3D-4

双対概念グラフとしての文書間関連分析

下司 義寛 † 和多 太樹 † 伊藤 希 ‡ 廣川 佐千男 ††

† 九州大学大学院システム情報科学府 ‡ 筑波大学生命環境科学研究科 †† 九州大学情報基盤センター

1 はじめに

従来の検索エンジンによる検索は, 大量の文書からユーザが必要とする文書を探すためのものであり, 巨大な文書集合を小さな文書集合に「狭める検索」と言える, そこでは文書検索の高い再現率や精度, ランキングアルゴリズムの質が求められる.

一方, ユーザがより多くの知識を獲得するための検索も考えられる. ユーザが文書を読み, そこから得た知識を用いてユーザの知識集合を「広げる検索」である. 広げる検索においては, 再現率や精度よりも検索結果文書の全体像の把握や各文書間の関連を理解することが重要となる.

本稿では, ユーザは有用な文書あるいは Web ページを 1 つ知っており, その文書と関連のある文書群を探し出すという状況を考える. このような広げる検索のために従来の検索エンジンを利用する場合, ユーザは検索クエリを考えなければならないが, ユーザが自分自身の興味や関心をクエリで正確に表現することは困難である. また, 一般の検索エンジンでは, 何度も検索クエリを変え検索しなおさなければならない. 検索結果はリストの形で与えられるので, そこから知識を得るためには結果の文書を読まなければならない. Vivisimo[5] のような結果をクラスタリングする検索エンジンでも文書同士の関連は分らない. KartOO[2] は検索結果を視覚的にグラフ表示し, 文書同士の関連の有無は分るが, 上位下位の関係は分らない.

本稿では, 文書群とそれらに含まれる単語群の対応関係を単語文書行列として表現し階層構造を抽出することで, 文書間の関連を可視化する検索システムを実装した. 入力を先のユーザにとって有用な文書あるいは Web ページとすることで, 複雑な検索クエリを考える必要を無くした. また, 結果を有向グラフで表現することで, 検索結果の全体像を把握しやすくし, 知識獲得を容易にした.

Relational Analysis of Documents on Dual Concept Graph

† Yoshihiro SHIMOJI(y-shimo@i.kyushu-u.ac.jp)
† Taiki WADA
‡ Nozomi ITOU
†† Sachio HIROKAWA
Graduate School of Information Science and Electrical Engineering, Kyushu University (†)
Computing and Communications Center, Kyushu University (††)

2 概念グラフ

筆者等は文書群に特徴的な単語を抽出し, その特徴語間の関連を有向グラフで表現する概念グラフシステムを構築してきた [1], [3], [4]. 文書集合 D における単語 w の特徴量 $s(w,D), s'(w,D)$ を全文書 U での文書頻度 $df(w,U)$ と D での文書頻度 $df(w,D)$ を用いて定義し, 特徴量が閾値 0.5 を超えるものを特徴語とした.

$$s(w,D) = \frac{df(w,D)}{df(w,U)}$$

$$s'(w,D) = \frac{df(w,D)}{|D|}$$

また特徴語 u,v について v から見た u の関連度 $r(u,v)$ を u,v の D での共起頻度 $df(u*v,D)$ を用いて定義し, 関連度が閾値 0.5 を超える u,v の組を上位下位関係として抽出した.

$$r(u,v) = \frac{df(u*v,D)}{df(v,D)}$$

特徴語を節とし, 特徴語間の上位下位関係を枝とする有向グラフが概念グラフである.

3 双対概念グラフ

概念グラフは単語の特徴量と関連度をそれが使われる文書集合に着目して定式化したものである. 本稿では, 文書の特徴量と関連度をそれが使う単語集合に着目して定式化する. 双対概念グラフでは文書 $d(q)$ をクエリとして受け取り d_q に含まれる単語集合 $W(d_q)$ を検索する. 単語集合 $W(d_q)$ の単語を含む文書 d の特徴量 s_d, s'_d を $W(d)$ と $W(d_q)$ を用いて定義し, 特徴量が閾値 0.5 を超えるものを特徴的な文書とする.

$$s_d(d, W(d_q)) = \frac{|W(d) \cap W(d_q)|}{|W(d)|}$$

$$s'_d(d, W(d_q)) = \frac{|W(d) \cap W(d_q)|}{|W(d_q)|}$$

また特徴的な文書 d_1, d_2 について d_2 から見た d_1 の関連度 r_d を $W(d_1), W(d_2)$ を用いて定義し, 関連度が閾値 0.5 を超える d_1, d_2 の組を文書の上位下位関係として抽出する.

$$r_d(d1, d2) = \frac{|W(d_1) \cap W(d_2) \cap W(d)|}{|W(d_2) \cap W(d)|}$$

特徴的な文書を節, 文書間の上位下位関係を枝とする有向グラフを双対概念グラフと呼ぶ.

