6J-4

情報爆発時代のための制約つきクラスタリングを用いた制約つきフィードバック手法の提案

Relevance feedback using constrained clustering for Information Explosion Era

辻下 卓見† 相澤 彰子‡ 高須 淳宏‡ 安達 淳‡

Takumi Tsujishita Akiko Aizawa Takasu Atsuhiro Jun Adachi

東京大学 大学院 / The University of Tokyo†

国立情報学研究所 / National Institute of Informatics‡

1. はじめに

World Wide Web(WWW) 上に流通する膨大な情報の中から目的を持って情報を得ようとする時、多くのユーザはキーワード検索型の Web サービスを利用する。しかし望むものが検索結果の上位に上らない場合も多い。その場合、ユーザは何らかの手段で検索結果の絞り込みを行う必要がある。現在使われている Web 検索サービスでは絞り込みはユーザによるキーワードの追加で行われるが、適切な結果を得られるキーワードを入力できるかどうかはユーザの検索能力によるところが多い。

こういった問題を解決するための手法は多数提案されているが、そのなかの一つがユーザとの対話を通じて、ユーザの嗜好を読み取り、検索結果を改善するやり方である。そのような手法の一つとして、検索で得られた結果から適合する文章と適合しない文章をユーザから入力してもらうことで検索結果を改善する適合性フィードバックがある。

適合性フィードバックは、検索結果が適合するかしないかを人手や自動で判定した結果を利用して検索結果を改善する手法である。しかし少数の例をのぞけば、現在のところ適合性フィードバックを使った一般的な検索サービスは普及していない。それは、良い結果が得られるのは多くの適合性判定情報を得たられているときであり、通常の検索においてはユーザから十分な量の適合性判定情報を得られることはまれであるからである。

そこで本研究では、近年研究の進んでいる制約付きクラスタリングを用いて適合性フィードバックを行う手法を提案する。

2. 制約つきクラスタリング

クラスタリングはデータを類似性に基づいてグループ化させるデータマイニング手法の一つである。一般的に用いられる教師付き機械学習の手法とは異なり、分類先のラベルや学習データをあらかじめ用意する必要がない非教師付機械学習である。しかし一方でクラスタリングは望ましいクラスタの基準を形式的に与えることが困難であるため、背景知識を使う様々なやり方が提案されてきている。

制約付きクラスタリングは背景知識を使う手法の一つで、あらかじめデータにいくつかの制約を与えてクラスタリングする手法である。提案手法ではいくつかある制約付きクラスタリングのうち、Constrained K-means という手法を用いる。

Constrained K-means は K-means を、制約付きで行えるように拡張した手法である。K-means は K 個の初期クラスタの中心をランダムに選び、その後以下のように中心を更新することで最終的なクラスタを得る。

1. すべてのデータと K 個のクラスタの中心との距離を計算し、最も近いクラスタに割り当てる。
2. 新たに形成されたクラスタの中心を計算する。

このステップはクラスタの中心が動かなくなるか、規定の回数に達するまで繰り返される。

Constrained K-means[1]は K-menas のすべてのデータをクラスタに割り当てる Step1 において、与えられた制約をチェックして制約を満たすようにデータを分類する。その際、Constrained K-means では分類対象の文章の組に対して以下の二つの制約を考慮する。

1. Must-link ：二つの文書は同じクラスタに分類されなければならない
2. Cannnot-link ： 二つの文書は同じくクラスタに分類されてはならない

ユーザから与えられた制約条件の集合がクラスタリングアルゴリズムに渡される。制約付クラスタリングはデータの集合とそれに対する制約条件を使ってクラスタリングする。

3. 提案手法

提案手法では次のような手順で検索結果を改善する。

1. Step 1:ユーザが入力した質問を既存の検索エンジンに投げた結果上位 N 件を表示する。
2. Step 2:検索結果 N 件のうち適合文章と不適合文書を人手でいくつか判定する。判定した結果から must-link と cannnot-link という二種類の制約条件を生成する。
3. Step 3:得られた制約条件を付けて、検索結果の上位 N 件を制約付きクラスタリングという手法で分類する。その結果として、適合文書クラスタ、不適合文書クラスタ集合、その他の集合が得られる。
4. Step 4:適合文書クラスタ、不適合文書クラスタ集合を用いて、通常の検索質問の修正手法にかけることで、新たな質問が得られる。その質問を用いて検索し直すことでよりよい検索結果が得られる。

以上を検索要求が満たされるまで繰り返す。

3.1. 制約の付与
適合性判定情報として、検索結果の文章のうち適合する文章と、適合しない文書それぞれが人手で与えられる。提案手法では、適合文章同士をすべて must-link でつなぎ、適合文書と不適合文書を cannot-link でつなぐ。得られた制約を制約付きクラスタリングに必要な制約条件とする。またクラスタリングして得られたクラスタ集合のうち、適合文章を含むものについては適合クラスタ、不適合文書を服喪のについては不適合クラスタとする。また、どちらも含まなクラスタは不明クラスタとして、適合性フィードバックには用いない。

