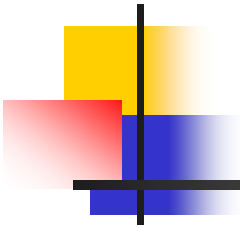
A2-2

地域情報検索のための リンク構造分析による ウェブページと地域の関係抽出

井上陽介* 李龍**
高倉弘喜*** 上林弥彦**

* 京都大学工学部情報学科
** 京都大学情報学研究科社会情報学専攻
*** 京都大学大型計算機センター



発表内容

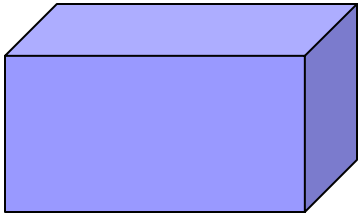
- 地域情報の特定の必要性
- ウェブページと地域との関係の2つの評価方法を提案
  - 地域での人気度
  - 地域指向性
- 提案する手法の評価実験

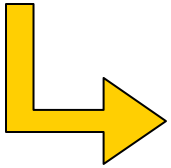
地域情報の特定の必要性

- インターネットの普及
  - グローバルな情報共有の実現
- 地域情報に対する需要は大きい
  - 地域情報検索における
    既存の検索システムの問題
- Web Mining
  - ウェブからの地域知識の発見

