Identification of non-functional requirements in textual specifications: A semi-supervised learning approach

Agustin Casamayor Daniela Godoy Marcelo Campo

Information and Software Technology
Volume 52, Issue 4, April 2010, Pages 436-445

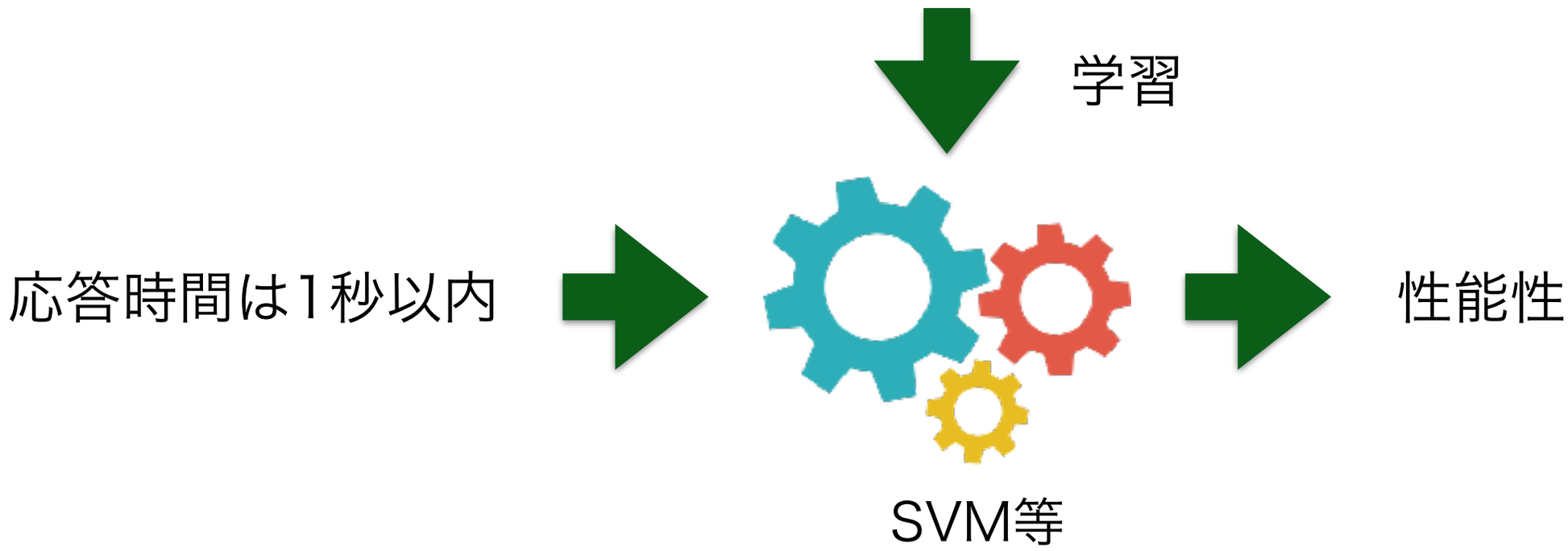
背景

- 非機能要件（NFR）を早期に特定することは設計やデザインの評価に重要である
- 人手のNFR特定は簡単そうで難しく時間がかかる
  - ドキュメントに記述されている要件は**機能要件（FR）についてがメイン**
  - NFRはFRの付属として記述されている
- テキスト分類技術を用いたNFR特定手法が提案されてきている

先行研究

- 教師あり学習を用いた手法が提案[1]

データの暗号化をすること	セキュリティ
応答時間は2秒以内	性能性
⋮	⋮



[1]A. Jalaji, R. Goff, M. Jackson, N. Jones, T. Menzies, Making sense of text: identifying nonfunctional requirements early. Technical Report, West Virginia University CSEE, 2006.

先行研究

- 教師あり学習を用いた手法が提案[1]

データの暗号化をすること	セキュリティ
応答時間は2秒以内	性能性
⋮	⋮

学習

大量の学習データが必要
学習データにない単語に対応できない

SVM等

[1]A. Jalaji, R. Goff, M. Jackson, N. Jones, T. Menzies, Making sense of text: identifying nonfunctional requirements early. Technical Report, West Virginia University CSEE, 2006.

