F_039

検索質問拡張に基づく伏せ語検索システムの試作
On a Turned Words Search System besed on Query Expansion

河合利政† 大囿忠親† 伊藤孝行† 新谷 虎松†
Toshimasa Kawai, Tadachika Ozono, Takayuki Ito, Toramatsu Shintani

1 はじめに

情報検索技術の1つに，検索式の同義語を考慮した同義語検索がある．外来語の表記の揺れや大文字と小文字，全角文字と半角文字を同一視した同義語検索は，多くの検索システムで既に実装されている．また外来語，省略形，及び通称を同一視した同義語検索については，研究が盛んに行われている．

一方で，伏せ字を使用して表記された語（以下，伏せ語と記す）については，伏せ語の本来の語（以下，原語と記す）の同義語として扱われてこなかった．伏せ語は，主にインターネット上の個人サイト及び掲示板において使用され，ポータルサイト，ショッピングサイト，及びニュースサイトといった商用サイトや法人サイトではほとんど使用されない．伏せ語を含むWeb文書は稀であるため，検索エンジンに及ぼす影響は極めて小さいと考えられる．しかし，Web文書を対象にした評判情報検索や情報クリッピングにおいては，伏せ語の存在が情報検索の妨げになっていると考えられる．そこで本研究では，WWWに存在する伏せ語を容易に検索可能な伏せ語検索システムを構築した．

本論文では，2章で検索式作成手法に基づく検索質問拡張について述べ，3章で検索結果のクラスタリング手法について述べる．最後に4章で本論文をまとめる．

2 検索式作成手法に基づく検索質問拡張

本研究では，既存の検索エンジンであるGoogle[1]の検索機能をプログラムから使用できる，Google Web APIs[2]を利用した伏せ語検索システムを構築した．Googleは2004年11月の時点で少なくとも80億ページのインデックスをもっている．そのため，Googleを利用することでWWWに広く点在する伏せ語を検索できると考えられる．

システムの概略図を図1に示す．本システムではシステム利用者がWebブラウザ上で原語を入力すると，システムが原語を元に伏せ語を作成する．作成した伏せ語をGoogle Web APIsを用いて検索する．そして結果を利用者に返している．

2.1 伏せ語の分類

一般的な同義語や類義語とは異なり，原語の同義語である伏せ語は，その使用者によってどのような形にもなり，特定の表記方法というものが存在しない．また，様々なオンラインコミュニティ及び個人サイトで日々新しい伏せ語が生み出されるため，WWW上に存在する伏せ語を網羅することは非常に困難である．しかし，いくつかのパターンに当てはまる伏せ語も実際に数多く存在する．本論文では，語の伏せ方に基づき伏せ語を表1のように分類する．本論文では表1の伏せ語を次のように定義する．

- 記号型伏せ語
  語を構成する任意の文字を，記号で置換した語．
- イニシャル型伏せ語
  語の1からk文字目を，語のイニシャルで置換した語．kは，語を構成する文字数-1以下である．

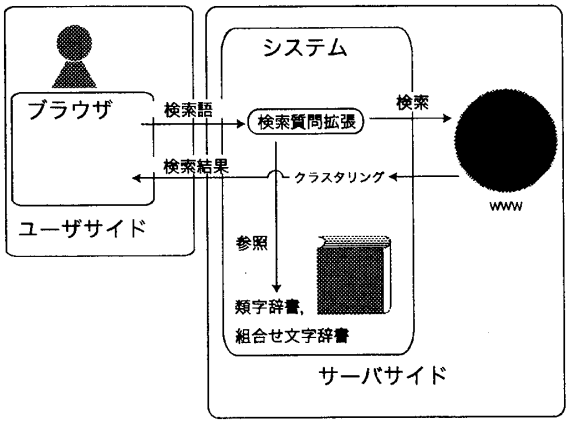

図1: システム構成

表1: 語の伏せ方に基づく伏せ語分類

分類名	用例
記号型伏せ語	名○屋市，名古屋○○大学
イニシャル型伏せ語	N古屋市，N市
類似文字型伏せ語	名古屋工業大学
組合わせ文字型伏せ語	糸且合わせ文字
当て字	南暮矢師（=名古屋市）
スラング	thx（=thanks=ありがとう）

