

# Expertise Identification using Email Communication

Christopher S. Campbell

Paul P. Maglio

Alex Cozzi

Byron Dom

# 電子メール通信による専門知識の識別

Christopher S. Campbell

Paul P. Maglio

Alex Cozzi

Byron Dom

# 背景と問題

- ・ **背景**

- ・ 組織内の知識は構成する人の技術 経験 専門知識で構成されている
- ・ 誰が何に挑戦しているかを知ることは重要
- ・ メールは知識が蓄積している

- ・ **問題**

- ・ 組織の中で誰が何をしているかを知ることは困難

# 先行研究

- ・ 共通の関心を持つ人物のグループを特定する

Schwartz and Wood [10]

- ・ 著名人を出発点としてメールの流れを見る
- ・ 順位のない人物のリスト

- ・ メール単語の出現頻度から専門知識をスコアリングする

ContactFinder system [6]

- ・ 単にメールを多く送った人物が表示

# 目的

- ・ 目的

- ・ メールから特定のトピックに対するエキスパートを探し出す新しい手法を提案する
- ・ 送信者と受信者に基づいた手法

# アプローチ

1. トピックに関するメールを全て収集する
2. 専門家のグラフを作成するための関連性があった全ての人のメールを分析する
3. 分析グラフを分析して、どの送信者と受信者が**HITSアルゴリズム**によるスコアリングにより知識のある特定者を特定する

# HITSアルゴリズム

- **Hyperlink-Induced Topic Search**

- GoogleやYahoo!などの検索エンジンでも使われているアルゴリズム

- **Authority :**

- 重要な情報を発信しているページ
- 優良なAuthorityほど優良なから被リンクを多く受けている

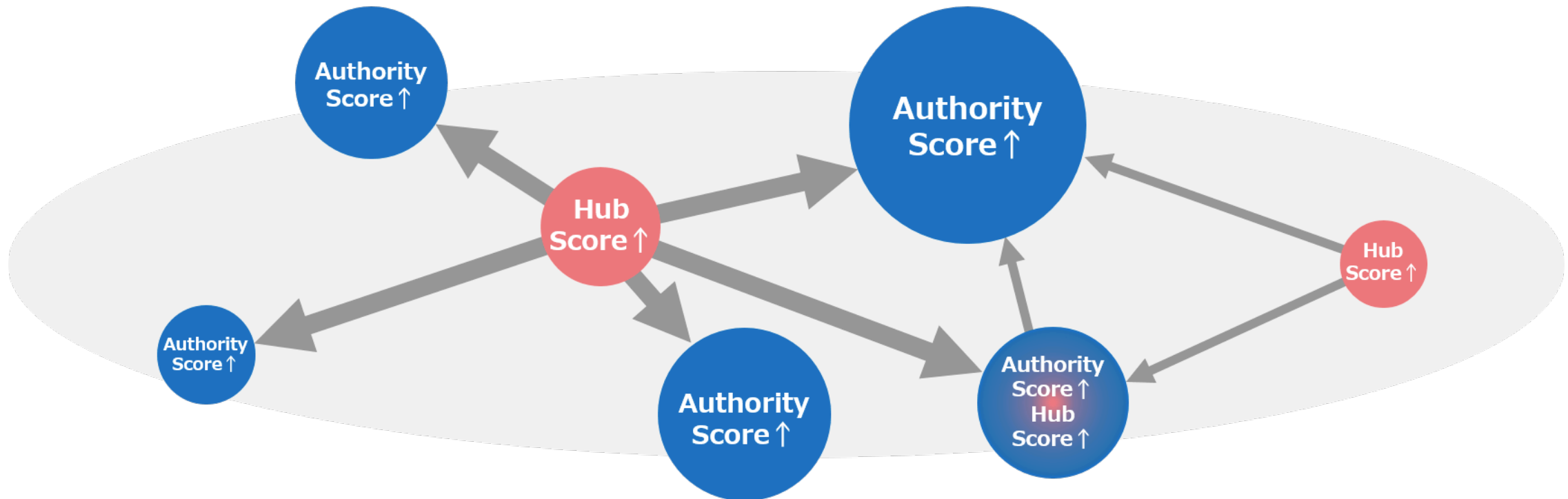
- **Hub :**

- 重要な情報を発信しているページに発リンクしているページ
- 優良なHubほど優良なAuthorityに発リンクしている

# HITSアルゴリズム

- ・ **Authorityスコア**
  - ・ そのページにリンクしている各ページのハブスコアの合計
- ・ **Hubスコア**
  - ・ そのページに リンクしている各ページのオーソリティスコアの合計

## こんな感じで入り組んでいるであろうイメージ図



# 評価方法

- ・ **メールのパターンによる情報（誰が送信して受信したか）が評価に役立つかどうか？**
  1. 単純なアルゴリズムHITSとの性能を比較する
  2. 各アルゴリズムを二つの異なる組織に適用
  3. 選ばれた候補者と多く交流のある評価者にアンケートを取る