研究の目的・アプローチ

- 目的
  - 学習データを減らしても精度のよいNFR特定手法を開発すること

- アプローチ
  - ナイーブベイズとEMアルゴリズムを用いた半教師あり学習
    - EMアルゴリズム：最尤推定を最大化するための手法

半教師あり学習

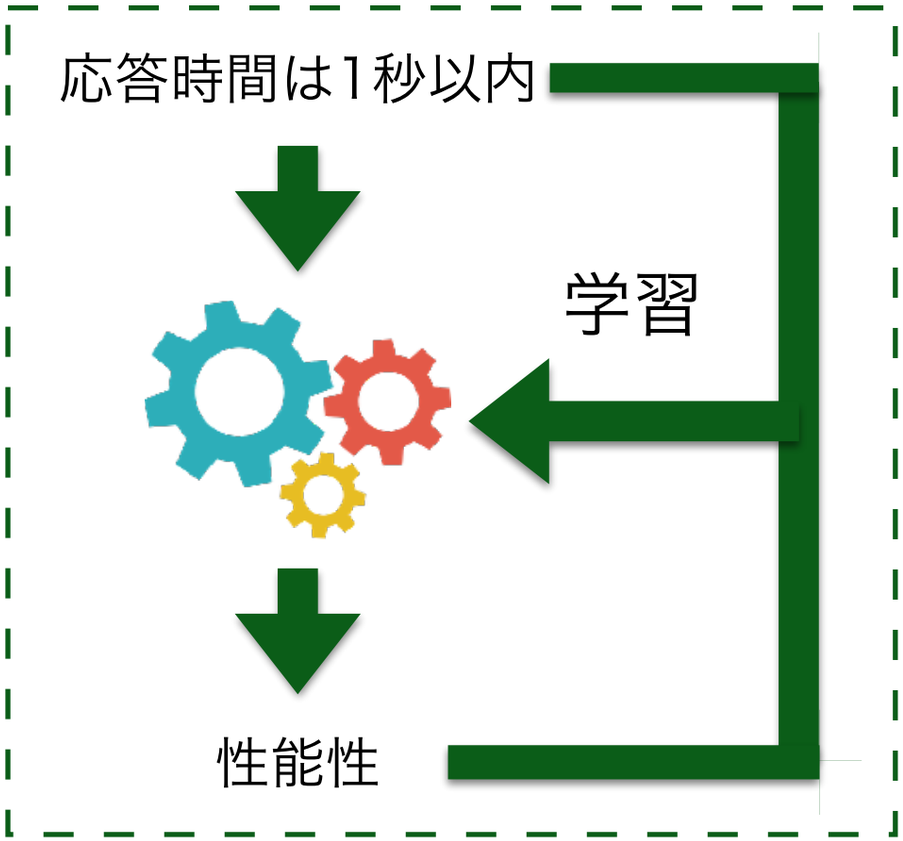
- ラベル有りデータとラベル無しデータを用いて学習を行う方法
  1. ラベル有りデータで分類器を作成
  2. 作成した分類器を用いてラベル無しデータを分類
  3. 分類結果を分類器が学習

1.

データの暗号化をすること	セキュリティ
応答時間は2秒以内	性能性
⋮	⋮

教師あり学習手法

2.3.



