

検索履歴を用いた Web 検索支援システム

竹内 星詞 (指導教員: 佐川 雄二)

名城大学 理工学部

1. はじめに

近年, インターネットの普及に伴い, 膨大な情報が Web 上に集まっている. それらの中には当然, 必ずしもユーザにとって有益な情報のみではなく, あるユーザにとっては興味のない情報も多くある.

現在の検索サービスでは, ユーザが検索要求を行うと, 膨大なデータベース中の各文書に対してクエリ (検索キーワードや検索式) との適合度を計算し, その値が閾値よりも大きい文書のみを適合度の降順に並べた文書のリストとしてユーザに提示している. そのため, ユーザにとっては知りたい情報が含まれている文書がランキングの下位に配置される場合がある. これは, ユーザによって知りたい情報は異なるが, キーワードのような検索方法では, それを十分あらわすことができないためである.

本研究では, 検索結果データベース中の文書に対してクラスタリングとラベリングを行うことによりユーザ情報をフィードバックすることで検索の精度・効率を向上させることを目指す.

2. システムの概要

以下にシステムの処理手順の概要を示す.

- ① 検索結果から上位の文書を抽出
- ② 抽出した文書を形態素解析
- ③ 文書をクラスタリング
- ④ 文書集合にラベルを付加
- ⑤ 評価結果をフィードバック

①では検索エンジン YAHOO! JAPAN(1) を用いユーザのクエリ要求に対する上位の文書を何件か抽出する. ②では, 抽出した文書に対して, 既存のツールである茶筌(2)を用いて形態素解析を行い, 以降の処理に必要な名詞のみを抽出する. 検索結果に対するユーザの評価を得るため, クラスタリングとラベリングを行い, 最終的にその結果をフィードバックすることにより, 次回以降の検索精度を向上させる.

3. 文書データベースをクラスタリング

検索により得られた文書の一つ一つに対してユーザが評価を与えなければならないのは現実的ではない. そこで, 文書を類似度によりクラスタリングしておき, クラスタ単位で評価を与える形にすれば, ユーザの負担は減少する.

ここでは, 上述のようにあらかじめ検索エンジンから抽出した上位の文書に対して形態素解析を行ったものを対象とする. その際には, ページ内における単語の出現頻度を

用い文書のベクトルを求め, 文書間類似度によりクラスタリングを行う.

4. ラベリング

クラスタリングされた文書に対して適合度の高いものに対してラベリングを行い, それらをデータベースに保存する. 例としては, ユーザが“PC”で検索した場合, “PC 販売”や“PC 構造”などのようにラベルを付加できるようにする.

5. 結果のフィードバック

結果をフィードバックできるようにする. ラベリングを行ったものに対してはデータベースに登録し, 次回以降の検索を行う際には呼び出せるようにする. 次回以降の検索では, 同様に検索結果をクラスタリングしたもののうち, データベースから呼び出したクラスタと類似したものを表示する. ここで類似度の判定は URL のうち“://”のあとの 7 文字が一致したかどうかにより行う. (多くの URL が www. で始まるため 7 文字とする.)

6. 実験

ここではユーザが ipod をネットで購入したあとに, 同じくパソコン関連製品の外付け HD を購入する場合を想定する.

1 回目: ipod で検索

“販売”, “ニュース”, “機能説明”などにクラスタリングされる.

2 回目: 外付け HD で検索 (1 回目の結果を利用)

クラスタリングされた結果のうち“販売”のクラスタを優先的に表示.

類似度の判定例

ここではそれぞれの検索結果のうち URL の最初の 7 文字 “kakaku.” が一致するページを含んでいたため類似クラスタとみなされている.

①: kakaku.com/item/01309511987

②: kakaku.com/item/05384013140

(①, ②はそれぞれ “ipod” と “外付け HD” の検索結果に含まれていたものである.)

参考文献

- (1) YAHOO! JAPAN, <http://www.yahoo.co.jp>
- (2) 茶筌, <http://chasen-legacy.sourceforge.jp/>