既存の検索システムの問題点

既存のウェブ検索エンジン





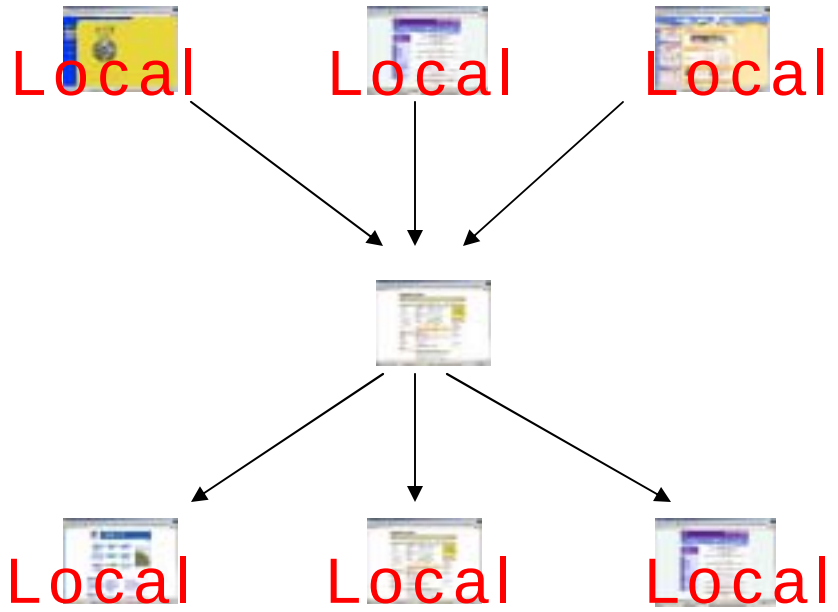
New York Times

New York City
の地域情報

地域の情報ではない

地域情報を持つかどうかの判定法が求められる

地域情報を含む重要なウェブページ



地域での人気度

ローカルなウェブページからリンクされている

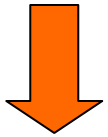
and

ローカルなウェブページにリンクしている

地域指向性

ウェブページの
ある特定の地域での人気度

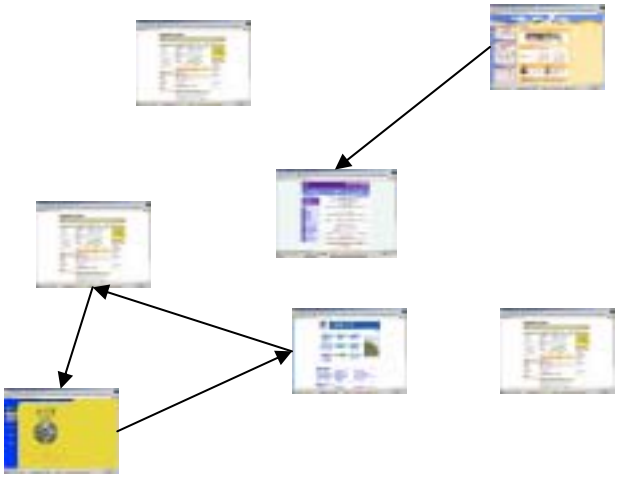
- ある特定の地域での人気度
  - 地域の中で重要なページから
    リンクされているページは
    地域の中で重要なページである
- GoogleのPageRankの手法を適用
  - リンク構造によるランキングアルゴリズム



ある地域の地名を含むウェブページ集合に適用

ウェブページの地域での人気度と部分集合の拡張

地域の地名を含むページによる部分集合に既存のランキングアルゴリズムを適用



とってきた部分集合

リンク構造が不十分

対象地域を拡大

リンク先を部分集合に追加

ウェブページの地域指向性

- リンクをたどってアクセスできる地域情報が多いページほど地域指向性が高い
- ユーザの行動を次のようにモデル化
  - 対象地域の地名を**含む**ページなら、ランダムにリンクを選択し、次のページを参照
  - 対象地域の地名を**含まない**ページなら、ブラウズを終了
- 各ページを、そのページから始めたときの平均訪問ページ数でランキング

実験方法

- 地域での人気度の評価
  - 既存の人気度計算法によるランキング
  - 地域での人気度によるランキング
  - 地域･リンクで拡張した場合の地域での人気度によるランキング