4 Wikiを用いた実験

Pukiwikiはだれでも編集可能なWebページであり, 個人のWebページよりも頻繁に編集, 作成, 削除が行なわれる. そのため, 長期に渡って利用されたWikiでは, ページの全体像を把握することが困難である. 事実, 筆者等の研究室のWikiでは, ユーザが既存のどのページに書き込むべきかもしくは新しいページを作成するべきか判断することが困難になっている. 自分の興味のあるトピックについてのページと関連のある文書全体を把握することができればこの問題は解決する. 本稿では, ユーザが興味を持つWikiページ名からそのページの双対概念グラフを構築するシステムを実装した.

4.1 結果と考察

筆者の2005年度の研究メモや成果についてのページ「下司」をクエリとした. このページには2779個の単語が出現しておりこの単語集合に特徴的な文書で50個以上の単語を使っている文書は31個存在した. クエリ文書「下司」についての概念グラフが図1である. 節の中には文書名, クエリ文書と共通する単語数, 単語異なり数が書かれている.

筆者の研究テーマである概念グラフに関係している文書, および検索, 概念束についての文書がクエリと関連のある文書として抽出されていることがわかる. また, 2005全体ゼミ/2006-01-06と2006全体ゼミ/2006-04-21は筆者が所属する研究室のセミナーのログである. 一見関係のない文書と考えられるが, このときのテーマが特徴語抽出法や概念グラフなど筆者の研究と関係するテーマであったため, クエリ文書と関係する文書と判断された.

しかし, 筆者の今年度の研究用ページは関連文書として抽出されないなど関連する全ての文書が抽出されるわけではなかった.

5 まとめと今後の課題

PukiWikiで作成された文書群について, あるページをクエリとし, そこに出現する単語集合に特徴的な文書を抽出した. さらに, 抽出した文書間の関連度を計算し, 文書をノード関連をエッジとするグラフを自動生成するシステムの実装を行なった. 筆者の研究のメモや成果についてのページをクエリとし実験を行なった. 筆者の研究テーマと関連する文書を抽出し, それらの上位下位の関係をグラフで表現した.

抽出された関連文書の適切性や再現性などの定性的評価や広げる検索に適しているかの評価が今後の課題である.

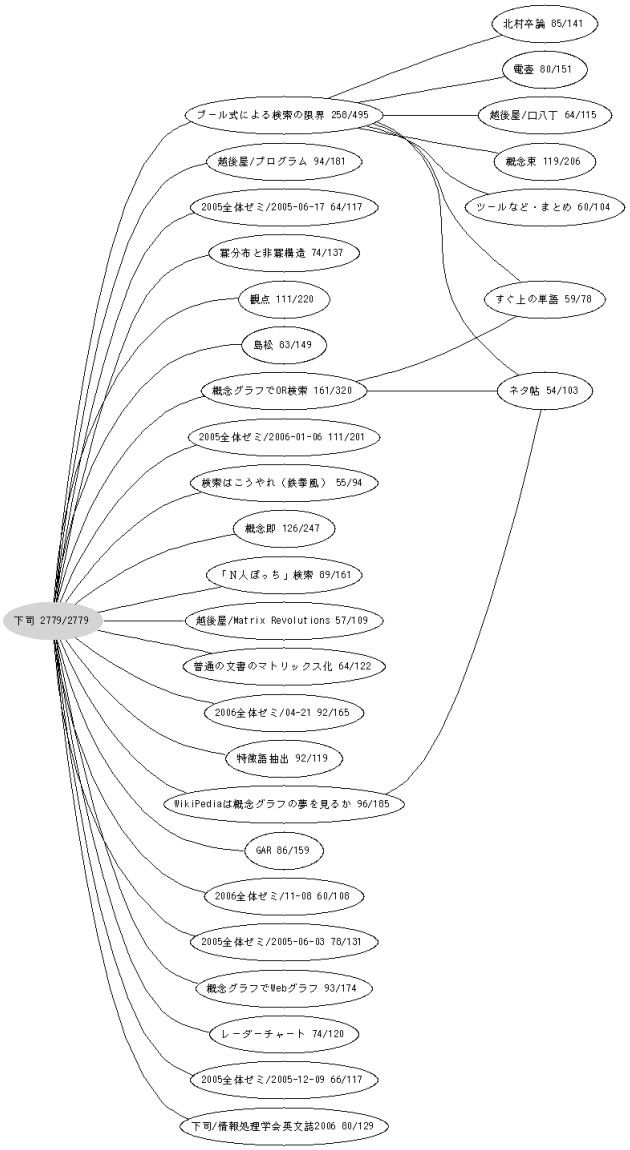

図 1: 下司の概念グラフ

参考文献

[1] 廣川佐千男, 下司義寛, 三輪眞木子, シラバスデータを使った分野ごとの概念マップの生成, 第68回情報処理学会全国大会講演論文集 3, pp.9-10, 2006

[2] KartOO, http://ww2.kartoo.com/en_index.htm/

[3] 下司義寛, 廣川佐千男, 学会講演データにおける著者やキーワードの関連分析システム, 人工知能学会 第63回 人工知能基本問題研究会, 2006

[4] 下司義寛, 和多太樹, 廣川佐千男, 英和辞典からの知識抽出, 第68回情報処理学会全国大会講演論文集 3, pp.19-20, 2006

[5] Vivisimo, http://vivisimo.com/