【WWW2014勉強会】

Session 15: Web Mining 2

担当：大島裕明（京都大学）

Session 15: Web Mining 2

Codewebs: Scalable Homework Search for Massive Open Online Programming Courses

- A. Nguyen, C. Piech, J. Huang, L. Guibas (Stanford Univ.)

Joint Question Clustering and Relevance Prediction for Open Domain Non-Factoid Question Answering

- S. Chaturvedi (Univ. Maryland), V. Castelli, R. Florian, R. Nallapti, H. Raghavan (IBM Watson Research)

Knowledge Base Completion via Search-Based Question Answering

- R. West (Stanford Univ.), E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, D. Lin (Google)

Codewebs: Scalable Homework Search for Massive Open Online Programming Courses

A. Nguyen, C. Piech, J. Huang, L. Guibas (Stanford Univ.)

- MOOCsにおける宿題の評価
  - 教師1名に対して1万人の学生 → 評価が困難

MOOCs宿題用プログラミングコード検索システム

  - 同じ言語のコード／同時／同じ問題のプログラム
    - 同じ問題を解く方法はいくつも存在する（例：forとwhile）
- プログラムの宿題の評価
  - 正しさは自動判別可能だが、創造性を評価する必要性
  - Mechanical Turk？評価はだれにでもできる訳ではない
  - 学生同士での相互評価はある程度成功したが…
- Stanford大学でのMachine Learningクラス
  - 12万人の登録者中10,405人が8回全ての宿題提出

Codewebs: Scalable Homework Search for Massive Open Online Programming Courses

- 宿題データ
  - 全42問
  - 提出総数：100万
  - 1問あたり提出数：2.4万
  - 1問あたり学生数：1.6万
  - 平均行数：16.44行
  - ASTの平均ノード数：164
- AST：Abstract Syntax Tree
  - プログラムを抽象化して木で表現
    - プログラムの総数に対して、**ユニークなASTはほぼ半数**
    - ノードはAssign、Statementなどのタイプを認識
      - それ以外は外部知識を利用せず

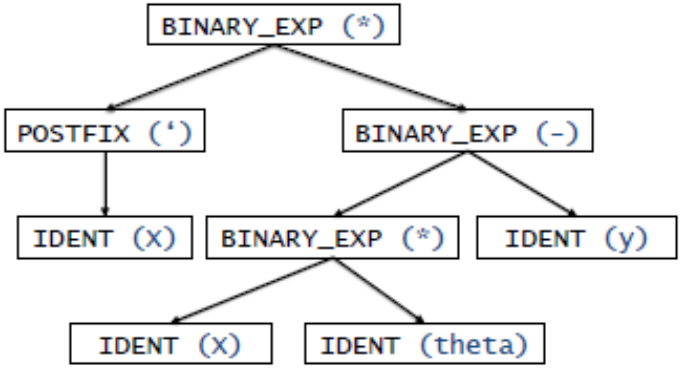
```
function [theta, J_history]
    = gradientDescent(X, y, theta, alpha, num_iters)

%GRADIENTDESCENT gradient descent to learn theta
%   updates theta by taking num_iters gradient
%   steps with learning rate alpha.

m = length(y); % number of training examples
J_history = zeros(num_iters, 1);

for iter = 1:num_iters
    theta = theta-alpha*1/m*(X'*(X*theta-y));
    J_history(iter) = computeCost(X, y, theta);
end
```

