3M-5

Wikipedia カテゴリおよび自己相互情報量に基づく関連検索キーワード生成による知識拡充支援

河内 沙織 † 豊田 哲也 †† 延原 肇 ††
† 筑波大学 工学システム学類 †† 筑波大学 大学院システム情報工学研究科

1 はじめに

ユーザがある程度知っているキーワードに関してさらなる知識を獲得したいと思った場合，検索エンジンに当該キーワードをクエリとし，第2検索キーワードを加えた検索が有効である．しかし，未知の領域にある知識を獲得するための適切な第2検索キーワードの想起は，当該領域自体に関する知識が不足しているため一般に難しい．さらに，検索エンジンの Yahoo や Google が検索キーワード作成の補助として提示している関連検索キーワードは，多くのユーザが興味を持ちやすい，一般的なキーワードが提示される傾向にあり，ユーザにとって既知であるものが提示される可能性が高い．

本研究では，クエリとの関連性があまり知られていない関連検索キーワードを生成する手法を提案することで，ユーザのクエリに関する知識を拡充することを目的とする．提案手法では，まず，Wikipedia 内のページ本文を解析し，候補となるキーワードを抽出する．次に，自己相互情報量（以下，PMI），tf-idf を適用して候補単語の評価値を計算し，評価値の高いものを生成キーワードとする．Google の関連検索キーワードとの比較を 11 人による主観評価実験を通して行い，提案手法の有効性を示す．

2 関連研究

本研究では，クエリとの関連性が一般的に知られていないものを関連検索キーワードとして生成するため，Wikipedia を利用する．Wikipedia はオンライン百科事典の1つで，言葉とページが一意に対応し，ページには必ず1個以上のカテゴリが付与され，ページ同士の繋がりを持っている．掘ら [1] は，Wikipedia の本文から内部リンクや単語間の距離を重要視したアルゴリズムを用いることで単語を評価し，関連語を抽出している．白川ら [2] は，Wikipedia のカテゴリネットワーク構造を利用してページとカテゴリの関連度を数値化し，それに基づいたベクトルを生成することでページの特徴を表現している．本研究では，カテゴリネットワークを利用してクエリと関連したページを選択し，それらのページから PMI，tf-idf を用いて単語の評価値を計算し，関連検索キーワードを生成する．

3 提案手法

提案手法の概要を図1に示す．Wikipedia ではフリーで DB データを提供しており，それらを利用する．単語抽出の精度を上げるため，あらかじめ，Wikpedia の管理に関するもの等の直接言葉の意味とは関係しないページやカテゴリのデータを除去する．

提案手法では，まず，ユーザがクエリ q を入力し，これに該当する Wikipedia のページを $page_q$ とする．$page_q$ が属している複数のカテゴリに，同じく属しているページが他に存在する場合，それらのページを $page_q$ の関連ページとする．存在しない場合，その複数のカテゴリのうち2つ以上に属しているページの中から，より多くに属している上位 10 %のページを $page_q$ の関連ページとする．

次に MeCab[3] を利用して形態素解析を行い，本文から単語を抽出する．基本的には名詞を取得し，本文中のカギ括弧で囲まれている文章は，固有な単語とみなしてそのまま抽出する．

提案手法では，2つの単語評価手法を採用する．1つ目の手法では，PMI を関連ページにあるカギ括弧内の単語に適用する．PMI は，

$$PMI(w, c_k) = \log \frac{\frac{c_k\text{内での } w \text{ の出現ページ数}}{\text{全ページ数}}}{\frac{w \text{ の出現ページ数}}{\text{全ページ数}} \frac{c_k\text{に属するページ数}}{\text{全ページ数}}} \ (\in \mathbb{R}) \quad (1)$$

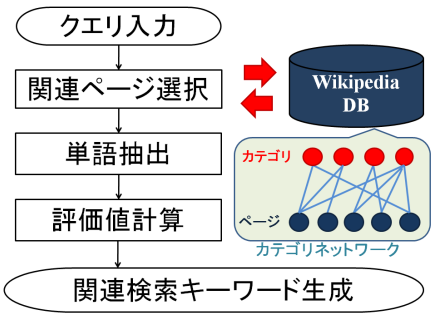


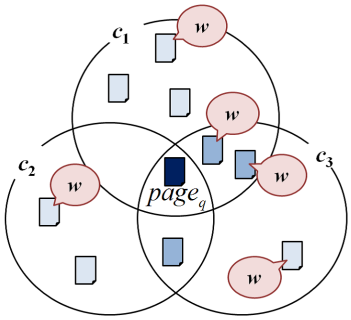
図 1: 提案システムのフローチャート

Knowledge expansion support by related search keyword generation based on Wikipedia category and pointwise mutual information

