3H7-01

要約情報の類似度を用いた WEB 検索支援システム

Web search support system using similarity of summary texts

國貞 暁[*1] 山本 けい子[*2] 田村 哲嗣[*3] 速水 悟[*3]
Akira Kunisada Keiko Yamamoto Satoshi Tamura Satoru Hayamizu

[*1] 岐阜大学大学院工学研究科
Graduate School of Engineering, Gifu University

[*2] 岐阜大学産官学融合センター
Collaborative Center for Academy/Industry/Government, Gifu University

[*3] 岐阜大学工学部
Faculty of Engineering, Gifu University

In the web search, the user usually repeats searches and finds similar pages in the search results. At each time, the user must select an appropriate page from the results. Moreover, there is a problem that those similar pages may not contain the web page which the user wants. In this paper, we propose a web search system having a functional button to present similar results to each item of the web search results. When the user clicks the functional button, the system shows sorted lists of the search results. Each list contains the links selected and sorted according to the similarity measures between summary texts. We compare the results of some methods for similarity, and study the effectiveness of the system.

1. はじめに

Web 上に存在する情報を検索するために，ロボット型の Web 検索がよく使われている．ロボット型検索エンジンは，ページの構成や，サイト間のリンク構造などを解析しページを順位付けしている．順位は検索ワードに対して順当に割り当てられるため，ページの並び方にページの類似性は反映されない．そのため，ある検索ワードに対して，複数の分野を含んでいたり，一部の話題にさらに言及しているページが存在したりすることは多い．このようなときに，現在は絞り込みキーワードを選び，絞り込み検索を行っている．本研究では，Web 検索の結果の一つの項目から，内容的に類似したページを集めることを目的とする．これにより，絞り込みの手間を省くことと，適切なページであるが絞り込みキーワードが含まれていないために除外されてしまうようなページの発見が可能となる．

2. Web 検索支援システム

Yahoo! のロボット検索の結果から，Web 検索結果を取得し，検索結果に付随するタイトルと要約テキストの類似度を算出し，ユーザの要求する類似ページを表示するシステムを作成した．Ajax の仕組みを用いて，クライアントの待ち時間を軽減した．作成したシステムは図 1 のように，指定した項目から動的に並び替える動作を行う．ユーザは一つの項目を指定すればよい．

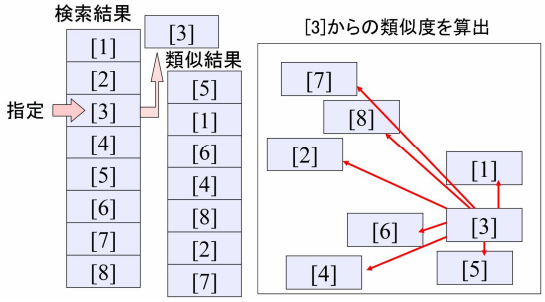

図 1 類似検索の概念

図 1 は，本システムの動作を簡略的に図示したものである．8 件の検索結果が得られると，項目間の類似性を動的に算出し，ユーザから類似結果を表示する要求があれば得られた類似度をもとに並び替えて表示する．

2.1 サーバ処理

サーバは，ユーザからの検索ワードを受け取ると，Yahoo! API[1]を用いて 800 件の検索結果を取得する．検索結果のタイトルと要約情報に対し Mecab[2]を用いて形態素解析を行い，TF*IDF 値によるベクトル空間を生成する．このベクトル空間内の類似度を保存しておく．ユーザから類似結果表示の要求があれば，800 件の検索結果を類似順に並び替えて，クライアントに受け渡す．

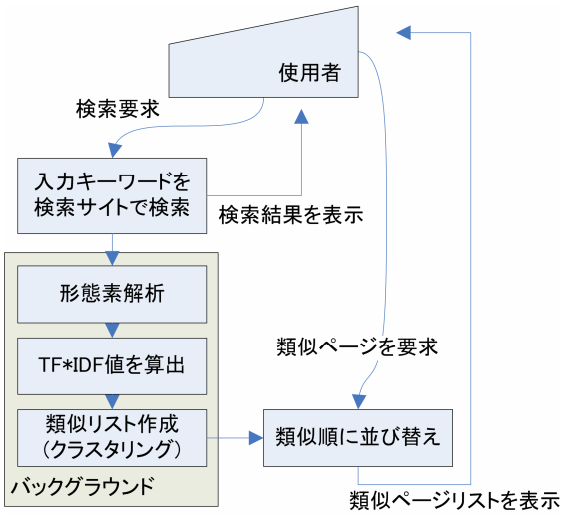

図 2 動作概要

ベクトル空間を生成するときに，名詞のみを利用している．計算コスト軽減のために上位 100 次元で処理を行う．本研究では

連絡先：國貞暁，岐阜大学大学院工学研究科応用情報学専攻，〒501-1193 岐阜市柳戸 1-1，kunisada@hym.info.gifu-u.ac.jp

[1] Yahoo! デベロッパーネットワーク http://developer.yahoo.co.jp/
[2] 形態素エンジン MeCab，工藤 拓 http://mecab.sourceforge.jp/

