

テキストマイニング(2)～言葉と数値の関係を探る

背景: テキスト中の数値

「入場料: 大人2000円、子供1000円」
 「社長に就任。富山県出身、58歳」
 「肺炎で死去、78歳」

動機2: 数値と言語の(共起)関係を捉えたい

年齢と呼称	典型的な値段	位置とイベント
3歳→男児(女児)	「コーヒー」→200円	1,000ft→「離陸」
20歳→若者	「サラダ」→500円	8,000ft→「落雷」
45歳→中年男性	「ランチ」→1000円	12,000ft→「揺れ」
80歳→老人	「ディナー」→3000円	25,000ft→「巡航」

動機1: 文字列としてでなく、数値として検索(範囲指定による検索等)したい

提案: 言語も数値も区別なく検索できるシステム

Query=「入場料は*」
 Query=「3～6歳の*」

数値の範囲を一つの語彙のように使える

手法: Number Suffix Array

Suffix Array: すべての部分文字列のソートされたリスト

犬だった。そのことは...
 犬との楽しい日々を...
 犬ではなく、...
 犬は嫌いです。でもドッグフードは...
 犬も歩けば棒に当たるというのは、...
 犬も歩けば棒に当たる！
 犬も猫も大好き！
 犬猫病院

構築

各数字の先頭に0を入れ、桁を揃える

「入場料: 大人00002000円、子供00001000円」
 「社長に就任。富山県出身、00000058歳」
 「肺炎で死去、00000078歳」

その後、普通にSuffix Arrayを構築する
 (ただし、数値の途中はインデックスしない)

数値範囲検索

数値範囲の最小値、最大値それぞれで検索
 →その間にあるインデックスをすべて取得

Query=大人500～30000円

大人0000500円...
 大人00001000円...
 大人00001800人...
 ...
 大人00030000円...

問題点: 数値の右にある文字列は、整列していない
 ⇒対策: 検索のたびにソートする

問題点: Query先頭が数値のとき、検索件数が膨大に
 ⇒対策1: 数値以外を先に検索し、ソートする
 →その後先頭の数値を検索
 ⇒対策2: 検索結果を記憶しておく

手法: 検索結果のクラスタリング

検索結果中の数値は、「最小値～最大値」の形でまとめる
 問題点: 数値コレクションの表現として著しく不適切な場合

1,2,3,4,1000,1001,1002→1～1002
 92年,94年,1993年,1999年→92～1999年

1,2,3,4,1000,1001,1002→1～4, 1000～1002
 92年,94年,1993年,1999年→92～94年, 1993～1999年

数値の動的クラスタリング

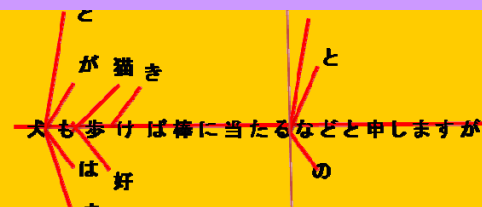
数値クラスタを、Dirichlet Process混合正規分布によりモデル化

$$\frac{\alpha^{|C|}}{\alpha^{(n)}} \frac{1}{\sqrt{2\pi} \sigma_0^2} \prod_{j=1}^{|C|} \frac{1}{(|C_j| - 1)!} \frac{1}{\sqrt{1 + |C_j| \left(\frac{\sigma_1}{\sigma_0}\right)^2}} \exp \left\{ -\frac{1}{2\sigma_0^2} \left\{ \sum_{i=1}^{|C_j|} x_{e_{j,i}}^2 - \sigma_1^2 + |C_j| \sigma_1^2 \left(\sum_{i=1}^{|C_j|} x_{e_{j,i}} \right)^2 \right\} \right\}$$

「クラスタリングの良さ」を、クラスタ数の違うクラスタリング同士で比較できる
 ⇒「良いクラスタリング」を、Greedy Algorithmにより求める
 (分割なしの状態から始めて、最適な分割を繰り返す)

応用: 用例検索と同義語抽出

Kiwi: よく使われる言い回しの提示システム



Query=「犬も歩けば*」→「棒に当たる」

Number-Kiwi

通常のKiwiに数値の取扱いを追加

Query=「入場料は*」→「大人1000～1500円、子供500～800円」
 Query=「3～6歳の*」→「男児|女児|児童|...」

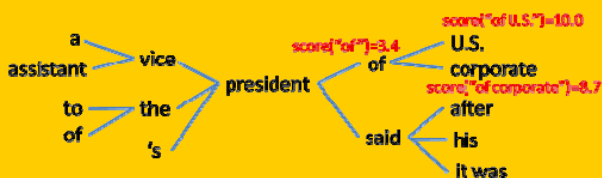
Number-Synonym

数値を入れると、それに類似する言葉を返す

10～20歳⇒[歳, 年, 中学生, 力月, 高校生, 小学生, ...]
 70～90歳⇒[歳, 人, 中年, 齢, 寝たきり, 年配, , 此, 被害者, 中学生, 痴ほう, ...]
 3～8歳⇒[歳の, 生の, 「, その, ,, この, 幼い, 人の, 年の, 若い,]

同義語抽出

「文脈をSuffix Arrayで検索することによる動的な同義語抽出



同義語抽出の精度向上

数値を文脈として使う

例: 「高度8000ftにて落雷」
 「高度7500ftで落雷」...

いままで:
 「落雷」の文脈=「高度:10, 8000ft:2, 7500ft:1, ...」
 (または、すべて0に正規化)

「落雷」の文脈=「高度:10, 3000～15000ft:20, ...」