3.2. クラスタリング
クラスタリングするために、文書をそれぞれ単語ベクトルに変換する。単語ベクトルは単語とその重みによって記述される。
提案手法では単語の重み付けには、BM25 TFIDF の式を用いる。BM25 TFIDF は文章検索で用いられる重み付け手法であり、文書 d における単語 t の重みは以下の式で与えられる。

$$w(t,d)=\log\frac{N}{df}\cdot\frac{(k+1)\cdot tf}{k\cdot\{(1-\alpha)+\alpha\cdot dl/avdl\}+tf}$$

ここで，w(t,d)は文書 d における単語 t の重み，k と α は定数，N は文書コレクションに含まれる文書の総数，tf は d における t の出現頻度，df は全文書における t を含む文書数，dl は d の長さ，avdl は文書の長さの平均値を表す.

3.3. 検索結果の修正
適合性フィードバックには様々な種類があるが、検索質問の修正はその中でも頻繁に使われるものの一つであり、いろいろな手法が提案されている。
提案手法ではよい結果が得られるという研究結果のある Roccho の手法を用いる。
ベクトル空間モデルの適合性フィードバックにおいて検索語の重みを修正する手法はいくつか提案されているが、どれも次式が基本となる。

$$Q_{i+1}=Q_i+\alpha\sum_{j=1}^{N^+}D_j^+-\beta\sum_{j=1}^{N^-}D_j^-$$

ここで、Q_i は i 回目の質問に対応する検索語の重みベクトルである。また、N^+ と N^- はそれぞれ、i 回目の検索結果における適合文書、不適合文書の数、D_j^+ と D_j^- はそれぞれ個々の適合文書不適合文書に対応する索引語の重みベクトルである。α と β はそれぞれ定数である。
Rocchio の式では α を $1/N^+$ と、β を $1/N^+$ とおく。すなわち、重みの調整分を適合文書、不適合文書で正規化している。

4. 評価
提案手法の評価のために、実際のデータを使って、実験を行った。実験の概要は次のとおりである。
初期検索の検索結果の上位五件に対して、正解データを参照して適合不適合の判定を行う。与えた適合性判定情報を基にして、適合性フィードバックを一度行い、性能を評価する。
比較対象として、Rocchio の手法を用いた。
実験にはテストコレクションとして TREC の TIPSTER 1 を、検索タスクに TREC-1 ad hoc & TREC-2 routing topics の 50 件を用いて行った。テストコレクションは事前に Poter stemmer を用いてステミングした。
実験は、初期検索の結果上位十件に適合文書が多いもの、少ないものをそれぞれ、簡単な検索タスク、難しい検索タスクと考え、別々に評価を行った。
評価は検索結果のランキング上位Ｎ件における精度を用いている。

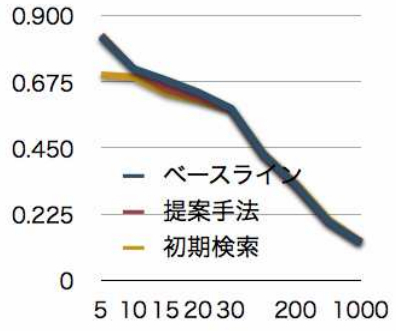

図 1 簡単な質問の結果

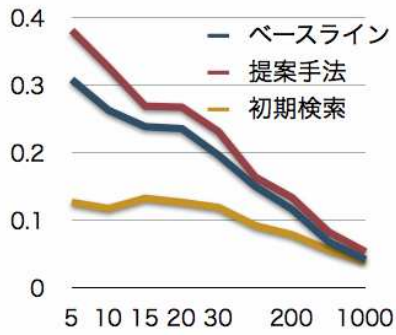

図 2 難しい質問の結果

グラフの縦軸は精度、横軸はＮを表す。
図 1 から、簡単な質問タスクの場合、提案手法ではあまり改善が見られないことがわかる。しかし、ベースラインもまた初期検索から向上が見られないため、初期検索の時点ですでに十分な性能が出ていると考えられる。
一方図 2 から、難しい質問の場合、提案手法はベースラインに比べ上位で 10%程度、全体で 8%程度の性能向上が見られ、難しい質問での有効性が認められた。

5. おわりに
本稿では、制約つきクラスタリングを用いて適合性フィードバックの精度を向上させる手法について提案し、テストコレクションを用いて手法の有効性を確認した。

参考文献
[1] Kiri Wagsta, Claire Cardie, Seth Rogers, and Stefan Schroedl. Constrained k-means clustering with background knowledge. In Proceedings of 18th International Conference on Machine Learning (ICML-01), pp. 577− 584,2001