1U-2

ユーザの興味に関連する文抽出についての一考察

横田 修作 † 金久保 正明 ‡ 菱沼 千明 ‡
† 東京工科大学大学院工学研究科 †† 東京工科大学コンピュータサイエンス学部

1 背景

現在インターネット上には, 膨大な量の情報が公開されている. Web 検索エンジンなどを利用し, 必要な情報のみを表示しても, すべてを見るのは大変な時間と労力が必要である. そこで情報を端的に表示する重要文抽出の研究が進められている. 現在主に進められている重要文抽出の研究は与えられた文章, または文章群のみを見た重要文の抽出である [1] 〜[3]. しかし, 人間が文章を見るときは, ある特定の着目点を持って見ることが多いと考えられる. つまり, 従来の重要文抽出ではユーザの要求を満たすことは困難である.

2 目的

本研究では, 特定の着目点を持った場合の文抽出方法として, TFIDF 法による語の評価に加え, ユーザの入力したクエリなどのユーザデータを考慮した重要文抽出システムを提案する.

3 提案方法概要

本研究で提案する方法は, 図 1 の (3) ようになっている. 入力されたクエリと文章を茶筌により形態素解析し, 形態素ごとに TFIDF と共起度により評価を行う. 次に算出された評価から文章の各文の評価値を計算し, 評価値の高いものから文を抽出する. 従来の研究では図 1 の (1),(2) のように TFIDF か共起度かどちらかによる評価がほとんどで, 共起度も頻出語との共起度などであったが, ユーザの着目点として入力されたクエリとの共起を評価することである特定の着目点に応じた文が抽出できるようになる. クエリとの共起関係による評価には, 分布の偏りを検定するために一般的に使われる χ^2 検定を用いる [4].

3.1 TFIDF の計算

TF と IDF はそれぞれ一般的に使われている以下の式により求める.

$$tf(d,t) = \frac{f(d,t)}{\Sigma f(d,t)}, \quad idf(t) = log\frac{N}{df(t)} + 1.0$$

A note on the sentence detection considering user's choice
Shusaku Yokota(y1104040ec@mf.teu.ac.jp) †
Masaaki Kanakubo ‡
Chiaki Hishinuma ‡
†Graduate School of Systems Engineering, Tokyo University of Technology
‡School of Computer Science, Tokyo University of Technology

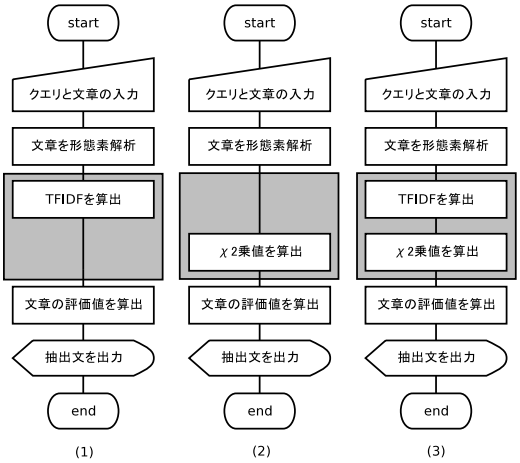

図 1: システム概要図

3.2 χ^2 値の計算

χ^2 値は $\sum \frac{(実現値-期待値)^2}{期待値}$ という式で求められる値で, 本来検定法として利用される. しかし, 本研究では単純に偏りの程度を表す値として使用する. 実現値を語 w とクエリ q∈Q との共起回数 (co_occur(q,w)), 期待値をクエリ q∈Q と語 w の共起する確率 (estab(q,w)) とすると, 以下の式で χ^2 を求められる.

$$\chi^2(w) = \sum_{q \in Q} \frac{(co\_occur(q,w) - estab(q,w))^2}{estab(q,w)}$$

4 評価

4.1 実験内容

製作したシステムの評価実験を, ある論文をもちいておこなった. 論文の内容は医学用語辞書を java の配列とデータベースにより実装し, それぞれにおける検索方法の高速化における一考察である. 実験の内容はいくつかのクエリを与え, 抽出されるキーワードと文を TFIDF のみの評価と比較した. 実験のためにクエリとして与えたものは検索アルゴリズムである"線形","2 分","ハッシュ"と, 研究の評価となる"処理時間"である. また,"ハッシュ法"というクエリを与えた文抽出結果と,TFIDF 法による文抽出結果を比較した.