医療分野での検索を想定し，病名と医療関係の単語を用いるためのフィルタリングを行う．病名フィルタは，ICD-10 標準病名[3]を形態素で切り分けたものを，医療フィルタは，医歯薬英語辞書[4]の医療用語を形態素で切り分けたものを利用している．

また，類似度算出方法を比較するために，ユークリッド距離(EUCLID)，コサイン尺度(COS)，主成分分析，サモンのマップ化，自己組織マップ(SOM) [T.Kohonen 05]によるクラスタリング[T.Abe 99]を個別に実装した．SOM は，可視化が目的ではなく，次元縮小という用途で用いるため，端の無いマップとした．また，計算時間を少なくするために，正確さよりも速度を優先して収束させている．

2.2 クライアント処理

クライアントは，サーバが検索結果を取得した後は，通常の Web 検索と同じような画面で操作可能となる．このとき，サーバ側では形態素解析や類似度計算処理が継続され，全ての結果が得られると，検索結果のそれぞれの項目に設置した，類似項目を表示するボタンが使用可能となる．このボタンを使用することで，指定した項目に類似しているページを見ることができる．

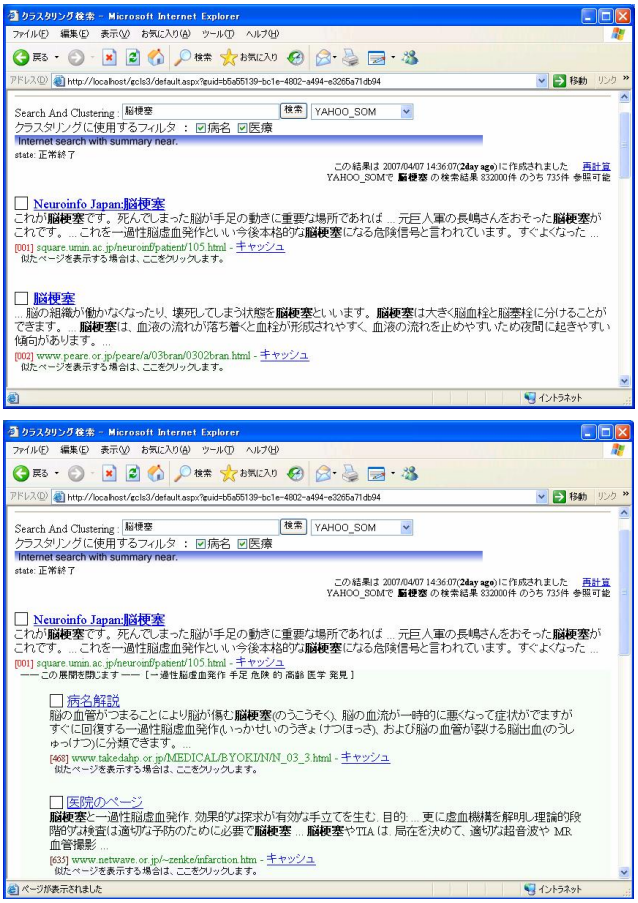

図 3 システム画面 (上:検索画面　下:類似結果)

図 3 において，上図が通常の検索結果画面であり，下図が類似結果を表示している図である．類似結果はそれぞれの検索項目から開閉でき，5 件ずつ，最大で 799 件表示できる．

[3] 財団法人医療情報システム開発センター (MEDIS-DC)
ICD10 対応電子カルテ用標準病名マスター　http://www.medis.or.jp/

[4] 医学英語研究所 (MEDO)
医歯薬英語辞書　http://www.medo.jp/

3. 評価実験

3.1 実験方法

システムの有効性を確認するために，1712 件の病名を用いて検索実験を行った．この病名は，ICD-10 標準病名のうち，Google でのヒット件数が上位であったものである．