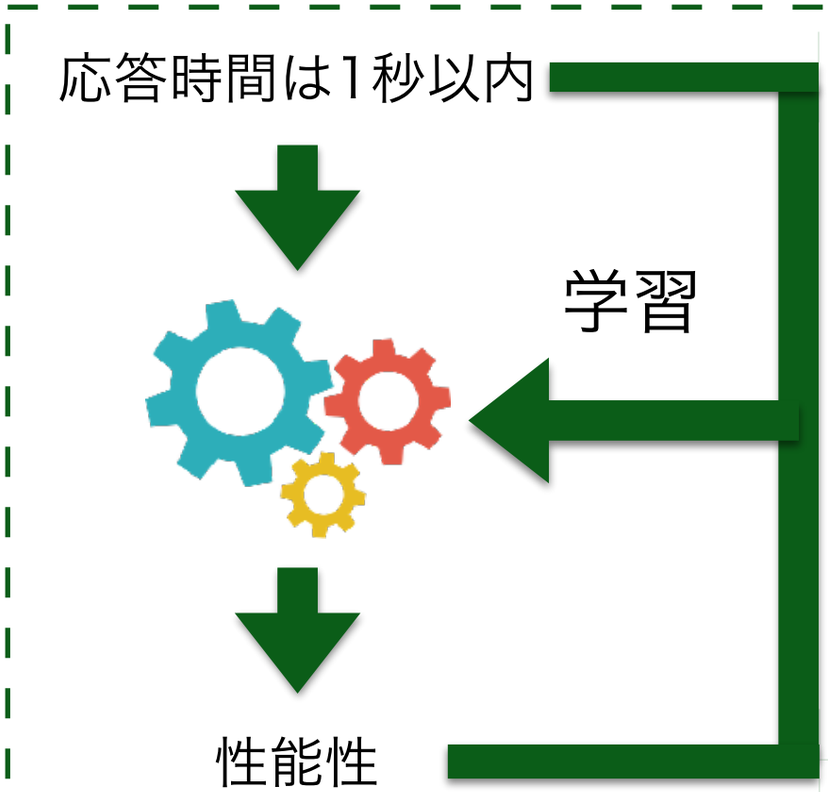
提案手法

1. ナイーブベイズを用いて分類器を作成

データの暗号化をすること	セキュリティ
応答時間は2秒以内	性能性
⋮	⋮

ナイーブベイズ

2. ラベル無しデータの学習にEMアルゴリズムを適用



前処理・分類器の作成

1. ストップワード（”is”,”the”といった高頻出の単語）の除去
2. ステミング処理
   - “done”→”do”, ”Apples”→”Apple”
3. tf-idfを適用しベクトル空間モデルを作成
4. ラベル有りデータをナイーブベイズに学習

$$\boxed{P(c_j|d_i;\widehat{\Theta})} = \frac{P(c_j|\widehat{\Theta})P(d_i|c_j;\widehat{\Theta})}{P(d_i|\widehat{\Theta})} = \frac{P(c_j|\widehat{\Theta})\prod_{k=1}^{|d_i|}P(w_{d_i,k}|c_j;\widehat{\Theta})}{\sum_{r=1}^{|C|}P(c_r|\widehat{\Theta})\prod_{k=1}^{|d_i|}P(\text{though } w_{d_i,k}|c_r;\widehat{\Theta})}$$