- 類似文字型伏せ語
  語を構成する任意の文字を，類似した形状の文字で置換した語．
- 組合わせ文字型伏せ語
  文字を構成する部首や旁といった部品に着目し，隣り合った2つ以上の文字の組を，1つの類似した文字で置換した語．または，1つの文字を，類似した2つ以上の文字の組で置換した語．
- 当て字
  語の読みに着目し，字義を無視して音のみを考慮して文字を置換した語．
- スラング
  大多数の一般人には理解できないような，特定の集団の中でのみ通用する隠語，略語，及び俗語．スラングが専門用語の扱いとなる場合がある．

2.2 伏せ語を検索可能な検索式作成手法

伏せ語の数は膨大であり，原語に対応する伏せ語の同定を個々の原語について手作業で行うのは現実的でない．そこで本論文では，原語から伏せ語を自動的に作成する手法を提案する．本手法に基づき，検索語として入力された原語について，その伏せ語を検索可能な検索式（以下，伏せ語検索式と記す）を作成する．イニシャル型伏せ語では，形態素解析ツール茶筌[4]を利用して原語の読みを取得し，アルファベットに変換した．類似文字型伏せ語及び組み合わせ文字型伏せ語では，英数字，平仮名，及び片仮名についての類似文字テー

† 名古屋工業大学 知能情報システム学科
[1] http://www.google.com/
[2] http://www.google.com/apis/index.html

マイクロソフト，アップル，小泉純一郎，富士通，
三木谷，松浦亜弥，凸版印刷，富士通テン，平山あや，
名古屋工業大学，孫正義，水樹奈々，ゴルゴ，韓国，
アンジェリーナジョリー

図 2: 評価実験で用いた検索語一覧

表 2: 検索結果上位 20 件での適合率

検索結果の分類	全て	記号型
適合率	0.554	0.967
イニシャル型	類字文字型	組合わせ文字型
0.392	0.638	0.850

ブルを手作業により作成した上で，既存の漢字データベースを利用し，漢字についての類似文字セットを機械的に作成した．利用した漢字データベースは「漢字構造情報データベース[3]」及び「東雲ビットマップフォント[4]」の 12 ドットゴシック体東雲フォントデータである．

3 検索結果のクラスタリング手法

Web 検索結果のクラスタリングに適したアルゴリズムとして，Zamir らによって提案された Suffix Tree Clustering アルゴリズム（以下，STC と記す）[1] が知られている．STC は，高速，かつ単語をベースにしたクラスタリング手法よりもクラスタの内容を分かりやすく表示できるという利点をもつ一方，トピック分離性能が低いことが岡部らによって指摘されている [3]．岡部らは，トピック分離性能の向上という観点から STC を改善する手法を提案している．本研究では，岡部らによって提案された STC の改善手法を基にして，検索結果のクラスタリングを行う．

伏せ語検索によって得られる検索結果について，伏せ語の共起語情報に基づく文書クラスタリングを行うことで，伏せ語の意味の違いに基づく検索結果の分類を試みる．文書クラスタリングに基づき検索結果を分類し，ユーザに提示することで，同形異義の伏せ語の除去を試みる．

4 評価実験

4.1 実験手法及び評価方法

本研究では，伏せ語検索システムを評価する尺度として適合率を用いる．適合率は (1) 式で表したものを用いる．

$$\text{適合率} = \frac{\text{検索された文書中の適合文書数}}{\text{検索された文書数}} \quad (1)$$

全ての検索結果，記号型伏せ語，イニシャル型伏せ語，類似文字型伏せ語，及び組合わせ文字型伏せ語の検索結果について，上位 20 件の適合率を求める．評価実験に用いる検索語には図 2 に示す 15 個の語を用いる．本検索語は，本システムの利用者によって実際に入力された検索語である．評価実験では，検索式として原語のみを入力し，検索条件となるその他の検索式は入力しない．また，日本語のページを検索範囲として，伏せ語を検索する．