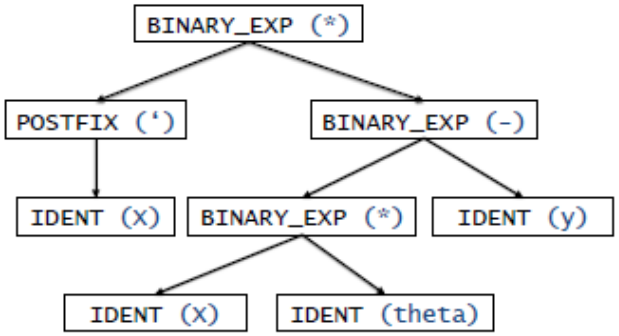
最急降下法の関数



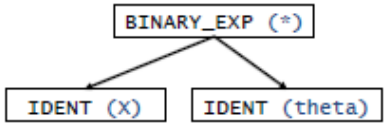
ASTで
$X' * (X * \theta - y)$
を表現

Codewebs: Scalable Homework Search for Massive Open Online Programming Courses

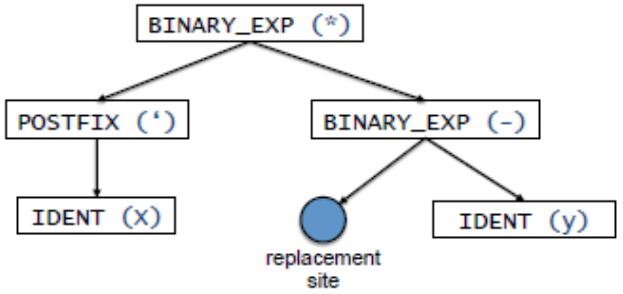
- 検索クエリ
  - ×テキストや正規表現 → 結局、コードを読むが必要ある
  - Code Phrases：ASTのサブグラフ
    - 3つの形式
      - Subtree：ASTの部分木
      - Subforest：「StatementタイプがルートのSubtree」の列
      - Context：ASTからあるSubtreeの部分を●に置換したもの



AST：A



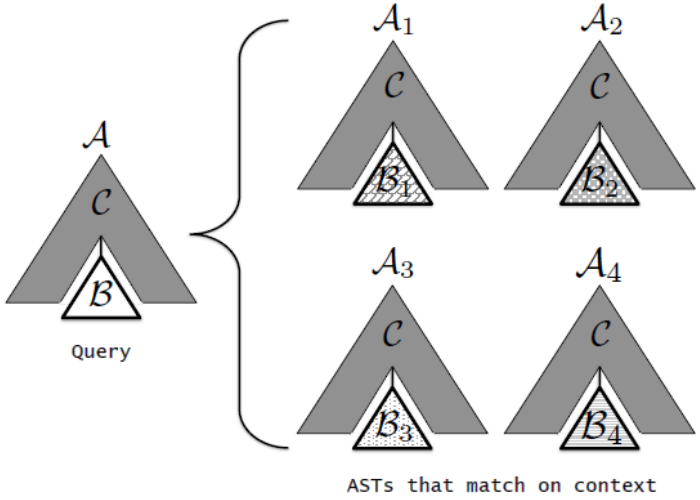
AのSubtree



AのContext

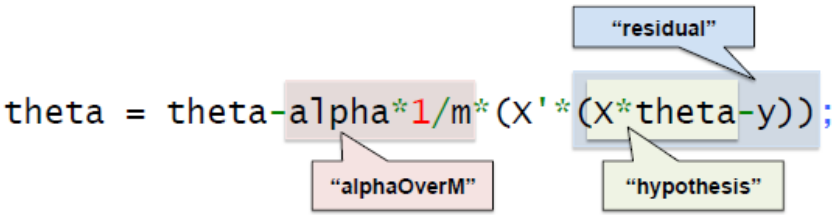
Codewebs: Scalable Homework Search for Massive Open Online Programming Courses

- どのようなASTをほぼ同一とみなすか
  - あるSubtree Bを含むすべてのASTで、BをB′に置き換えても結果が変わらないなら、BとB′を同一とみなす
    - 緩和策
      - どんな入力でも結果不変 → ユニットテストで結果同じ
      - Contextが十分に類似する2つのサブツリーで考える
  - BとB′の類似度
    - 「Bを含むA」と「B′を含むA′」で、「AからBを取り除いた部分」＝「A′からB′を取り除いた部分」で、AとA′のユニットテストの結果が同じなら、類似



Codewebs: Scalable Homework Search for Massive Open Online Programming Courses

- 結果例



{m}

m	rows (X)	rows (y)
size (X, 1)	length (y)	size (y, 1)
length (x (:, 1))	length (X)	size (X) (1)

{alphaOverM}

alpha / {m}	1 * alpha / {m}	alpha .* 1 / {m}
1 / {m} * alpha	alpha .* (1 / {m})	alpha ./ {m}
alpha * inv ({m})	alpha * pinv ({m})	1 .* alpha ./ {m}
alpha * (1 ./ {m})	alpha * 1 ./ {m}	alpha * (1 / {m})
.01 / {m}	alpha .* (1 ./ {m})	alpha * {m} ^ -1

{hypothesis}

```
(X * theta)
(theta' * X')'
[X] * theta
(X * theta (:))
theta(1) + theta (2) * X (:, 2)
                ⋮
sum(X.*repmat(theta',{m},1), 2)
```

{residual}

```
(X * theta - y)
(theta' * X' - y')'
({hypothesis} - y)
({hypothesis}' - y')'
[{hypothesis} - y]
            ⋮
sum({hypothesis} - y, 2)
```