これらの病名に対応する症状のリストを評価に使用した．例を表 1 に示す．症状リストは，病名に対して Web から症状を取り出している[日高 06]．

表 1 評価に使用した病名と症状リストの例

かぜ	熱,鼻汁,咳,下痢,のどの痛み,頭痛,くしゃみ,痰,寒気,むくむ
高脂血症	高血圧,心筋梗塞,ストレス,狭心症,脳卒中,出血,貧血,熱,代謝異常,腫瘍
高山病	頭痛,吐き気,熱,熟睡できない,嘔気,食欲不振,下痢,むくむ,息苦しい,めまい

各病名について検索を行い，病名に対する症状が，結果項目のタイトルあるいは要約中に含まれていれば，その項目は適合と判定するようにした．上位の適合項目から類似結果を表示し，再現率(Recall)と精度(Precision)を算出した．上位の適合項目とは，適合と判定された上位 3 件以内かつ，全体の項目のうち 60 位以内に相当するものである．

$$Recall = \frac{\text{検索できた適合文書数}}{\text{適合文書数}} \qquad Precision = \frac{\text{検索できた適合文書数}}{\text{検索文書数}}$$

なお，60 件とした基準は，アンケート結果より設定した．アンケートは Web 検索を行うユーザが検索結果を確認する件数を尋ねた．アンケートの対象は情報系の学生 15 名である．

3.2 実験結果

図 4 は，5 件ずつ類似結果を 60 件まで表示した際の再現率と精度を 1712 病名で平均したものである．類似結果を 5 件ずつ増やしていくので，再現率は単調に増加する．各系列において，再現率の一番低いプロットが最初の 5 件での精度を示し，以後は 5 件ずつプロットしている．再現率が増えても高い精度が保たれているほど良い結果を表す．

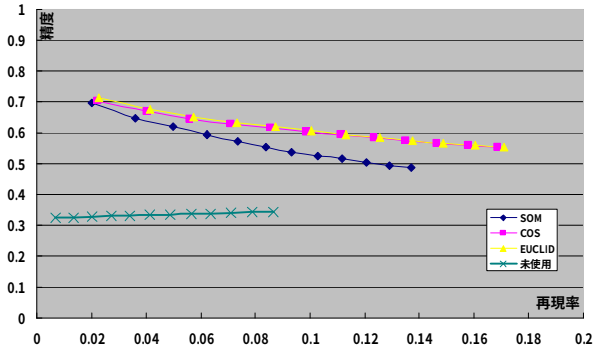

図 4 再現率-精度グラフ

適合文書を 5 個得られるまでの件数を調査した．図 5 は，5 個の適合文書を発見するまでに確認した数の中央値をとったものである．システムを使用した場合は，最初の適合項目からの類似結果を調べている．未使用の場合は，5 個目の適合項目の検索順位になる．

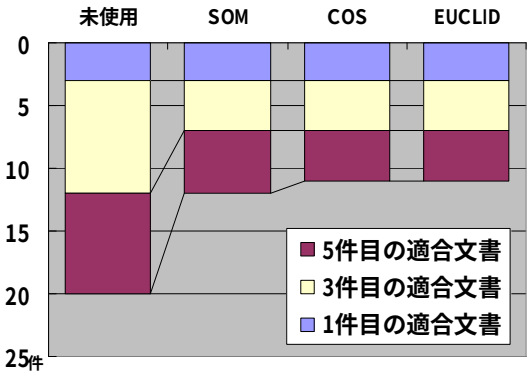

図 5　5 個の適合文書までの確認数

図 6 は，ユークリッド距離で類似度を判定した際に，適合文書を 5 個見つけるまでの，確認数の分布である．最初の適合文書で類似リストを要求した際に得られる 5 件の中に，適合文書が 4 個含まれていれば 5 件以内であり，続く 10 件目まで表示させた際に 4 個含まれていれば，6〜11 件目になる．

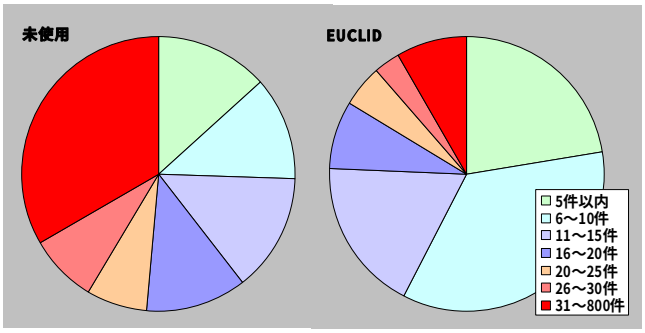

図 6　確認数の内訳

図 7 は，1712 件の検索結果において，EUCLID の類似結果 60 件中の精度が良い順に並び替えたものである．SOM や COS が，EUCLID に沿っていれば同じ傾向にあるといえる．逆に，沿っていなかったり，変動が大きかったりする場合は EUCLID と傾向が異なっているといえる．

