

【VLDB2011勉強会】

Session 3: Web

担当：加藤 誠(京都大学)

まとめ

▶ Output URL Bidding

- ▶ Panagiotis Papadimitriou, Hector Garcia-Molina (Stanford University), Ali Dasdan, Santanu Kolay (Ebay Inc)
- ▶ 「検索クエリ」ではなく「検索結果に表れるURL」に広告をうつ

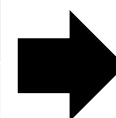
▶ Automatic Wrappers for Large Scale Web Extraction

- ▶ Nilesh Dalvi, Ravi Kumar (Yahoo!), Mohamed Soliman (U. of Waterloo)
- ▶ 自動的なラッパーの生成

▶ Recovering Semantics of Tables on the Web

- ▶ Petros Venetis (Stanford University), Alon Halevy, Jayant Madhavan, Marius Pasca, Warren Shen, Fei Wu (Google Inc), Gengxin Miao (University of California, Santa Barbara), Chung Wu (Google)

2	Entity Matching	白川(阪大)
3	Web	加藤(京大)



セッション3 (セッション番号):
Web (セッション名) の担当は
加藤 (人名)

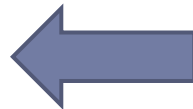
Output URL Bidding

Panagiotis Papadimitriou, Hector Garcia-Molina (Stanford University), Ali Dasdan, Santanu Kolay (Ebay Inc)

▶ 背景

▶ たくさんのクエリに広告をうつのは大変

パイレーツオブカリビアン 生命の泉	
プリンセストヨミ	



[パイレーツオブカリビアン生命の泉](#)

アマゾンでは常時無料配送／一部除くお急ぎ便利用で当日、翌日にお届け。

www.amazon.co.jp

...全映画分で数千クエリ?

▶ 提案

▶ 検索結果中に出てくるURLに広告をうったらどうか?

「movie.goo.ne.jp」が出てきたら広告を出す



[パイレーツオブカリビアン生命の泉](#)

アマゾンでは常時無料配送／一部除くお急ぎ便利用で当日、翌日にお届け。

www.amazon.co.jp

[プリンセストヨミ - goo 映画](#) [Translate this page](#)

映画『プリンセストヨミ』大ボラ見た(笑)～雑感です。2011年11月7日 3時01分 ** (yutake イヴのモノローグ) ** 映画:プリンセストヨミ 2011年7月17日 0時19分 よしなしごと プリンセストヨミ 2011年7月3日 1時56分 映画の庭(分館)

movie.goo.ne.jp/contents/movies/MOVCSTD17853 - Cached page

[パイレーツ・オブ・カリビアン／生命\(いのち\)の泉 - goo 映画](#) [Translate this page](#)

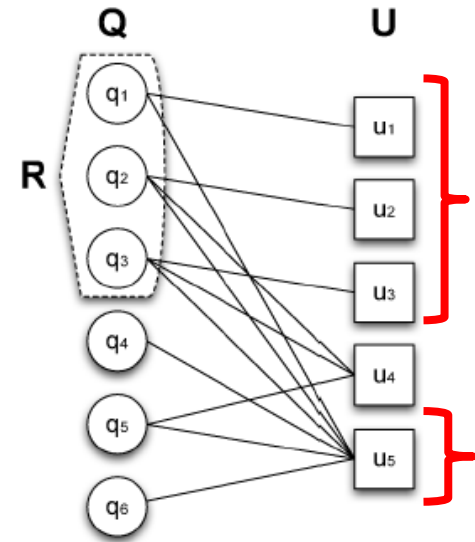
古い仲間のギブスを絞首台から救うべく、ロンドンに足を踏み入れたジャック・スパロウ。そこでは、「ジャックが『生命の泉』を目指すため乗組員を集めている」という噂が流れていた。偽のジャックを求めて行った先には、かつて彼が愛した女海賊アンジェリカがいた。彼女に捕えられたジャックは、驚くべき告白を聞く。アンジェリカの父は残忍で知られる海賊 ...

movie.goo.ne.jp/contents/movies/MOVCSTD17553

数URLを指定しておけば十分カバーできる
(コンパクト!)