4.2 実験結果及び評価

先に述べた実験手法に基づき，伏せ語検索システムの評価実験を行った．実験結果を表 2 に示す．

評価実験によると，本システムは検索する伏せ語の種類により適合率が大きく異なる．記号型伏せ語の適合率が非常に高い理由としては，本システムにおいて行っている検索結果に対するフィルタリングの効果が考えられる．本システムでは，記号型伏せ語の検索結果について，記号を含まない検索結果を全て除去しているため，適合率が高くなると考えられる．しかし，検索結果の多くを除去しているため，再現率は非常に低い可能性がある．

イニシャル型，類似文字型，及び組合わせ文字型の伏せ語については，伏せ語検索式作成機構で作成された検索式にヒットした検索結果を全て表示している．そのため，記号型伏せ語の検索結果に比べ適合率が低いと考えられる．

しかしながら，本アプローチの特筆すべき点は，類字文字型及び組合わせ文字型伏せ語の適合率が，イニシャル型に比べて高い点にある．類字文字型及び組合せ文字型伏せ語はそれぞれの辞書を作成したうえで，この辞書を参照している．このため，辞書を用いないイニシャル型に対し，高い適合率になったと考えられる．また，イニシャル型伏せ語の適合率が特に低い点については，ヒューリスティクスを用いて，適度な長さの伏せ語に制限することにより改善できると考えられる．

以上より，上記の伏せ語の検索結果の適合率を高めるためには，伏せ語検索式作成機構を改良し，検索式から非適合語を除去する必要がある．

5 終わりに

本論文では，Google が収集した大量の Web 文書中に存在する伏せ語を検索する手法を提案し，伏せ語検索システムを構築した．Web 上の文書全体に占める伏せ語の割合は非常に小さいと考えられるが，企業名，製品名，及び個人名といった固有名詞の伏せ語は実際に数多く存在する．これらについて伏せ語検索を行うことは，企業にとって，リスクマネージメントの観点から十分有益性があると考えられる．

今後の課題として，アルファベット表記の伏せ語の検索，及び伏せ語辞書の自動更新があげられる．本システムでは，日本語表記の検索語に対する伏せ語検索に比べ，アルファベット表記の検索語に対する伏せ語検索の精度が低い．本研究では，日本語表記の伏せ語の検索を最優先の目的と考え，1 章で行った伏せ語の分類についても，英語圏の国ではどのような伏せ語が使われるのかを考察していない．しかし，日本語の Web 文書ではアルファベット表記の語も頻繁に使用されるため，システムの改良による本問題の解決が必要である．また，原語及び伏せ語の対応関係を記したデータベースである伏せ語辞書の自動更新手法の確立，及びその手法を実装した伏せ語辞書自動更新機構の構築が望まれる．

参考文献

[1] Oren Zamir, Oren Etzioni, "Web document clustering: a feasibility demonstration", the 21st International ACM SIGIR Conference, 1998.

[2] 市瀬龍太郎，武田英明，本位田真一，"階層的知識間の調整規則の学習"，人工知能学会論文誌，Vol.17, No.3, pp.230-238, 1998.

[3] 岡部正幸，ビクタークリサノフ，角所考，"少数フレーズに基づく文書クラスタリングと Web 検索への適用"，第 18 回人工知能学会全国大会，2004.

[4] 松本裕治，"形態素解析システム「茶筌」"，情報処理学会誌，Vol.41, No.11, pp.1208-1214, 2004.

[3] http://kanji.zinbun.kyoto-u.ac.jp/projects/chise/ids/
[4] http://openlab.jp/efont/shinonome/