. ACM Conf. on Information and Knowledge Management (CIKM 2012), 2012.

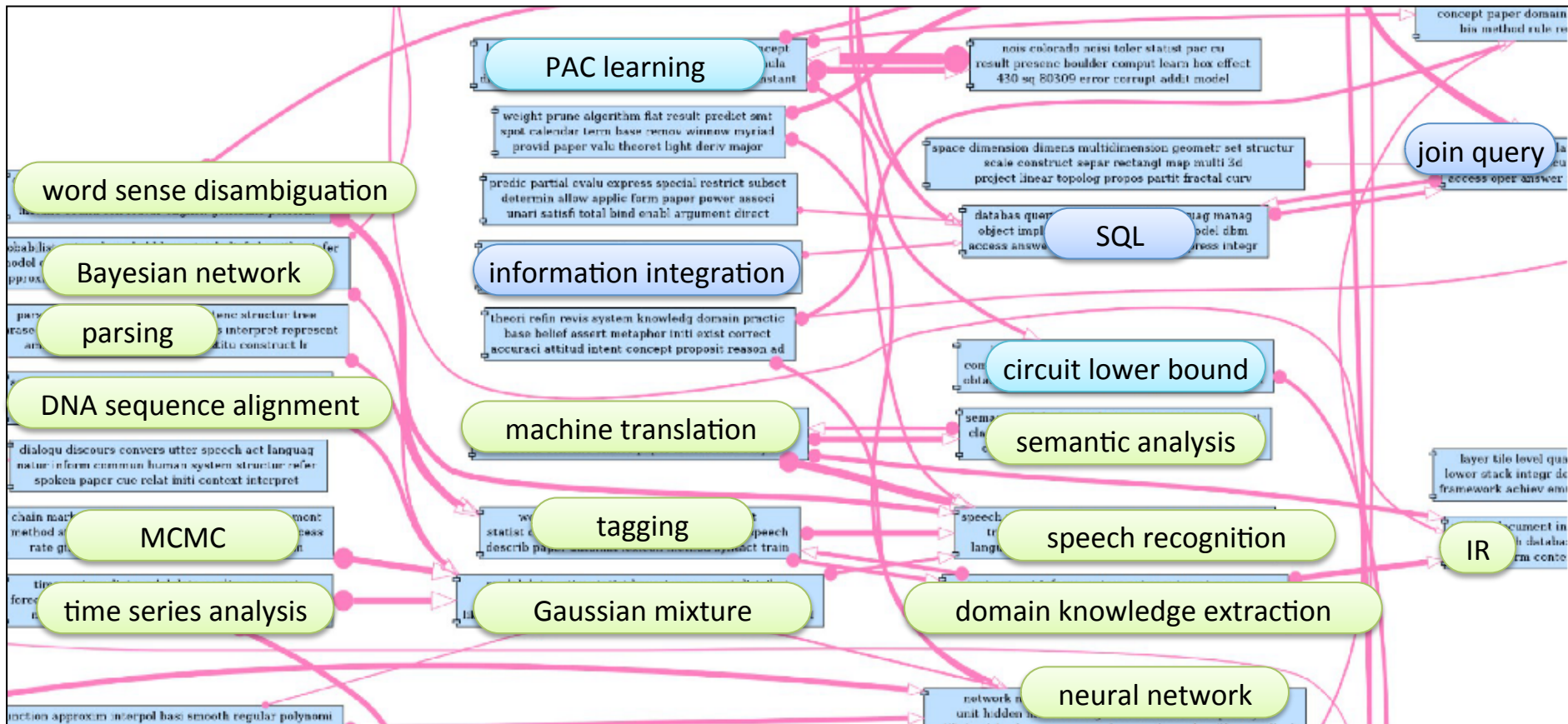
TERESAの確率モデル

- トピック確率分布を以下のように設定





研究トピックのトレンド分析例 (Cora data set)



まとめ

- **学術情報の活用のための情報抽出・統合の取組を紹介**
 - 大規模学術情報を活用するための多様性に対処可能な技術の研究
- **学術情報の俯瞰を支援する情報分析法の紹介**
 - 潜在構造の推定に基づく分析技術