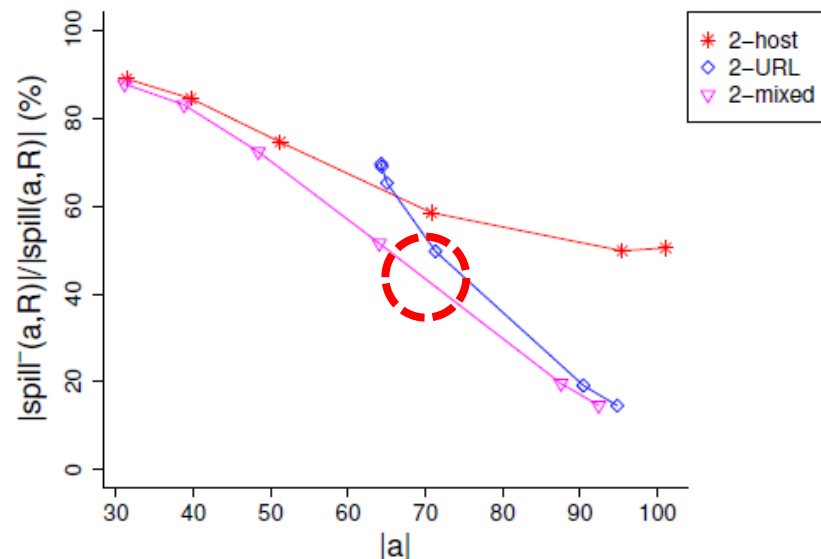
Output URL Biddingはどれくらい現実的か？

- ▶ Adクエリ集合をカバーする AdURL集合に変換（最適化問題）
 - ▶ コンパクト性: AdURLの数
 - ▶ 冗長性: AdURL集合でカバーしてしまう Adクエリ集合以外のクエリ数
- ▶ 既存のAdクエリ集合をコンパクトで非冗長なAdURL集合に変換できたら現実的



- ▶ 実験
 - ▶ 2,251 ads, 13M queries, 63M URLs
 - ▶ Adクエリの1/3 ($|a| = 70$)ほどで問題なくカバー可能

冗長なクエリのうち半数が潜在的に適合したクエリ



Automatic Wrappers for Large Scale Web Extraction

Nilesh Dalvi, Ravi Kumar (Yahoo!), Mohamed Soliman (U. of Waterloo)

▶ 背景

- ▶ アノテーションなしで情報抽出ラッパーを作りたい
 - ▶ ただし、「抽出したい情報」はおおよそルール化できる
 - 大学名: (.+[大])
 - ▶ ルールを使って自動でアノテーションできる
- ▶ 正確なアノテーションがないので、ノイズが含まれてしまう場合がある
 - ▶ 従来のモデルはノイズを想定していない

▶ 提案モデル

アノテーションの正確さ

繰り返し構造のきれいさ

```
<tr>
  <td>2</td>
  <td>Entity Matching</td>
  <td>白川(阪大)</td>
</tr>
<tr>
  <td>3</td>
  <td>Web</td>
  <td>加藤(京大)j</td>
</tr>
<tr>
  <td>5</td>
  <td>Uncertain Data</td>
  <td>高阪大地(名大)</td>
</tr>
```

