Web 文書検索のための非排他的クラスタリング手法の提案

~ NOCTURNE (New Overlapping Clustering Tool Using Ranking iNformation of search Engines) ~

成田 宏和† 太田 学† 片山 薫† 石川 博†

†東京都立大学 大学院 工学研究科 〒192-0397 東京都八王子市南大沢 1-1

†E-mail: {narita,ohta,katayama,ishikawa}@hikendbs.eei.metro-u.ac.jp

あらまし Web に存在する情報は、量・種類共に目覚ましい成長を続けている。この大量かつ多様な情報の中から、必要な情報にアクセスするための手段として、検索エンジンが多く用いられているが、その殆どは検索結果をリスト形式で表示するものである。一方、ユーザの検索意図は様々であり、それぞれに適した検索結果の提示方法も異なるはずである。時には、大雑把な検索語で検索し、検索結果全体の概要を知りたいと思うこともある。この場合、リスト形式で提示された検索結果から把握するのは困難である。これに対し、類似した検索結果を、適切なラベルを持つクラスタに分類して表示すれば、全体の把握が容易になる。また、このような意味的まとまりを持つクラスタは、互いに排他的なものとは限らない。これらを踏まえ、本稿では要素の重複を許容した、非排他的クラスタリング手法を提案し、実装システムとして NOCTURNE を開発した。

キーワード Web 文書クラスタリング,非排他的クラスタリング,メタサーチエンジン

Overlapping Clustering Method for Web Document Search

―NOCTURNE (New Overlapping Clustering Tool Using Ranking iNformation of search Engines)―

Hirokazu NARITA† Manabu OHTA† Kaoru KATAYAMA† Hiroshi ISHIKAWA†

†Graduate School of Engineering, Tokyo Metropolitan University, 1-1 Minami-Osawa, Hachioji-shi Tokyo, 192-0397

†E-mail: {narita,ohta,katayama,ishikawa}@hikendbs.eei.metro-u.ac.jp

Abstract The variety and volume of information on the Web are remarkably growing up. Most search engines rank search results and show them as a list, which is not always the best way to present the results. In other words, how to present search results should vary depending on user's retrieval purposes. While a list of ranked search results may help us find what we look for, clustered search results may help us grasp a whole view of retrieval topics. If search results are provided with proper labeled clusters, we can easily recognize what kinds of category compose a retrieval topic. Each search result can belong to two or more clusters. So we propose an overlapping clustering method for web document search.

Keyword Web document clustering, overlapping clustering, Meta Search engine

1. はじめに

web 文書の検索結果に対するクラスタリングは、教師無しテキスト分類問題と捉えることが出来る。テキスト分類問題に関する研究の多くは、文書集合の排他的分割、入れ子構造を持つ階層化を目的としてきた。本研究でも、入れ子構造を持つ階層的クラスタリング[1]を行うシステム METAL[10]を開発しており、他系統の枝のクラスタ同士は要素の重複がない方式であった。ところが、先にも述べた通り意味的まとまりを持ったクラスタは必ずしも排他的ではない。実際、クラスタ要素の正当性の指標として**クラスタ適合率**(4.2.1 参照)、或るクラスタに対する正当な要素の、網羅性の指標として**クラスタ再現率**(4.2.1 参照)を定義し、METAL に対する評価を行ったが、クラスタ生成プロセスの後の方で作成されたクラスタでは、含まれるべき要素が既に他のクラスタの要素となってしまい、クラスタ再現率が低くなることが確認できた。これを踏まえ、本稿では要素の重複を許容した非排他的クラスタリング手法を提案し、実装システムとして NOCTURNE を開発した。本提案手法の特徴は次の 4 点である。

1. 検索結果が順位付けされた文書集合であるということを利用している。
2. 検索要求の度に動的にクラスタリングを行う。
3. 非排他的クラスタリングである。
4. 各々のクラスタには、適当なラベルを付与する。

以降、2 章では関連研究と本提案手法との位置づけ、3 章では本提案手法について、4 章では評価方法、5 章でまとめと今後の課題について述べる。

2. 関連研究と提案手法の比較

検索結果をクラスタリングする研究として、福原は、新聞記事に対しキーワード検索を行い、検索結果をクラスタリングして、クラスタに含まれる複数の文書を同時に要約し、見出しと要約をクラスタと共に表示するという手法を提案している[2]。この手法は、新聞記事のように長い文章に対しては有効と思われるが、Web 文書は比較的文字が少なく視覚的なページも多いので、要約を読むよりも実際そのページを見た方が分かりやすいこともある。また、web ページの検索結果に表示されるサマリは、既に要約された文書だが、上記の理由で全文読まれないことも多々あると思われる。これに対し、本研究ではクラスタには、1 単語からなるラベルのみを付与し、文章を読む煩わしさを回避している。