# 分析対象

- ・ 対象

- ・ 研究機関：A 13.417通

- ・ ソフトウェア開発グループ：B 15.928通

- ・ トピックは各グループで行われた業務に基づいて、関連するメールを読むことにより決定した

# 結果：感度

- ・ 信号検出論の感度( $d'$ )により比較（情報検索の分野ではよく使われる指標??）
- ・ 感度：信号とノイズが分類される程度
- ・ 大抵の閾値でHTISの方が高い感度を持っていることがわかる

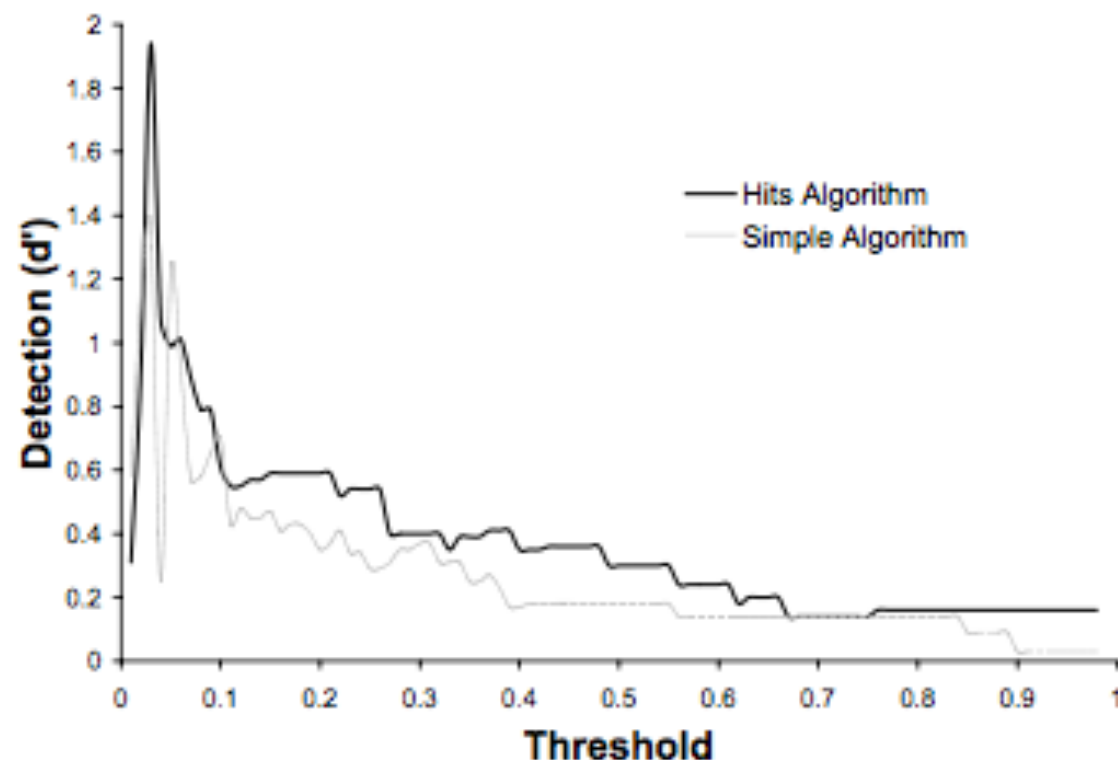


**Figure 2: Detection ( $d'$ ) is higher for HITS across most thresholds for OrgA.**

# 結果：正答率

- ・ 正答率は単純なアルゴリズムとHITSで差はほとんどない
- ・ 単純なアルゴリズムは誤報告が多い
- ・ そのためPrecision(精確さ)  $d'$  (感度) が単純なアルゴリズムは低くなっている

|                      | OrgA |        | OrgB |        |
|----------------------|------|--------|------|--------|
|                      | HITS | Simple | HITS | Simple |
| Percent Correct      | 38   | 44     | 33   | 31     |
| Percent False Alarms | 35   | 71     | 55   | 64     |
| Precision            | 52   | 38     | 67   | 50     |
| $d'$                 | .39  | .28    | .63  | .44    |

Table 1: Results for all topics and thresholds.

# まとめ

- ・ メール分析においてHITSアルゴリズムの有用性を証明した
- ・ HITSアルゴリズムは単純なアルゴリズムよりも精確である
- ・ 正答率には違いは少ない

# 所感

- ・ トピックの決め方には改良の余地が多い
- ・ トピックの決め方が実用的ではない（アルゴリズムの有用性を提案するためだからいいのかもしれない）
- ・  $HITS + \alpha$  でなにかできないか？
- ・ 古い論文なので最近のまでさかのぼる必要がある