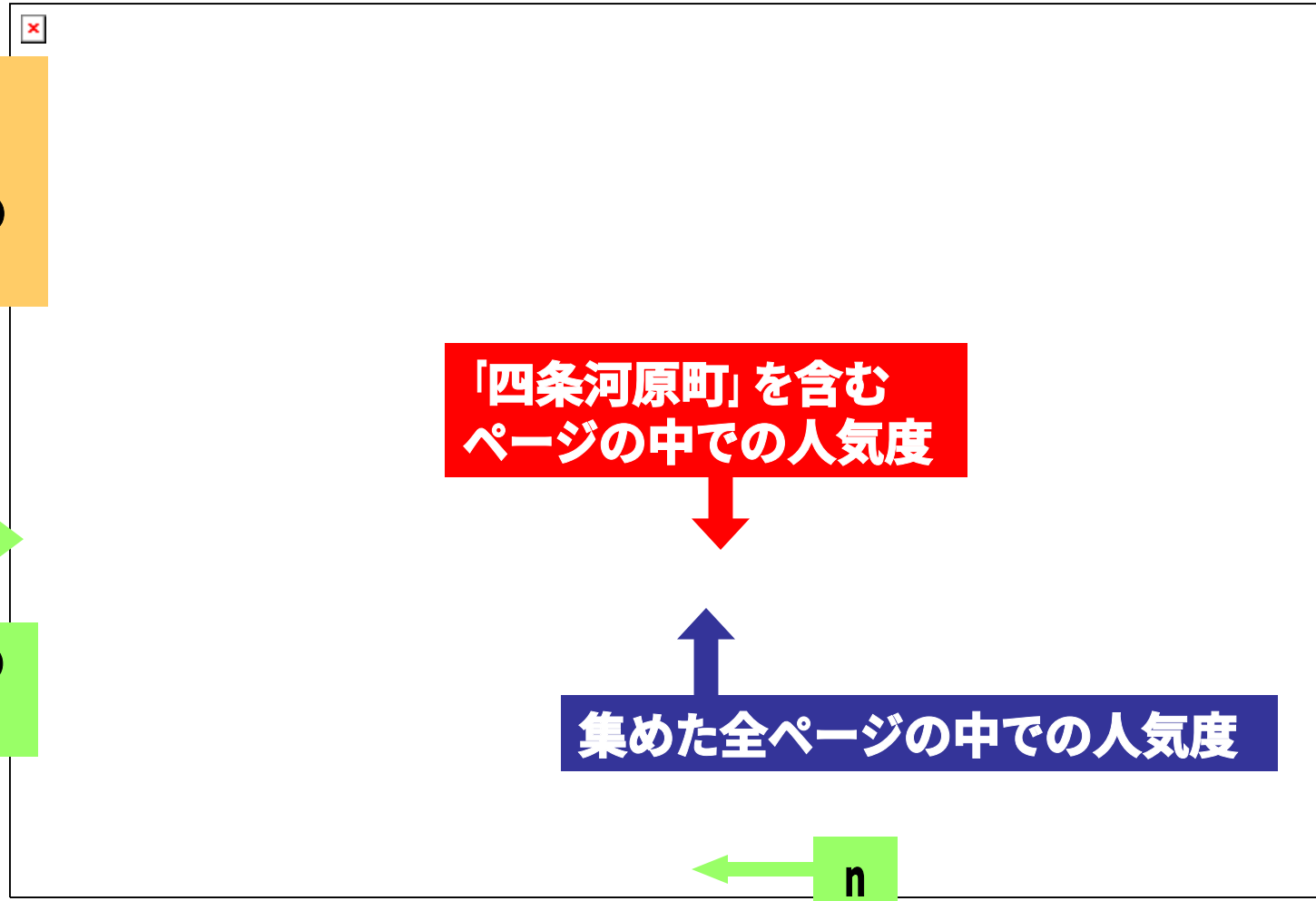
比較

- 地域指向性の評価
- 両者を組み合わせた場合の効果

実験：評価方法

- 著者の主観による評価結果と比較
- 比較対照データ
  - 著者の主観により各ウェブページを
    0～5の6段階にランク付けする
- 各手法によるランキング結果の
  上位nページの平均点により、結果を比較

