Webページ間のキーワード類似度を用いた検索支援システム

An Information Retrieval Support System based on Keyword Similarity between Web Pages

中川尊雄，上野秀剛

奈良工業高等専門学校　情報工学科

1. はじめに

日本国内においてインターネットの利用者数は10年間で5倍に増えており，あらゆる年齢層に爆発的に普及している[1]．一方で，一般的に初心者ユーザはWeb検索に必要な知識が十分でない．

本研究では，複数の正解ページが存在するような検索を対象に，初心者ユーザの支援を行うためのWebページ推薦手法を提案する．提案手法はユーザの検索結果と，ユーザが提示した正解ページを用いて，正解ページに類似したWebページのリストを推薦する．提案手法を実装し，被験者実験によって有効性の確認を行う．

2. システム

本研究では，TF-IDF 法と形態素解析を利用してWebページを特徴づける単語を抽出，ページ間類似度を測定することで類似ページを推薦する．以下でそれらの手法について説明する．

2.1. 日本語形態素解析

文書に含まれる単語の出現回数を数えるためには，その文書から自動で単語を抽出しなくてはならない．本研究では，Webページから自動で単語を切り出すために，日本語形態素解析を用いる．形態素解析とは，自然言語で書かれた文章を形態素という言語の中で意味を持つ最小単位に分割し，それぞれの品詞を特定する[3]手法である．

本研究では，この形態素解析を自動で行うために，日本語形態素解析エンジンMeCabを用いる．

2.2. TF-IDF 法

TF-IDF 法とは，ある文書の集合（文書セット）の中に含まれる一つの文書に注目したとき，その文書が文書セットの中でどういった単語(term)で特徴付けられるか調べる手法である．

この手法では，1)注目している文書にある単語が何度出現するかを表す TF(Term Frequency)と，2)その単語が文書セット中のいくつの文書に含まれているかを表す DF(Document Frequency)の逆数の対数をとった IDF(Inversed DF)の二つの値に注目する．ある文書中に含まれる単語ごとの特徴度(その単語が特定の文書をどの程度特徴付けているかを示す値）を式(1)を用いて算出する．

$$tf-idf_{term,doc} = tf_{term,doc} \times \log\left(\frac{N}{df_{term}}\right) \quad (1)$$

$tf\text{-}idf_{term,doc}$:文書 doc の中における単語 term の TF-IDF 値

$tf_{term,doc}$:文書 doc の中における単語 term の出現回数

df_{term}:文書セット全体に含まれる，単語 term が含まれる文書数

N:文書セットに含まれる文書の総数

この式で算出される TF-IDF の値は，高ければ高いほどその単語の特徴度が高いことを示している．

本研究ではこの TF-IDF を用いて，複数の Web ページ間の類似度を測定する．

2.3. 提案手法

前節までに説明した方法を用いて，ユーザの提示した正解の見本データから，類似ページを推薦するシステムを提案する．

図1に提案システムの概要図を示す．

提案手法では，まず，検索結果の中からユーザが見本にしたいページ（たとえば，大和郡山市の中華店A）と，検索結果で上位に表示されたWebページをデータベースとしたもの（「郡山　中華」で検索した結果の Web ページ群）を入力とし，それぞれのWebページを形態素解析にかけ，出てきた単語に対して TF-IDF を用いて特徴語の抽出を行う．

次に，見本ページとデータベースに含まれる Web ページの特徴語の一致度から見本ページとの類似度を算出し，データベース内の Web ページを類似度順に整列して推薦する．

この手法では，TF-IDF の算出に用いるデータベースを検索結果から作成する．そのため，検索結果の中でよく用いられる語（「中華」や「郡山」）は TF-IDF 値が低くなり，見本ページの特徴語から取り