文 d_i が NFR c_j に属する確率

$P(c_j|\widehat{\Theta})$ NFR c_j の確率

$P(d_i|c_j;\widehat{\Theta})$ NFR c_j の時の文 d_i の確率

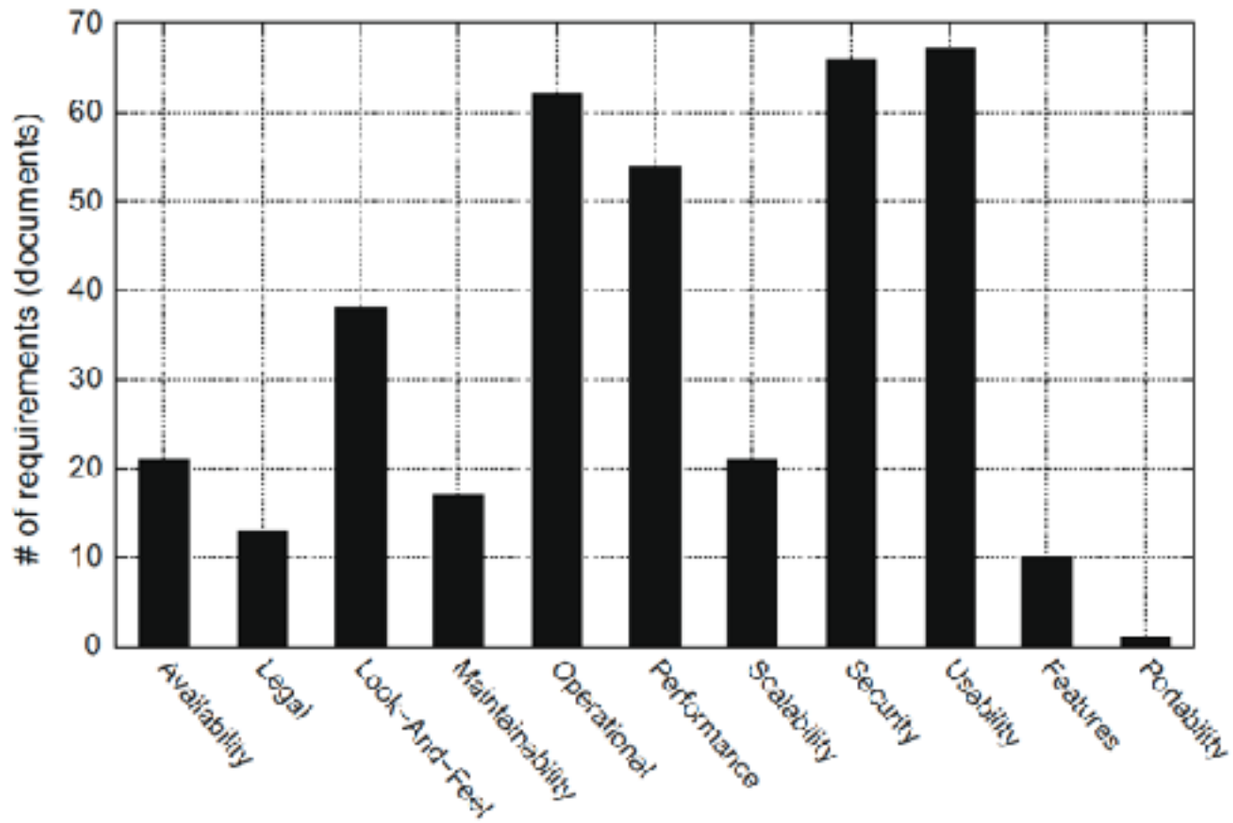
$\widehat{\Theta}$ 学習データの数

ラベル無しデータの学習

1. 作成した分類器にラベル無しデータを入力しNFRの確率を求める
2. 1の結果と既に学習したデータを再度学習して分類器を作成する
    1. ラベル有りデータについては$P(c_j|d_i;\widehat{\Theta})$の値を, c_jの時に1，j≠kの時に0として学習
    2. ラベル無しデータは1の結果をそのまま学習
3. さっきのページの式の各パラメータが安定するまで, 1.2の作業を繰り返す