実験：
地域での人気度によるランキング



実験：
対象地域を拡張した場合のランキング

適切な範囲で地域を拡大することで
より地域に密着した情報を得られる

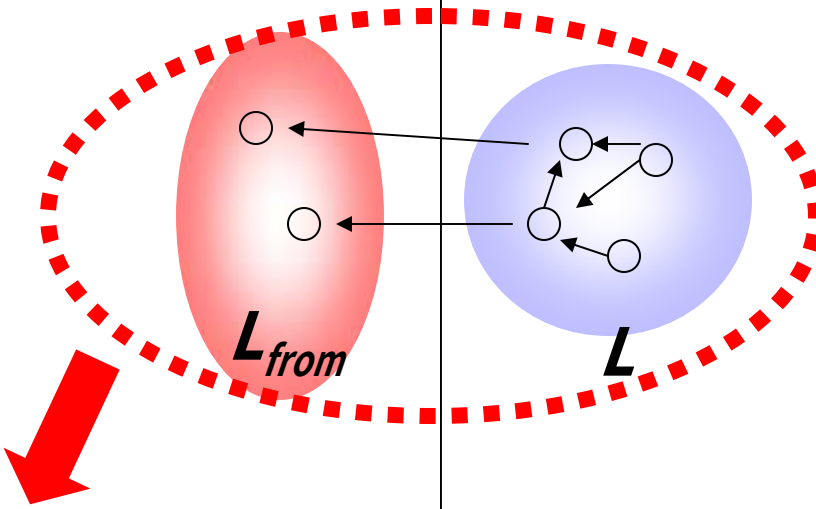
四条河原町から1km以内の地名を
含むページ集合での人気度

四条河原町から2km以内の地名を
含むページ集合での人気度

「四条河原町」を含む
ページの中での人気度

実験：
リンクで拡張した場合のランキング

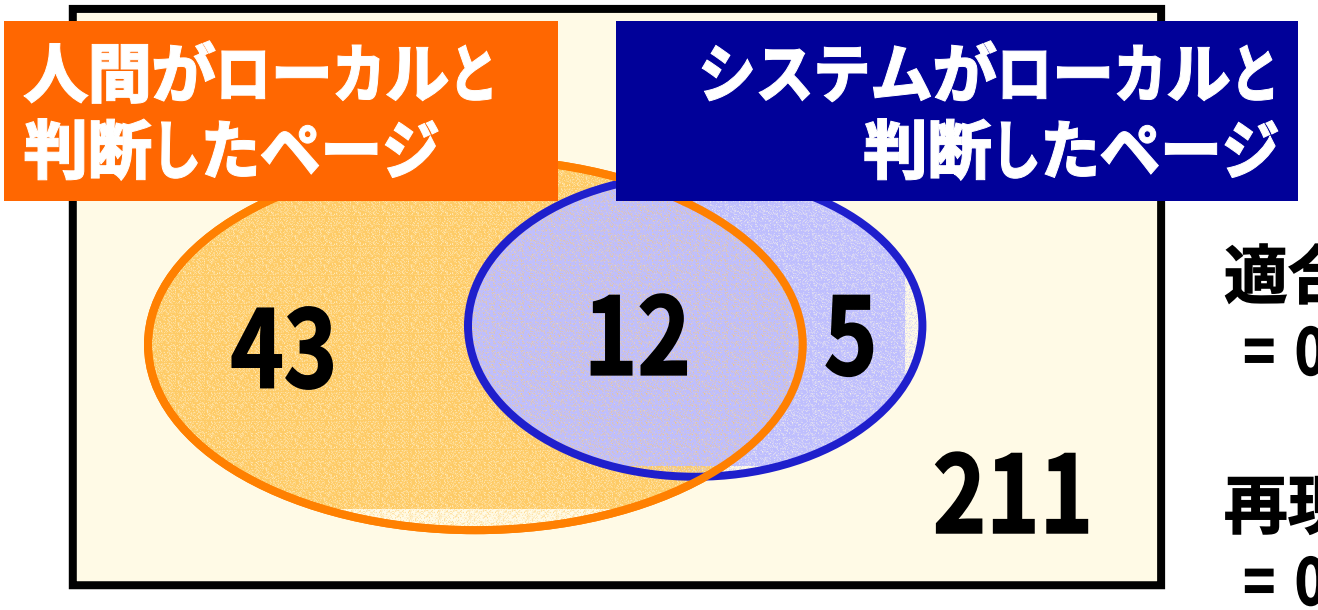
リンク先を加えることで
やや結果がよくなる



もとのページ集合からリンクされているページを対象に追加した場合

実験：地域指向性による ローカルなページの判断

地域指向性が一定の閾値以上のページをローカルと判断し、人間の主観と比較

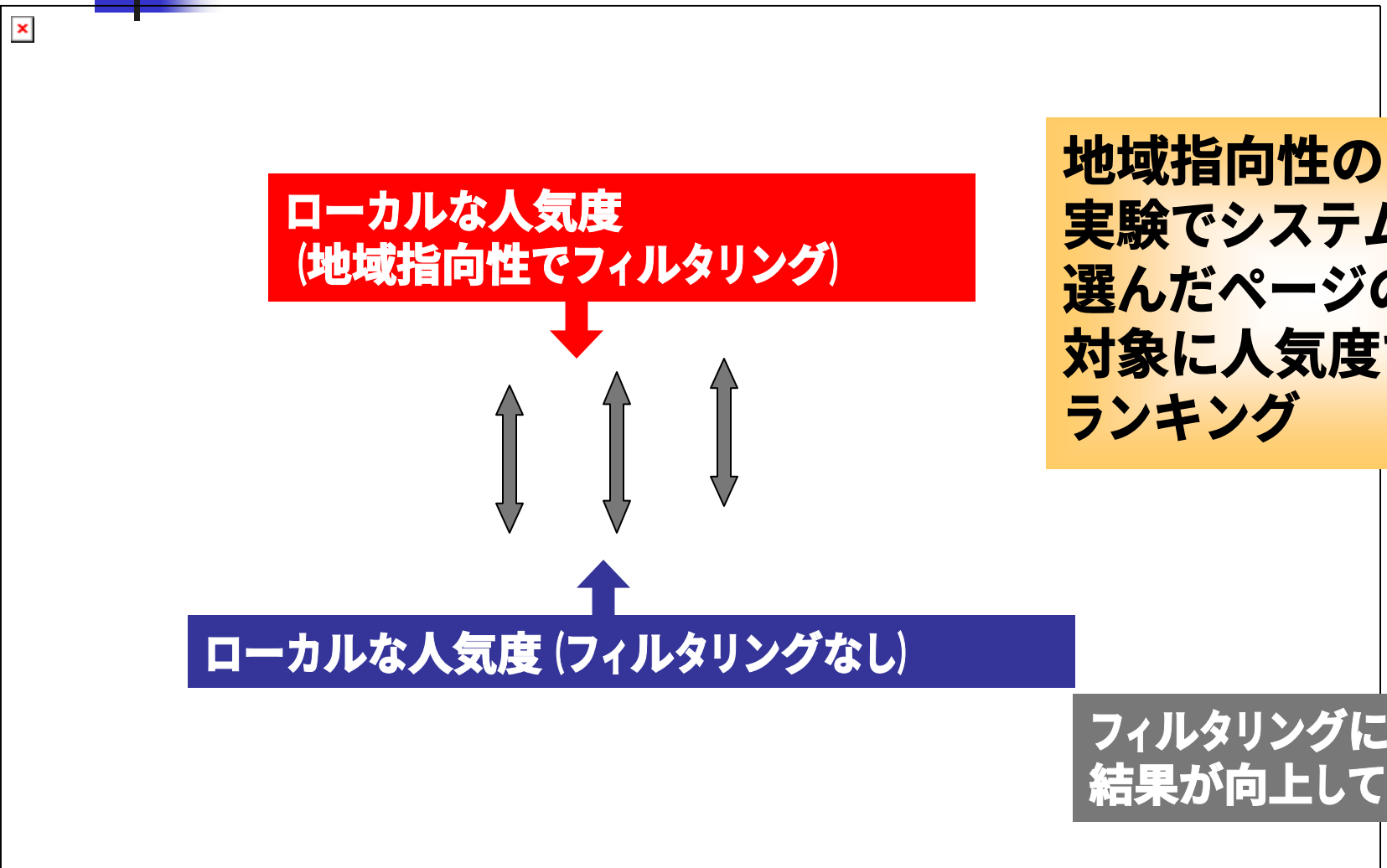


適合率
= 0.706 (12 / 17)

再現率
= 0.218 (12 / 55)

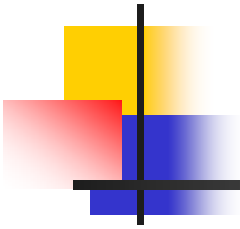
「銀閣寺」に関するウェブページでの実験結果

実験：地域指向性によるフィルタリングと地域での人気度によるランキング



地域指向性の
実験でシステムの
選んだページのみを
対象に人気度で
ランキング

フィルタリングにより
結果が向上している



まとめ

- ウェブページと地域の関係を測る２つの尺度の提案
  - 地域での人気度
  - 地域指向性
- ウェブページの部分集合の拡張法の提案
  - 地域による拡張
  - リンクによる拡張
- 実験による提案手法の有用性の説明

今後の課題

- 再計算が必要
  - 再計算のいらない近似法の開発

- 今後は両者をより密接に融合し、
  より効果的な地域情報検索の実現を目指す
  - 人気度計算におけるリンクの遷移確率に
    地域指向性を利用