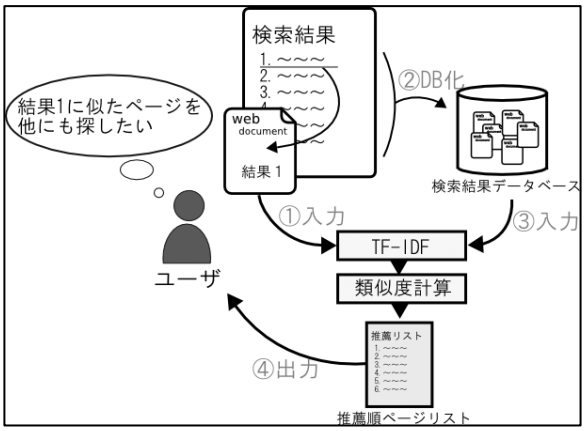

図 1. 提案システムの概要図

除かれる．一方，ユーザが検索結果からそのページを選択した要因を示す単語（例えば「ラーメン」）は TF-IDF 値が高くなるため，特徴語として選択される．

本研究では，提案手法を Linux 上で動作するシステムとして実装し，システムの有効性を確認する為に，設定したタスクに基づいて被験者実験を行った．

3. 実験

実験は，5 つのタスクを設定し，インターネットの利用経験が 3 年以上ある 10 人(男性 9 人，女性 1 人)を対象にして行った．

実験では，各タスクで指定されたキーワードを Google で検索した結果を被験者に提示し，そこに含まれるページが，タスクで設定された検索目的にどの程度沿っているかを選択式のアンケートで答えてもらう．アンケート結果を，同じタスクをわたしたときの推薦結果と比較して，被験者の回答とシステムの結果がどの程度一致するか分析する．

分析においては，システムが推薦したページと被験者が検索目的に一致していると判断したページがどの程度一致しているかをあらわす精度(precision)と，被験者が検索目的に沿うと回答したページのうちどのぐらいを推薦できているかをあらわす再現率(recall)を求めた．

4. 結果

アンケートを実施した結果得られたデータを，表 1 に集計した．この表は，横が被験者の回答，縦がシステムの推薦結果をあらわす．

結果，システムの精度は 0.603 で，推薦全体の 60%が有用であり，推薦したページが被験者にとって不要であることは少ないことがわかった．一方，

表 1. 実験結果の集計表

	有用	不要	合計
推薦する	241	159	400
除去する	234	566	800
合計	475	725	1200

再現率は 0.507 と，被験者が有用と判断したデータのうち半分は取りこぼしていた．

システムは，被験者の要求に沿った動作をしており，また不要な Web サイトを除去する能力が高いが，有用な Web サイトを推薦しきれていないといえる．

また，被験者の回答とシステムの推薦結果の間に関連性があるかどうかを統計的に評価する為，χ^2 検定を行って検証した．検定の結果，システムの推薦結果と被験者の回答の間に有意水準 99%で有意な関連が見られ，システムの推薦結果が有用であることが示唆された．

5. おわりに

本研究では，複数の正解ページを持つ検索の支援を目的に，TF-IDF 法と形態素解析によって，ユーザの提示した見本ページに類似した Web ページを推薦するシステムを提案した．またシステムの実装を行い，被験者実験による有効性の確認を行った．

その結果，このシステムを用いることでユーザの提示した見本ページに基づいて，目的とするページを推薦することができることを示した．

本研究では，実験に著者が主観で定めたタスクを用いた．今後，検索エンジンの有効性を確認するための文書セットデータベースである NTCIR のテストコレクション[4]を用いた実験を行うことで，より有効性の高い評価が可能になる．

参考文献

[1]総務省　情報通信白書(平成 21 年度版)　図表 4-1-1-3

[2]G. Salton, M.J. McGill: "Introduction to modern information retrieval," McGraw-Hill, New York (1983).

[3]長尾真: "言語の機械処理,"　三省堂(1984)

[4] Oyama,K., Eguchi,K., Ishikawa,H., Aizawa,A.: Overview of the NTCIR-4 WEB Navigational Retrieval Task 1; Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization (2004).