評価実験

- PROMISEデータセット[2] のNFR370件・FR255件

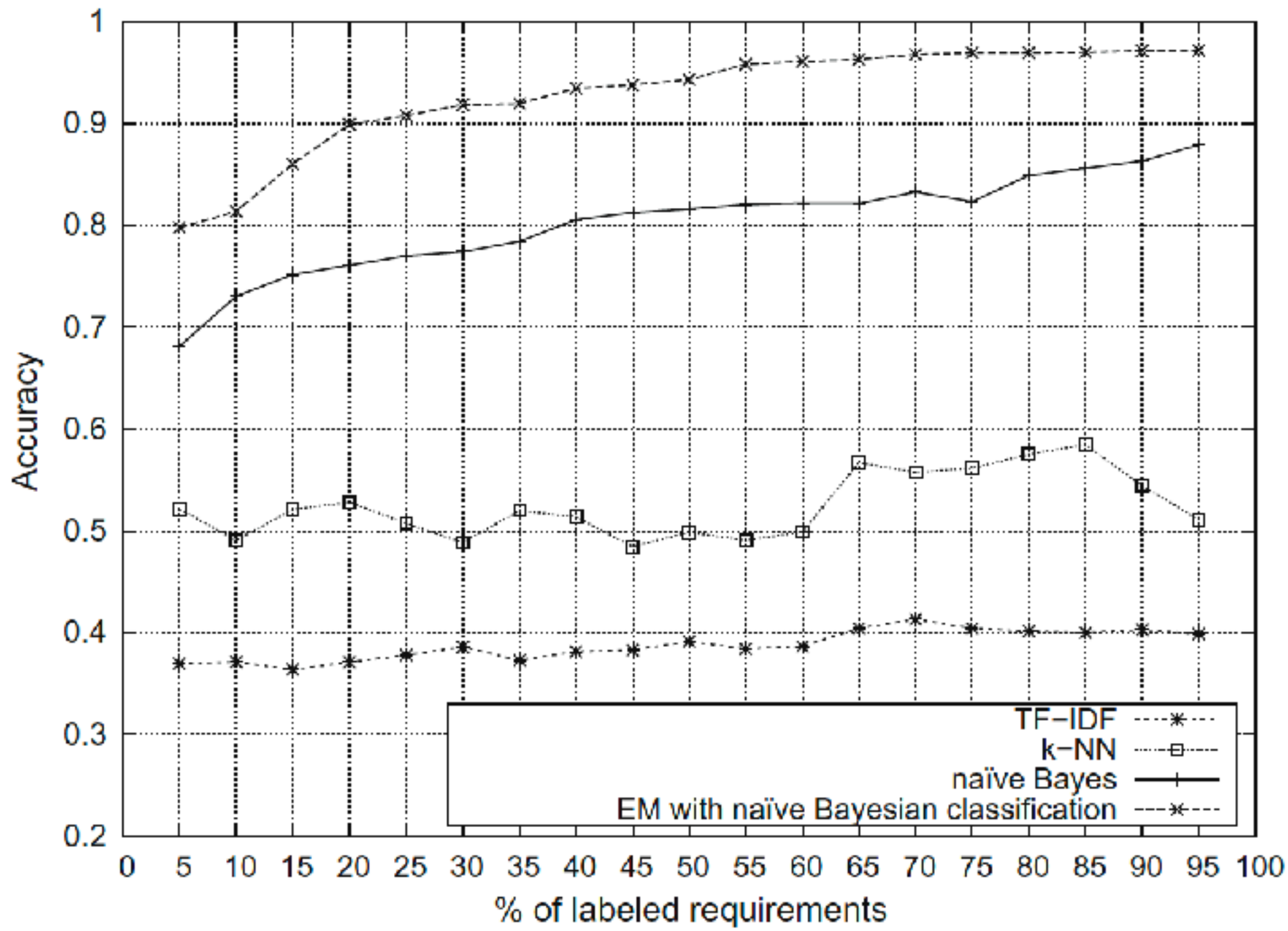


- 10交差検証実施：データの75%学習・25%テスト
  - 学習するラベル有りデータの数を変えて精度の変動も調査
  - Accuracy・Precision・Recall・F値で精度を評価

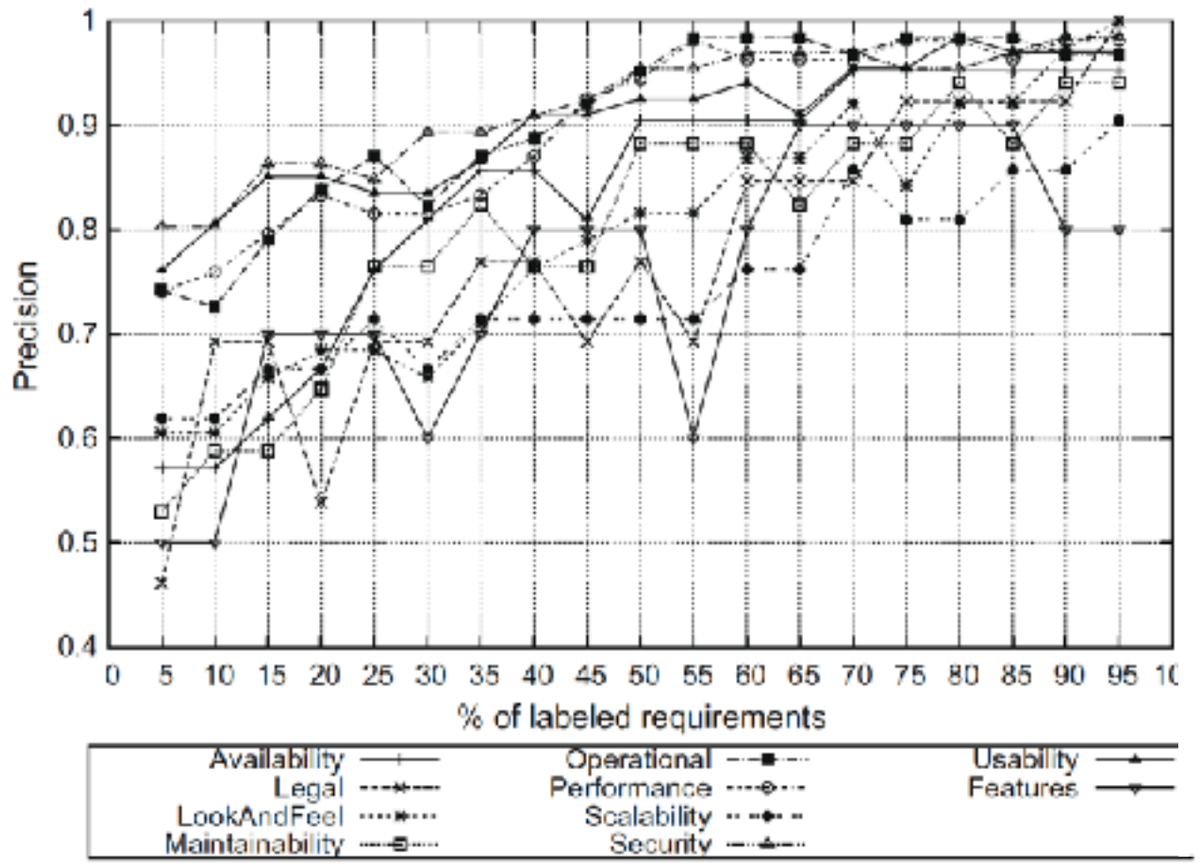
[2]http://promisedata.org/?p=38.