文書を非排他的クラスタリングする手法として、相澤は、多数のテキスト文書から、文書とその出現語を同時にクラスタリングすることで、互いに重なり合う比較的小さなクラスタ集合(マイクロクラスタ)を自動生成する手法を提案している[3]。文書と語を同時クラスタリングしているので、マイクロクラスタには、語と文書が含まれている。この内、語をクラスタラベルとして用いることにより、ラベル付きの文書クラスタとすることが可能だと思われる。しかし、この手法は、インデックス手法の一つとして応用が考えられており、計算量が多いので動的処理には不向きと思われる。

3. 提案手法

本稿で提案する非排他的クラスタリング手法は、次に示す 6 つのステップからなる。

STEP1：Parse

検索エンジンから取得した、検索結果の HTML ソースを、独自に用意した Parser で処理し、HTML タグを除去した上で 1 件ずつの検索結果に分割する。Parser にはそれぞれの検索エンジンに適したパラメータを与える必要がある。この分割された検索結果の集合を R とする。また、1 つの検索結果 $r_i \in R$ は、タイトル・サマリ・URL の属性に分かれている。

STEP2：形態素解析

タイトル・サマリを、茶筅[4]を利用して形態素解析する。このとき得られた品詞情報を利用して、名詞・未知語を抽出し、これらを語集合 T とする。

STEP3：クリーニング

不要語の排除と、語の整形の目的で T をクリーニングし、特徴語集合 F を作成する。

STEP4：LI(Local Importance)の計算

$r_i \in R$ のタイトル･サマリに出現した $f_j \in F$ について、「文章の先頭に近い語ほど重要」という仮定の下に、LI を計算し、特徴語集合 LF_i を作成する。

STEP5：GI(Global Importance)の計算

$f_j \in F$ について、STEP1 で生成した R 全体に於ける出現回数 TF_j (Term Frequency)、DF_j (Document Frequency)即ち R のうちで f_j を含む $r \in R$ の数、テキスト分類などで、語の重要度指標として一般的に用いられる TF_j･IDF_j (Inverse Document Frequency)、独自に定義した SP_j (Sine Point)と LP_j (Logarithm Point)を計算し、この内いずれかを f_j の GI とする。

STEP6：クラスタリング

F から GI の高い順に $f_j \in F$ を取り出し、R の要素の内、LI が閾値以上の f_j を LF に持つ $r \in F$ を全て集め、これらを要素とする f_j のラベルが付いたクラスタ $c(f_j)$ を生成する。

以降 STEP3～6 について詳しく説明する。

3.1. クリーニング

このステップでは、不要語の除外と表現の統一・整形をし、経験則を利用して適切な語に変換する事を目的としている。

3.1.1. 不要語の削除と表現の統一

半角・全角の違いや、括弧などで括られている等の理由で、同じ語が違う語として処理されることを防止するために、次の処理を行い表現の統一を行う。

1. アルファベットを半角大文字に変換
2. アラビア数字を半角に変換
3. 半角仮名を全角片仮名に変換
4. 語頭・尾の記号を除去

次に、不要語の除外を行う。対象とする検索結果の集合は、『検索キーワード』で検索されたものなので、その殆どが『検索キーワード』を含むことになる。よって、『検索キーワード』の形態素と一致する語は除外する。また、単独では意味の分からない語として、1 文字の非表意文字（UTF8 文字コードに於いて、4E00～9FFF に含まれない文字）からなる語、数字と記号のみからなる語が挙げられる。これに対し、表意文字の場合は、1 文字からなる語でも意味が分かることが多いので、除外しなかった。まとめると、以下に示す条件の内、一つでも当てはまる語を除外対象とする。

1. 一文字の非表意文字から構成される語
2. 数字・記号のみから構成される語
3. 検索キーワードの形態素と一致する語
4. 任意指定した無視リストにある語

3.1.2. 人名の結合

人名に対し茶筅を利用して形態素解析すると、姓名詞・名名詞に分割され、一人の人物を示す語がバラバラに扱われてしまう。そこで、姓名詞・名名詞が連続して出現した場合は、これを一人の人物を示す語として、結合して扱う。以降、姓名詞・名名詞・結合された姓名名詞を総称して人名名詞と呼ぶことにする。