Joint Question Clustering and Relevance Prediction for Open Domain Non-Factoid Question Answering

S. Chaturvedi (Univ. Maryland), V. Castelli, R. Florian, R. Nallapti, H. Raghavan (IBM)

- 背景
  - 自然言語による検索：質問応答を目的とする
    - Factoid：事実について
      - 例：アメリカの大統領はだれ？
    - Non-Factoid：意見、教えて、レビュー
      - 例：中絶は道徳的か？
      - 例：NYのストップアンドフリスクのインパクトは？
    - Factoidは分類がある（人？場所？天気？定義？）
    - Non-Factoidは主観的で分類難しい

Factoid-styleではない質問の分類体系の構築

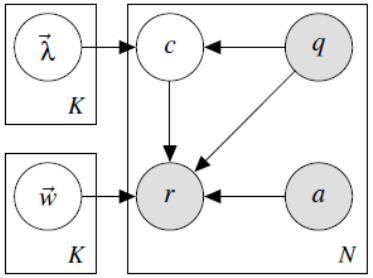
  - 回答をクラスタリングしてData Drivenに構築
  - ロジスティック回帰のモデルを提案

Joint Question Clustering and Relevance Prediction for Open Domain Non-Factoid Question Answering

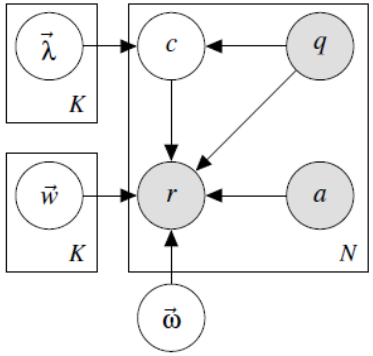
- 問題設定
  - データ（質問q、回答a、適合性r）
    - 質問q：Non-Factoidの質問
    - 回答a：スニペット（文など）
    - 適合性r：aはqの回答として適合（1）か非適合（0）
  - モデル
    - カテゴリc：潜在的に存在するカテゴリ（K個）
      - r=1のときのみ、質問aとカテゴリcは関係づけられる
  - 特殊
    - h：Global（0）かLocal（1）かを表す項
- 特徴ベクトル**（かなり特殊（3.4節））**
  - f_q：質問のみから作成された特徴ベクトル
  - f_{qa}：質問と回答から作成された特徴ベクトル

Joint Question Clustering and Relevance Prediction for Open Domain Non-Factoid Question Answering

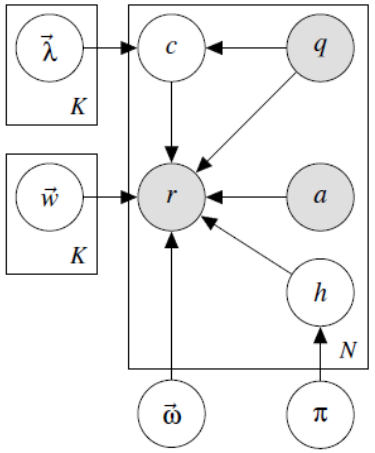
- 提案手法：3つのモデル



(a) Logistic Regression Mixture (LRM)



(b) Glocal Logistic Regression Mixture (G-LRM)



(c) Mixture Glocal Logistic Regression Mixture (MG-LRM)

  - LRM：cごとに異なる適合性推定モデルを用いる
  - G-LRM： cのためのGlobal変数ωを用いる
  - MG-LRM：Globalのときは、qとaとωを使い、
    Localのときは、qとaとcとwを使う

- 目的

3つのモデルで共通

$$P(r|q,a) = \sum_{c}^{K} P(c|q,a)P(r|q,a,c)$$

$$\approx \sum_{c}^{K} P(c|q)P(r|q,a,c)$$

$$P(c|q) = \frac{e^{\vec{\lambda}_c \vec{f}_q}}{\sum_k e^{\vec{\lambda}_k \vec{f}_q}}$$

