WEB検索における人名の曖昧性解消技術の動向

関根　聡（ニューヨーク大学）

発表者：サ　ミンソン

なぜ人名の曖昧性解消が重要か

- Googleで山口百恵を検索・・・
  ー元歌手の山口百恵さんが大量に表示
  ーフットサル選手の山口百恵さんは12位

・佐々木元で検索した場合・・・
  ー上位10件は、NECの会長、東北大学の教授、プロのBMXライダー、翻訳者の4名
  ー60年代の映画監督の人は82位

つづき

- 検索結果の各場面で、ユーザが探す対象と検索結果の順番には常に相関関係があるわけではない　→　非常に不便

- この問題を解くために、同姓同名を人物ごとに振り分ける技術が必要になってくる

検索結果の要約（スニペット）

- 山口百恵の検索結果のスニペットに「女優」、「元歌手」、「コンサート」などの単語が見つかれば、検索ユーザは歌手の山口百恵と気づく

- 上記のようなキーワードをテキストの中で認識し、各テキストの特徴を得るという方法が考えられる

評価プロジェクトWePS

- ある人名で検索されたWeb検索結果の100ドキュメントまでのセットを渡され、その中に含まれるページを人物ごとにクラスタリングする
- データの作る際の人名は3つの違った種類のリソースからランダムにサンプリング

①国勢調査に挙がった人名

②Wikipediaの項目に挙がっている人名

③学会発表者に挙がっている人名

つづき

- Webページは、YahooのAPIを利用してそれぞれの人名に対して、100ページを上限として収集

トレーニングデータ		
リソース	人物数	ドキュメント数
wikipedia	23.14	99.00
ECDL06	15.30	99.20
WEB03	5.90	47.20
平均	10.76	71.20
テストデータ		
wikipedia	56.50	99.30
ACL06	31.00	98.40
国勢調査	50.30	99.10
平均	45.93	98.93

参加システムの技術

- 各システムは、HTMLのドキュメントからテキストを抽出
- 抽出したテキストに対してWeb解析ツールを使った前処理を行う
- ドキュメント間の類似度を計算
- 類似度を元にクラスタリングを行う

評価

- 世界中から16チームが参加
- 結果は、purity(各クラスタにおいて最も多い正解のラベルの割合の加重平均)とInverse purity(各正解のクラスタにおいて、最も多く含まれたクラスタに属する正解クラスタの要素数の割合の加重平均)に基づいたF値で評価された。
- F値が最高であったチームは0.75の値
- 人間の評価は0.98の値

問題点

- トレーニングデータとテストデータの平均人物数の間に大きな乖離があり、トレーニングデータを利用して決定する閾値にはこの乖離が問題となる

- 例えば、クラスタ数がトレーニングデータの平均の11になったらクラスタリングを終了するという単純なシステムには大きな問題

考察　－人物の属性－

- 曖昧性の解消のために、人物の属性を認識することが重要（ex:職業名、作品など）
- WePSで使用されたドキュメントの156個から人物の属性と考えられるもの16種類を選んだ

今後の展開

- 2回目の評価プロジェクトで人物の属性抽出のサブタスクを行い、各ページから属性をどのくらい抽出できるかで評価を行う

- また、このタスクに対しては、固有表現抽出、テキストマイニング、パターンマッチング、情報抽出などさまざまな技術が展開され利用されることが期待できる

中間発表

- 韓国の有名人の同姓同名をGoogleで検索し、上位300件の検索結果のスニペットをデータとする
- 検索した名前：박명수

A(芸能人)	264	88%
B(プロゲーマー)	28	9.3%
スニペットではわからない	8	2.6%

one class svm

- Support Vector Machine(SVM)を用いて外れ値検出を行う。
- データを多次元空間内の点とみなしたときに、データ集合の領域を求め、その領域に入っていないデータを外れ値と見なす。

つづき

- One Class SVM では，学習データが入る小さな領域内で+1をとり，それ以外の領域で−1をとるような関数fを，以下の考えによって生成

①学習データを原点からの最大マージンによって分類

②RBFカーネルを使用すると存在確率の低いデータが原点の近くにくる特性を利用

つづき

- 特徴空間上の学習データを原点から最大マージンによって分類するため，以下の2 次計画問題を解く

$$\min_{w\in F, E\in R^l, p\in R} \quad \frac{1}{2}\|w\|^2 + \frac{1}{vl}\sum_i E_i - p$$

$$subject\ to \quad (w\cdot\theta(x_i)) \geq p - E_i, \quad E_i \geq 0$$

- この2 次計画法を解くことで，Structural Risk を最小化するような分類超平面を見つけることができる.
つまり，正常状態データと正常状態でないデータ（外れ値）を分ける超平面を見つけることができる.