種々の単語間類似度を用いたネットワーク記事検索

3F－2

城風敏彦　羽生田博美　木下哲男
新情報処理開発機構　新機能沖研究室

1. はじめに

本稿では，いくつかの単語間類似度を用いたネットワーク記事のキーワード検索法を検討する．

近年インターネットに代表されるネットワークのニュース記事は量，質ともに充実しており，特にFAQに代表されるハードウエア／ソフトウエアの購入，操作についての体験談や質問／応答などの情報をネットワークから獲得できることは，利用に際して大きなメリットとなっている．

ただし，ニュース記事はあまりに膨大であるため望みの記事を検索することが非常に難しくなっている．検索技法として実用化しているのはキーワード検索システムと，全文検索システムであるが完全なキーワードやフレーズを与えることは苦痛であり，対策としてシソーラスが有効であるとされている．

本稿では従来シソーラスの自動作成に用いられてきたキーワードの共起確率[1]に加え，キーワードの連接確率，キーワード表記の類似度をシソーラスとした検索システムの報告をする．

2.キーワード抽出

通常の自然言語処理システムの形態素解析アルゴリズムをネットワーク記事のキーワード抽出に用いると，辞書にない単語を未定義語として出力してくれれば良いのだが，本来複合語として出力して欲しいキーワードの一部が辞書に登録されている単語だった場合，無意味なキーワードが抽出されてしまうので，まず字種の区切りを用い，次に辞書を分割辞書として用いる以下のキーワード抽出法を用いた．

まずヘッダ部から発信者，サブジェクト，発信時刻など抽出して，記事からヘッダ部を取り除く．次ぎにシグネーチャや引用部分を抽出し，改行コードの使い方に注意して１次キーワード抽出をする．

１次キーワード抽出は，漢字，かたかな，ひらがな，英字，数字を字種として，字種ごとのまとまりを１次キーワードとする．ただし英字については記号や数字との結合によって特別な意味（時間，金額など）を持つのでこれをルール化している．前処理として全角の英数字を半角に変換する．

次に１次キーワードをコード順にソートし辞書と一致する語頭の副詞を分割する．次に語尾の動詞を分割する．副詞，動詞だけでなく形容詞，名詞についても同様のルールをキーワード分割ルールベースに格納しておき，完全に分割できるキーワードだけを分割し，専門用語などの複合語は極力分割しないようにする．

このアルゴリズムでキーワード抽出した文書数とキーワード数との関係を図1に示す．文書数とともにキーワード数の増加率は減少するが収束の兆しは見られない．これは日々発生する新しい専門用語があることを考えれば当然といえる．実験に用いた記事はワークステーションやパソコンを扱ったニュースグループである．

3. 3種のキーワード間類似度を用いた検索システム

利用者が思い付くキーワードには限界があり，十分な検索条件を指定できるくらいであれば，問題はほとんど解決したも同然であるといえる．利用者が思い付くキーワードはほとんどの場合数個であり，表記の揺らぎや間違いも多い，メイルを例にとれば長音（メール），英語表記と日本語漢字（Mailと郵便），スペルミス（Meil），他のキーワードとの結合（Email,E-Mail,waismail,cc mail）などが挙げられる．またメールと文書も近い関係にある．これらをあらかじめ想定して従来のキーワード検索を行うこ

A Retrieval Method of Network News Documents Using Various Similarities between keywords
SHIROKAZE Toshihiko, HANIUDA Hiromi, KINOSHITA Tetsuo
RWCP, Novel Functions Oki Laboratory
c/o Multimedia Lab., Oki Electric Industry Co.,Ltd. 10-3, Shibaura 4-Chome,Minato-ku,Tokyo 108 Japan

とは利用者の負担が大きくなりすぎて現実的でない．ここでは以下の３種の単語間の関係をシソーラスとして用いることにする．

まずシソーラスの１として共起確率を用いるが，１文書内共起確率を用いるよりも，グループ内共起確率を用いた法が良いという報告もあるが，ネットワークニュースでグループを考えると，ニュースグループはあまりに巨大であるし，同じサブジェクトをもつ文献は文献数の分散やサブジェクトと関係ない内容の記事の存在などの問題があるので，１文書内共起確率を用いている．

シソーラスの２としてキーワード間連接関係を用いる．ハードウエア名(ex. Work Station) など２単語に別れていても実際は１単語として扱うべきものと係り受け関係（ex. 速いマシン）があるものを吸収するためである．連接関係にあれば当然共起関係にあるので，利用は共起確率だけを用いたのではキーワードを拡張しすぎる場合のキーワード絞り込みにも使える．また，共起確率をもちいた文書確度の計算には時間がかかりすぎる（キーワード数１万，文書数1000で３分程度，80MIPSのWSで計算）ので，連接関係で十分な場合に，小さいシソーラスとして用いる．

シソーラスの３として単語表記類似度を用いる．検索要求のキーワードも検索対象の記事にも多くの表記の間違いが混じっている．eudoraというメイラーがあるがeudraと書く人も多い(eudoraの２割程度，これはMacintoshのquadraからの影響が大きいと思われる）．２次の連結確率と動的計画法を用いて計算し，付加と脱落のペナルティを与えて表記の類似度を計算して1字程度の脱落，付加，置換に対応できるようにする．

4. 検索実験

インターネット接続関係のニュース記事1000に対して５通りの検索要求を出して，３種のシソーラスを用いその再現率と適合率の変化を見た．検索要求は[メール,マック,曖昧,Eudora]の４つのキーワードの単体と２つのキーワードのANDの計10個である．表１に従来のキーワード検索との比較を示す．適合率は少々下がっているが，再現率が向上している．全体に適合率が高めであるのは類義語の少ない[Eudora]の影響と思われる．

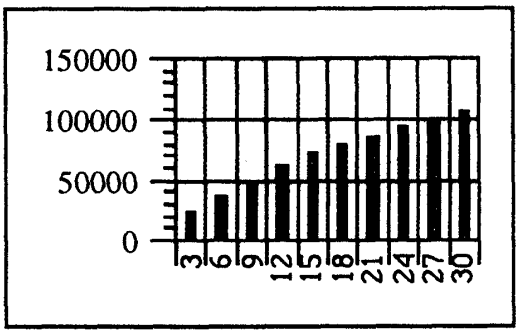
5.まとめ

従来の共起確率によるシソーラスに加えて連接確率，表記類似度適合率は多少犠牲になるが，共起確率を用いたもので46%，連接確率で5%，表記類似度で13%の向上があった．今後はこれらを検索用途に応じて動的に組み合わせて用いることにより検索率を高める検討をする予定である．

［参考文献］

[1]日本ファジィ学会編：”ファジィ・データベースと情報検索”,日刊工業新聞社(1993)

[2]新開，村岡：”グループ内共起関係を利用したキーワード間類似度計算法”,情処研資情報学基礎34-2(1994)



■ 図１文書数とキーワード数の関係(横軸単位千)

表1 各シソーラスによる検索率(単位%)

	再現率	適合率
キーワード	37	57
共起確率	54	54
連接確率	39	57
表記類似度	42	55