対数線型モデルで求める

残り

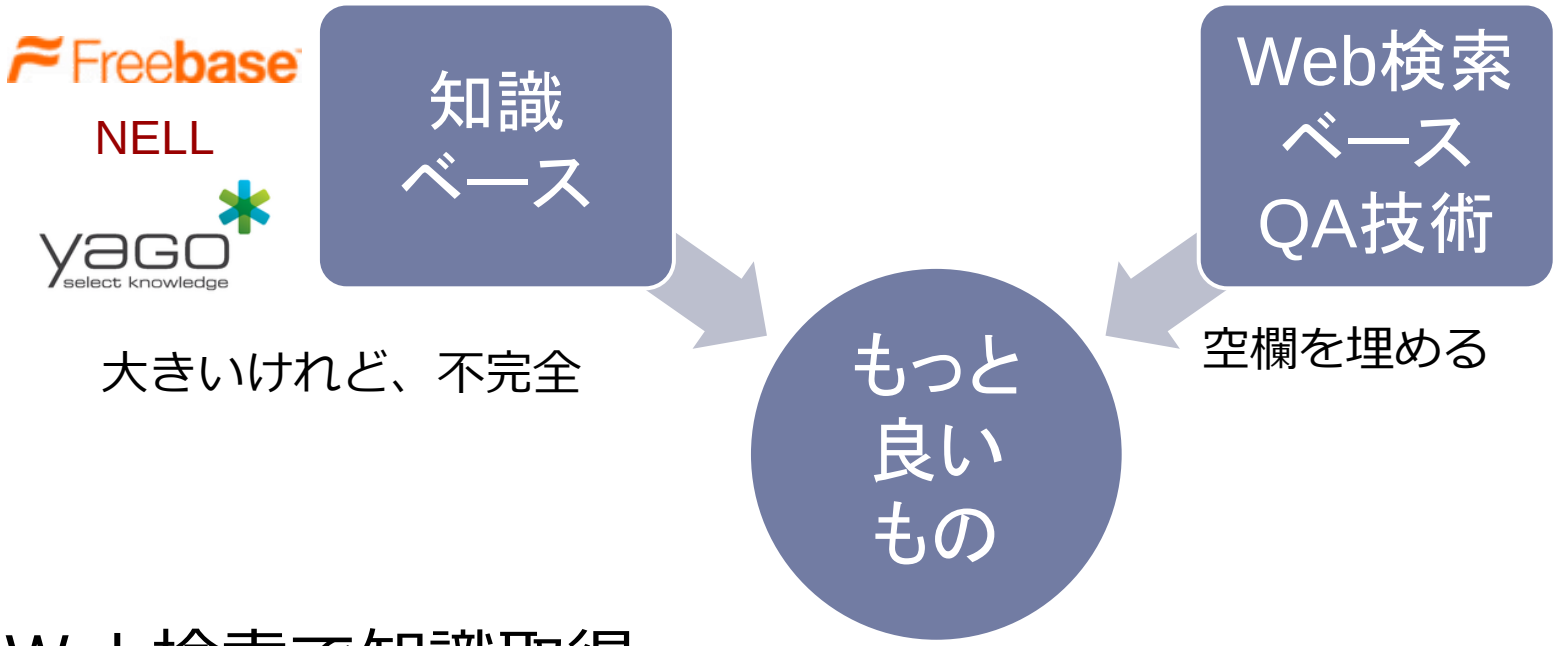
Joint Question Clustering and Relevance Prediction for Open Domain Non-Factoid Question Answering

- 実験：**適合性を推定できるかどうか**
  - データセット
    - BOLTデータセット（DARPA **B**road **O**perational **L**anguage **T**ranslation program）
    - TACデータセット
  - 人手による適合性付与
  - 5/6で訓練1/6でテスト

Model	BOLT Data				TAC Data			
	K	P	R	F	K	P	R	F
Oracle	–	63.73	42.33	50.87	–	–	–	–
LR	–	61.33	36.79	45.99	–	24.79	65.66	35.99
LR AnsTypeTruth	–	50.12	44.69	47.25	–	–	–	–
J48 [28]	–	44.60	43.20	43.88	–	–	–	–
LRM	4	59.36	39.03	47.09	9	24.43	71.41	36.40
G-LRM	4	59.97	40.36	48.25	2	24.41	71.80	36.44
G-LRM + Init	7	61.01	42.15	**49.85**	13	24.34	71.80	36.40
MG-LRM	6	59.66	41.75	49.12	18	26.31	58.88	36.37
MG-LRM + Init	7	64.48	39.43	48.93	2	24.53	72.98	**36.72**

Knowledge Base Completion via Search-Based Question Answering

R. West (Stanford Univ.), E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, D. Lin (Google)



- Web検索で知識取得
  - Frank Zappaの母の名前を知りたい
    - Query「who is the mother of Frank Zappa」で検索
    - →「The Mother of Invention」って、それはバンド名や！
    - 曖昧性をなくす語を追加する（例：Frank Zappaの出身地）