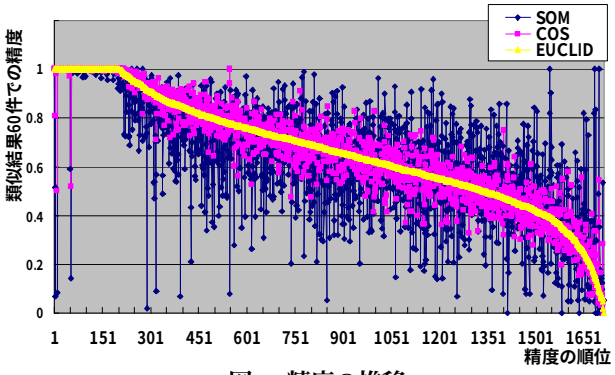

図 7　精度の推移

この計算にかかった処理の一回あたりの平均時間は表 2 の通りである．Response Time は現在の Web 検索と同等の結果が得られるまでの時間であり，Turn Around Time は類似結果が表示可能になるまでの時間を示す．Turn Around Time は，検索要求を出してから類似要求が可能になるまでの時間であるので，Response Time を含んでいる．Similarity Response Time は，一つの類似結果の表示にかかる時間である．

表 2　処理時間

	SOM	COS	EUCLID
Response Time	2500ms		
Turn Around Time	5600ms	2700ms	
Similarity Response Time	2ms	51ms	11ms

3.3 考察

図 4 の結果から，類似結果を表示すると適合文書を集められることが分かった．今回の実験では，類似計算の方法にユークリッド距離を用いると，一番良い結果となった．図 4 では示さなかったが，主成分分析及び，サモンのマップ化による精度は，SOM と未使用の間の精度を取る．傾向として SOM,主成分分析，サモンのマップ化など，クラスタリングを用いる方法は類似件数の表示数がある量を超えると精度が急激に悪くなっていた．図 5 では，5 件の適合文書を得るために，文書を確認する回数が，システムを使用しない場合に比べて減少することを示した．図 7 をみると，大まかには COS も SOM も EUCLID の結果と同じように推移している．しかし，SOM は振り幅が COS と比べて大きいことが分かる．これは，SOM が得意な部分と EUCLID や COS が得意な部分が異なっていることを示している．EUCLID が苦手としている部分でも，SOM が得意であることがあるため，類似度算出方法を使い分けると，より良い精度が期待できる．

処理時間に関しては，Response Time に 2.5 秒かかっているが，これは本システムが Yahoo! API を用いた二次的な検索システムであるためである．通常の結果が得られてから，類似結果を得るまでの計算時間は SOM であれば 2 秒ほどあるが，Turn Around Time をユーザが意識することは，ほとんどない．

4. まとめ

タイトルと要約情報を用いて Web 検索結果の類似結果を集めて表示するシステムを構築した．今回の実験では，ユークリッド距離による類似度算出がもっとも良かった．また，5 個の適合文書を得る際に，1 個目の適合項目から類似リストを表示させた場合と，表示させずに探した場合とでは，半分程度になることを示した．類似度算出方法での比較を行うと，ユークリッド距離が苦手としている検索質問では，自己組織化マップを用いたクラスタリングが得意な場合もあった．類似度算出方法によって，良い精度が出る傾向が異なっていることを示した．

今後の課題として，本システムは Yahoo! API を用いているため，Yahoo!の検索結果に大きく依存している．そのため，数週間後に同一ワードで検索を行っても，Yahoo!の検索結果に変動が生じて挙動が変化してしまうことがある．このような依存性を排除することが課題である．また，本システムは基本的には，絞り込み検索と相反するものではない．多数の絞り込みキーワードを追加すると，タイトルと要約中から得られる情報が減少してしまうので影響が出るが，数語程度なら影響は少ない．そのため，絞り込みキーワードの選択機能を追加することも検討できる．

参考文献

[T.Kohonen 05] T.Kohonen：自己組織化マップ，シュプリンガー・フェアラーク東京，2005 年.

[T.Abe 99] T. Abe, S. Kanaya, M. Kinouchi, Y. Kudo, H. Mori, H. Matsuda, D. C. Carlos, T. Ikemura：Gene classification method based on batch-learning SOM, Genome Inf. Series, No. 10, 314-315 ,1999.

[日高 06] 日高幸範，山本けい子，速水悟，田村哲嗣：医療分野における Web 文書からの情報抽出，第 26 回医療情報学連合大会，2006 年.