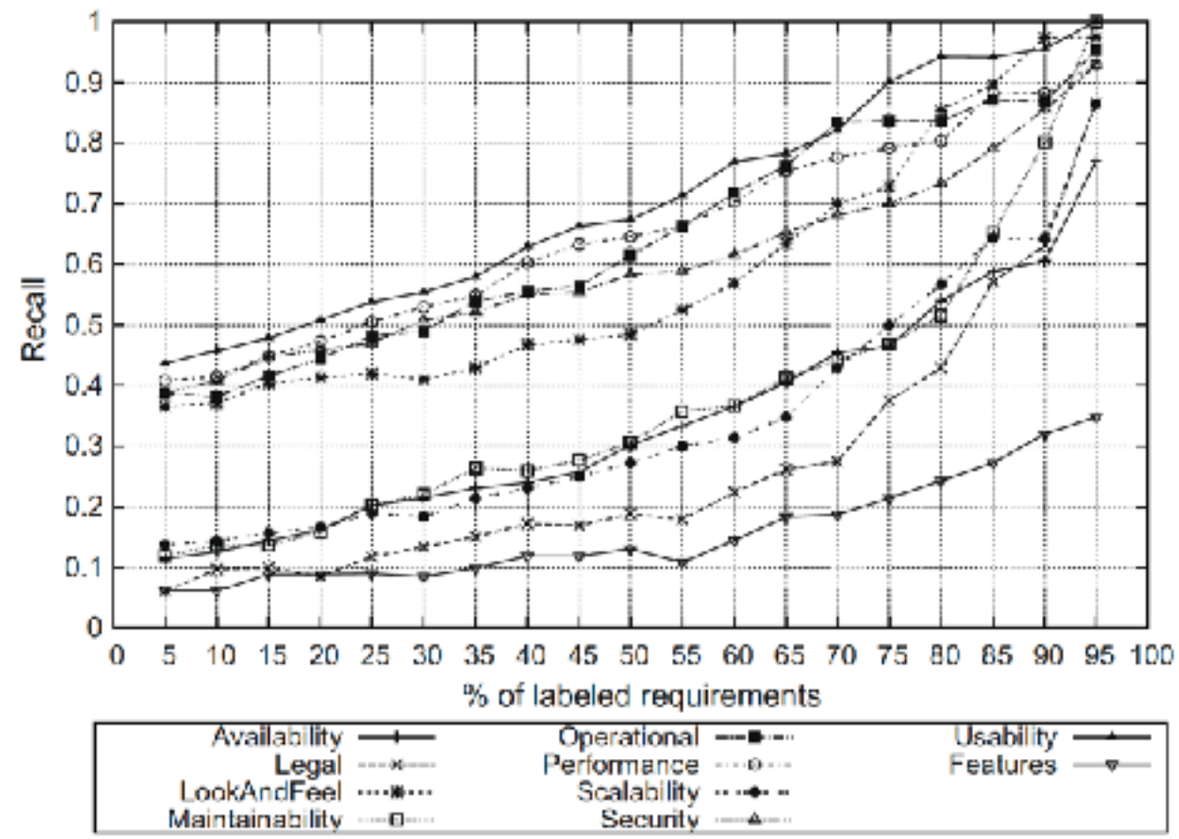
実験結果（全体）

- 提案手法の方が教師あり学習より精度が良かった



実験結果（各NFRごと）





- データ数の少ないNFRやFRは精度が低いが，それ以外は精度良い

まとめ・所感

- まとめ
  - ナイーブベイズとEMアルゴリズムを用いた半教師ありNFR特定手法を提案
    - 既存の教師あり学習の手法よりも精度が良かった
    - ラベル有り学習データが増えるにつれて精度も良くなる
- 所感
  - そんなに精度よくなる？
    - ラベル有りデータを増やしたら教師ありの方が精度が良いイメージ
    - 反共師あり学習の方が過学習になりにくいから？