Knowledge Base Completion via Search-Based Question Answering

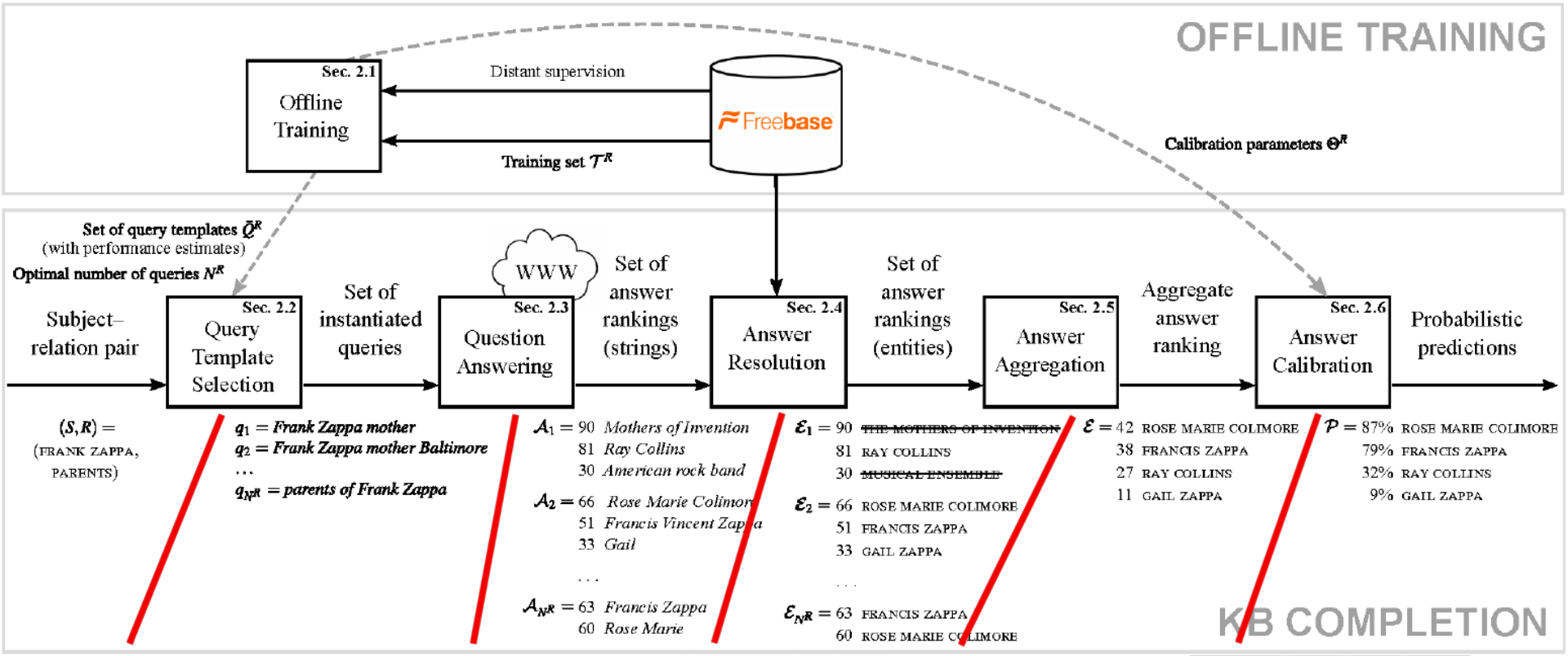
- 知識ベースの欠落情報
  - Freebaseにおける「人」情報の属性ごとの情報欠落率

関係	全300万	上位10万
職業	68%	24%
出生地	71%	13%
国籍	75%	21%
学歴	91%	63%
配偶者	92%	68%
両親	94%	77%
子ども	94%	80%
兄弟	96%	83%
民族	99%	86%

- 言語テンプレート
  - 該当の関係を表すテンプレートを10個取得
  - 両親
    - name of __'s father
    - __ mother
  - 出生地
    - what city was __ born in
- クエリの生成
  - 適当な語を追加して、disambiguation
    - Frank Zappa mother **Baltimore**
    - Birthplace of Michael Jackson **World Guide to Beer**

Knowledge Base Completion via Search-Based Question Answering

- 手法の概要



クエリ選択
あるテンプレートで共起しやすい関係を求める

詳しくいえないけど、Web検索ベース

取得された語の正体を判別

複数のクエリで取得された語を集約してランキング

ロジスティック回帰で確率っぽい数値に変換