† Saori KAWAUCHI(kawauchi@cmu.iit.tsukuba.ac.jp)
†† Tetuya TOYOTA(toyota@cmu.iit.tsukuba.ac.jp)
†† Hajime NOBUHARA(nobuhara@iit.tsukuba.ac.jp)
University of Tsukuba, College of Engineering Systems, (†)
University of Tsukuba, Graduate School of Systems and Information Engineering (††)

表 1: 実験結果

手法	ある程度知っているクエリ			あまり知らないクエリ		
	想定外の割合	関連度平均	有益度平均	想定外の割合	関連度平均	有益度平均
PMI	0.670	2.591	2.363	0.736	2.586	2.193
tf-idf と PMI	0.302	3.644	2.934	0.539	4.038	2.872
google	0.159	3.605	3.249	0.273	3.198	2.808



$$PMI(w,c_1) = \log \frac{\frac{3}{10}}{\frac{5}{10}\times\frac{6}{10}}$$

図 2: PMI の計算例

で定義し，w は単語，c_k は $page_q$ が属しているカテゴリ中の任意のカテゴリとする．全ページ数は，$page_q$ が属しているカテゴリそれぞれに属している全てのページとする．図 2 に PMI の計算例を示す．単語と $page_q$ が属しているカテゴリそれぞれとの PMI を算出し，それらの平均値を単語の評価値として用いる．これは，$page_q$ 内にある単語は既知である単語が多いと仮定し，他ページの単語からクエリと関連がある単語を抽出することで，ユーザにとって知らない単語を提示する手法である．

2 つ目の手法では，$page_q$ 内にある単語は既知ではない単語も多いと仮定し，$page_q$ 内の単語に tf-idf を適用する．tf-idf は，

$$tf\text{-}idf(w) = w \text{ の出現回数} \times \log \frac{\text{関連ページ数}}{w \text{ の出現ページ数}} \quad (2)$$

で定義し，$page_q$ 内での w の出現回数，関連ページ内での w の出現ページ数とする．カギ括弧内にある単語は固有なものとして重要であるとみなし，tf-idf の値と $page_q$ のカギ括弧内にある単語の PMI 値に 0.8 をかけた結果とを足しあわせたものを評価値とする．ここで，0.8 とは事前実験により決定した値であり，tf-idf, PMI の値はそれぞれ正規化して扱う．

4 評価実験

提案手法の有効性を確認するため，被験者 11 人に対し，主観評価実験を行う．評価実験では，クエリを入力して提示された関連検索キーワードがユーザにとって想定内であるか否かについて判断してもらう．さらに想定外の関連検索キーワードについては，関連性があるか，有益な情報か，の 2 点について実際に第 2 検索キーワードとして検索を行った結果に基づき，評価してもらう．提示する関連検索キーワードは，本提案手法の 2 種類と google 関連検索キーワードの 3 種類であり，被験者にはそれぞれをランダムで提示する．

被験者に入力してもらったクエリは,「ある程度知っているクエリ」と「あまり知らないクエリ」それぞれ 5 個ずつである．結果を表 1 に示す．提案手法では，ユーザにとって想定外のキーワードが多く抽出できている．特に PMI の手法において有効であることが分かる．クエリとの関連度においても，tf-idf と PMI の手法が最も高く，$page_q$ から得られる情報が有効であることが分かった．ただし，有益度ではクエリが既知の場合で既存手法の方が有効であるため，提案手法の改善が必要である．

5 おわりに

本論文では，Wikipedia を利用し，抽出した単語に tf-idf や PMI を適用することで，ユーザにあまり知られていない関連性のある有益な関連検索キーワードを提示するシステムを提案した．実験結果より，あまり知らない単語を調べたい場合に，この提案手法がユーザの知識拡充に貢献できていることが分かった．今後はページ間の関連性の数値化を単語の評価値計算に適用させ，提案手法の精度向上を目指す予定である．

参考文献

[1] 堀憲太郎, 大石哲也, 峯恒憲, 長谷川隆三, 藤田博, 越村三幸. “Wikipedia からの拡張クエリ生成による Web 検索とその評価”. 人工知能学会 セマンティックウェブとオントロジー第 20 回.

[2] 白川真澄, 中山浩太郎, 原隆浩, 西尾章治郎. “Wikipedia のカテゴリネットワークを用いた概念のベクトル化手法”. 情報処理学会研究報告, Vol. 2008, No. 56, 2008.

[3] “MeCab”. http://mecab.sourceforge.net.