$$\operatorname{argmax}_X P(L|X)P(X)$$

X: 抽出された情報のリスト
L: 正解ラベル

モデルの直感的説明

n1	a1	z1	p1
n2	a2	z2	p2
n3	a3	z3	p3
n4	a4	z4	p4

$P(L|X)$
 $P(X)$

n1	a1	z1	p1
n2	a2	z2	p2
n3	a3	z3	p3
n4	a4	z4	p4

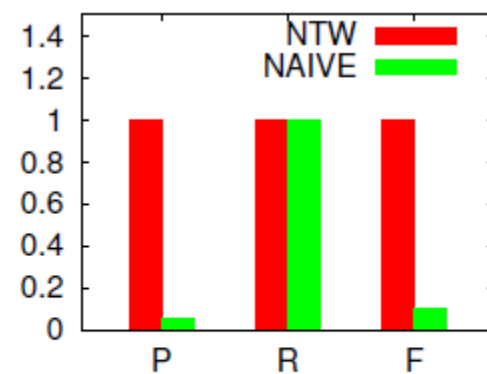
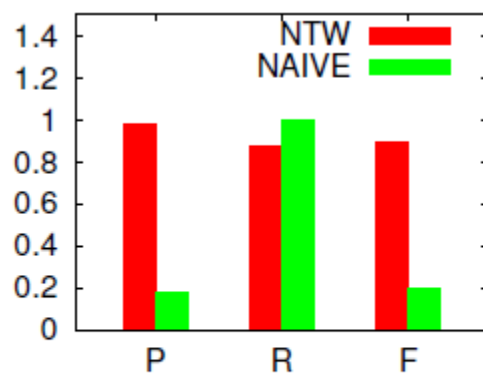
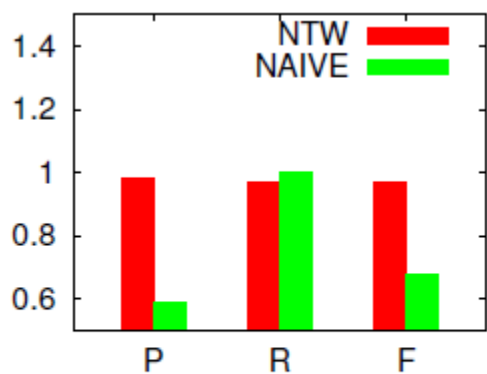
$P(L|X)$
 $P(X)$

n1	a1	z1	p1
n2	a2	z2	p2
n3	a3	z3	p3
n4	a4	z4	p4

$P(L|X)$
 $P(X)$

例はVLDB2011の発表スライド21P目から引用

実験



(d) Accuracy of XPATH on DEAL- (e) Accuracy of LR on DEALERS. (f) Accuracy of XPATH on DISC.ERS.

Recovering Semantics of Tables on the Web

Petros Venetis (Stanford University), Alon Halevy, Jayant Madhavan, Marius Pasca, Warren Shen, Fei Wu (Google Inc),
Gengxin Miao (University of California, Santa Barbara), Chung Wu (Google)

▶ 背景

- ▶ Web検索でテーブルを探す機会が多い
- ▶ ただしテーブルの意味が明示的にわからないときもある

▶ 目的

- ▶ カラム情報の復元(カラムにあるデータは何か)
- ▶ 関係情報の復元(横に並んでいるデータはどのような関係か)

カラム情報

セッションID

11:45-12:45 関西会場発表1(4件, 60分)

2	Entity Matching	白川(阪大)
3	Web	加藤(京大)
5	Uncertain Data	石川研学生(名大)
6	Database Design	櫻 惇志 (NAIST)

カラム情報

人名

セッション名

関係情報

[人名]は[セッション名]を担当

カラム情報の復元

日本	東京	アジア
中国	北京	アジア
アメリカ	ワシントンDC	北米
ポーランド	ワルシャワ	ヨーロッパ

- ▶ カラム情報の復元 → is-a関係抽出
 - ▶ 言語パターンの利用: **cities** *such as* **Tokyo** and **Beijing**
 - ▶ 推定
 - ▶ 「cities such as Warszawa」という表記は少ない...
 - ▶ →「東京」「ワルシャワ」の特徴量(前後の語など)が似ていればワルシャワもcity (尤度最大のところへ分類)

関係情報の復元

日本	東京	アジア
中国	北京	アジア
アメリカ	ワシントンDC	北米
ポーランド	ワルシャワ	ヨーロッパ

- ▶ 関係情報の復元 → 3つ組発見
 - ▶ TextRunner [Banko IJCAI 07]
 - ▶ Conditional Random Fieldを使った手法
 - ▶ 特徴量: POS, 前後の語など
 - ▶ 例: Tokyo, which is located at the center of Japan, **is the capital of** Japan
→ < Tokyo, **is the capital of**, Japan >

主題となるカラム検出・テーブル検索

▶ 主題となるカラム検出

11:45-12:45 関西会場発表1(4件, 60分)

2	Entity Matching	白川(阪大)
3	Web	加藤(京大)
5	Uncertain Data	石川研学生(名大)
6	Database Design	樺 惇志 (NAIST)

SVM分類器精度: 94%

[83% (一番左のカラムを選択するベースライン)]

▶ テーブル検索

- ▶ カラム情報・関係情報を復元しその情報でインデックスを作りテーブル検索を実装
- クエリ例: *<presidents, political party>* *<laptops, price>*
- ▶ (当たり前だが...) Googleの検索結果よりもPrecision/Recallが高い