4.2 実験結果と評価

実験結果の上位 10 位までの語は表 1,2 のようになった. 論文には延べ 5978 の語が出現している. また, 最も出現頻度の高い語は"情報"で出現回数は 93 回, 次に"ハッシュ"が 90 回,"検索"が 85 回となっている. "

線分”,”2 分”,”ハッシュ”は論文全般に使われている語であるため, 結果も論文に全般的に出現する語が抽出されている. 次に,”処理時間”の結果は,”かかる”,”実験”,”高速”など実験にかかわるもの,”回”のような処理時間と関係のありそうなものが抽出された. このように, クエリに関係の高い語を抽出することができた.

表 1: ”線形”,”2 分”,”ハッシュ”による処理結果

順位	抽出語	χ^2 値	出現頻度
1	探索	72.347	48
2	英略	53.714	5
3	関数	52.500	48
4	法	41.691	56
5	昨年度	34.865	8
6	医学	30.519	79
7	用語	29.106	75
8	語	27.326	24
9	方法	26.857	9
10	一つ	26.857	7

表 2: ”処理時間”による処理結果

順位	抽出語	χ^2 値	出現頻度
1	処理	21.517	39
2	かかる	9.413	9
3	実験	9.413	15
4	法	8.558	56
5	掲載	8.068	17
6	用語	7.482	75
7	回	6.742	18
8	探索	5.661	48
9	計測	5.379	4
10	高速	4.375	17

文抽出の結果は以下のようになった.

・システムの抽出した文例

(例 1) しかし, 辞書への掲載数が 2 回,3 回と増えるごとにハッシュ法をもちいない SELECT 文との処理時間差が広がり,13 回掲載されている用語になると, ハッシュ法をもちいない SELECT 文との探索時間の比は 3 倍から 4 倍にもなってしまう.

(例 2)2 分探索法とハッシュ法を実装したプログラムで, 昨年度 Java のテーブルとして作成され,746 行登録されている医学略語辞書による変換処理時間を計測し, 比較した (表 5).

(例 3) 医学略語辞書の探索を高速化するために,2 分探索法, ハッシュ法の二つの探索アルゴリズムで処理速度のテストし, 考察と改良をおこなう.

・TFIDF によって抽出された文例

(例 1) さらに, サイトに公開されている情報は多くの場合, 正確さを審査されることがないため, 検索された情報には目的とする情報だけでなく, 無責任に公開された情報や, 発信者が誤解したまま公開された情報, 閲覧者を騙そうという悪意のもとに公開された情報なども含まれる可能性がある.

(例 2) また, 一つ目のパラメタから欧語正式表記「artery」は英語であること, 二つ目と三つ目のパラメタから最新・医学略語辞典には掲載されているが, ステッドマン医学略語辞典には掲載されていないこと, 四つ目のパラメタから登録者の判断で参照元から変更はしていないことがわかる.

(例 3) そのため「欧語正式表記が何語であるか」,「最新・医学略語辞典に掲載されているか」,「ステッドマン医学略語辞典に掲載されているか」,「データを登録者の判断で参照元から変更したか」をパラメタとして入力していく.

TFIDF による文抽出は当然クエリに関係なく, 語の頻度のみで評価をしているので論文の内容に添ったものが抽出されるとは限らない. しかし, システムにより抽出した文は, ハッシュ法の実験結果, 実験内容などクエリに関連した文が抽出されている.

5 今後の課題

今後の主な課題として, ユーザ情報の追加. ダブルコーテーションなどでくくられたフレーズへの対応. クエリと共起度の高いの共起語と共起している語, つまり 2 段階目の共起語の評価への導入. 現在の共起度の測定方法では発見できないような語間の隠れた共起情報を抽出できるようなフィルタリング処理の提案と導入. 照応関係, 特に代名詞照応の解析をおこなうことを検討している.

参考文献

[1] 伊藤潤, 酒井哲也, 平澤茂一:「係り受け木を用いた日本語文書の重要部分抽出」, 電子情報通信学会,2003 Nov.

[2] 砂山渡, 谷内田正彦:「文章の特徴を表すキーワードを発見して重要文を抽出する展望台システム」, 電子情報通信学会,Vol.J84-D-I No.2 pp.146-154 2001 Feb.

[3] 大竹清敬, 岡本大吾, 児玉充, 増山繁:「重要文抽出, 自由生成要約に対応した新聞記事要約システム YELLOW」, 情報処理学会論文誌,Vol.34 No.SIG2(TOD13) 2002 Mar.

[4] 松尾豊, 石塚満:「語の共起度の統計情報に基づく文書からのキーワード抽出アルゴリズム」, 人工知能学会論文誌,Vol.17 No.3 2002.