2007 年度卒業研究概要

索引自動作成システムの研究

大谷　紀子研究室
0432225 山本洋平

１．研究の背景と目的

仕事や研究においてレポートや論文を書く機会は非常に多い。しかし、レポートや論文では目次を作成しても、索引は作成しないのが一般的であり、読者が目的の単語を探すのに手間がかかる。また、書物の索引を作るときは、著者や編集者が文章を入念にチェックする必要があり、多くの時間を費やしている。既存のソフトでの索引作成機能を用いても，著者自身が索引とすべき単語を入力しなければならないため、索引作成の労力はあまり軽減されない。

索引がないと、読者は目的の単語を見つけることが困難で、非常に不便である。本研究ではレポートや論文に索引を付けて読者の利便性向上を図ることと、大量の文章から索引を自動作成して著者の手間を軽減させることを目的とする。自動索引作成システムを構築し、評価実験によりその有用性を示す。

２．システム概要

文章に含まれる名詞と未分類語から、索引とすべき単語の候補を抽出するシステムを構築する。形態素解析には ChaSen[1]を用いる。ChaSen の性質上“ オブジェクト指向プログラミング ” などの複合語は “ オブジェクト ”・“ 指向 ”・“ プログラミング ” と分かれて形態素解析される。単語が分離された状態では索引語として不適切であるため、名詞や未分類語が連続して出現した場合には、連結して一単語とする。索引とすべき単語を決定する際に用いる重みの算出方法を以下に記す。

① 文書中の出現頻度に基づき、局所的重み l_{ij} を与える。文書中に頻繁に出現する単語は重要度が高く、索引に含まれるべきとの考えから、頻出単語に大きな重みが与えられるよう，対数化索引語頻度[2]を用いる。主に再現率向上を目的とした重みである。対数化索引語頻度には以下の式で算出される。

$l_{ij} = \log(1+f_{ij})$　　*f_{ij}(索引語頻度)：索引語 w_i の文書 D_i における出現頻度

しかし、上記の計算式だけでは日常的に用いる単語にまで大きな重みを与える可能性がある。

② 文書全体における特定の単語の分布を考慮し、大域的重み g_i を算出する。ある特定の文書にのみ集中して出現する単語は、その文書の特徴を表すことが多いとの考えから、他の文書には存在せず特定の文書にのみ存在する単語には大きな重みを与えられるよう、文書頻度の逆数[2]を用いる。主に適合率向上を目的とした重みである。文書頻度の逆数には以下の計算式を用いる。

$g_i = \log(n/n_i)$　*n(文書数):文書集合中の文書の総数　*n_i(文書頻度):索引語 w_i を含む文書数

上記の計算式では、多くの文章を本システムに読み込ませることで、より正確な結果が表示される。そのために、あらかじめ多数の文書データを本システムに蓄積する。文書データとしては、wikipedia より無作為に抽出した文書を使用した。

③　①、②の 2 つの指標から、以下の式により索引語の重み d_{ij} を決定する。

$$d_{ij} = l_{ij} \times g_i$$

抽出された単語全てに③の重みを求める。索引語とすべき単語は、重み d_{ij} が一定値以上の単語のみとし、重みの大きい順に表示する。基準となる値は文章の長さ、文書データ数等に左右される。本システムにおいては $d_{ij} > 1.36$ とする。

３．評価実験と結果

本システムを用いて書籍の索引を作成し，実際に掲載されている索引と比較する。なお比較対象の書籍にはアルゴリズム入門[3]を用いる。

実験の結果を表１に示す。適合率は 23.9%、再現率は 41.1%となった。

表 1：評価実験の結果

本システムの算出した索引語数	実際の索引語数	正解数
451	263	108

４．考察

実験の結果では、適合率は 23.9%、再現率は 41.1%にとどまったが、重みの大きな単語では実際の索引語と合致することが比較的多かった点が評価できる。特定の文章中に集中して出現する単語には、高い精度で索引語が合致したといえる。しかし様々な問題点が浮き彫りとなり、改善の余地は多い。まず、抽出された候補のうち約 2 割近くが数式や、変数、プログラム言語であった点である。数式などの単語は文章中に数多く存在するが、他の文章で全く同じ数式が使用される場合は少ないので、索引語として抽出されることが多かった。しかし専門書でない限り、実際に索引として使われる可能性は低い。数式や変数において多用される英数字に対しては必要に応じて処理をする必要がある。次に類似する言葉が頻出する点である。結果として、“ アルゴリズム ”、“ 各アルゴリズム ”、“ 整列アルゴリズム ”、“ 線形探索アルゴリズム ”などといった、類似しているが索引には現れない語が多数出現した。さらには “ こと ”、“ とき ”、“ もの ” 等の文章中に非常に多く現れる単語は、たとえ他の文章でよく使われていても大きな重みが付くことが判明した。文書中数多く現れる単語にも対応できる計算式、あるいは NG ワードの設定などが必要となる。

抽出に失敗した事例としては、助詞を含む索引語が挙げられる。実際の索引には “ 自然数の和 ” や “ ハノイの塔 ” 等、途中に助詞を含む言葉がある。本システムでは、名詞と未分類語のみを索引対象としているため、言葉の途中に助詞を含むと候補の対象外となる。しかし、助詞も索引対象にすると索引候補が急激に増加し、精度のさらなる低下が予測できる。

以上より本システムにはまだまだ改善すべき点が多々あり、現段階では完璧な索引を作成することは難しい。しかし、システムが抽出した単語を参考にすることで索引作成時間を短縮できることを考慮すると、本研究は非常に有意義であるといえる。今後は文章の長さの影響も考慮した文書正規化、文章中の単語出現の分布や偏りを考慮した計算式などが課題となる。

参考文献

[1] ChaSen's Wiki, http://chasen.naist.jp/hiki/ChaSen/

[2] 北 研二 / 津田 和彦 / 獅々堀 正幹　“ 情報検索アルゴリズム ”, 共立出版, pp33-45, 2002

[3] 大谷 紀子 / 志村正道　“ アルゴリズム入門 “ , コロナ社, 2004