3.1.3. 経験則の利用

表現の統一を図るため、同じものを示す語は 1 つの語に統一し、統一後の語が出現したものとして扱う。また、経験上他の言葉に置き換えた方が分かり易いものは、置き換えた語が出現したものとして扱う。また、URL 中に出現する文字列から、経験的にそれが何のページか分かる場合は、それを示す語を生成して、その文書に出現した語として扱う。次のような例がある。

- **タイトル/サマリ中の文字列**

BBS/投稿	⇒	掲示板
jpg/gif/png	⇒	画像
ISBN	⇒	書籍

- **URL 中の文字列**

bbs/board	⇒	掲示板
syllabus	⇒	シラバス
.ac/.edu	⇒	学術機関

3.2. LI の計算

このステップでは、$r_i \in R$ のタイトル・サマリそれぞれについて、$f_j \in F$ の r_i に於ける重要度の指標として LI_j を設定する。現段階では、タイトル・サマリいずれについても、先頭から f_j までの形態素数を k として、$LI_j=10-0.5k$ としている。ただし k>20 の場合は $LI_j=0$ である。

3.3. GI の計算

このステップでは $f_j \in F$ の R 全体に於ける重要度の指標として GI を計算する。GI の候補として、テキスト文書集合に対する、出現単語の統計値として一般的に利用されている TF, DF, TF・IDF に加えて、独自に定義した SP, LP を計算する。ここで、R の要素数を N、$r_i \in R$ に於ける $f_j \in F$ の出現頻度を TF_{ij}、検索エンジンでの $r_i \in R$ の順位を n_i として SP_j, LP_j を次のように定義し、①,②の下線部をそれぞれ SP,LP 重みと呼ぶ。

$$SP_j = \sum_{i=1}^{N}\left[TF_{ij} \times \underline{\sin\left\{\frac{p}{1+\sqrt{n_i}}\right\}}\right] \times IDF_j \quad \cdots ①$$

$$LP_j = \sum_{i=1}^{N}\left[TF_{ij} \times \underline{\{1-\log_N(n_i)\}}\right] \times IDF_j \cdots ②$$

この定義により、出現回数が多ければ多いほど、出現した文書の順位が高ければ高いほど GI が高くなる。出現する文書による重み付けとして、線形的な重み付けとしなかった理由として、上位の検索結果ほどスコアが高いという性質はあるものの、順位とスコアとの間には、一般に線形的な性質は無いからである。図 1 は『東京都立大学』と『片山津温泉』を検索語として、Infoseek Japan で検索を行ったときの、上位 50 件のスコア[%]のプロットと、SP,LP 重みのグラフを 1～50 位

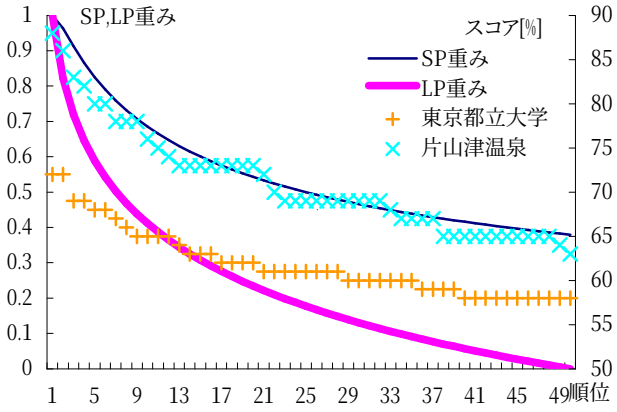

図1:スコアと重要度評価関数

について描いたものである。スコアの近似曲線を描くと、いずれの検索語の場合でも下に凸の形状となっていることが分かる。今回定義した、SP, LP 重みも同様に下に凸の形状となっており、SP 重みについては、スケールは違うもののグラフ形状は HIT 率のそれと類似していることが分かる。LP 重みは、順位減少による重要度重みの減少もより顕著にした定義である。

3.4. クラスタリング

GI の高い順に $f_j \in F$ を選び、次の 1～5 の手順を行う。$f_j \in F$ の GI には、SP_j, LP_j, DF_j, TF_j・IDF_j のいずれかを用い、それぞれに閾値を設け、閾値以下の特徴語は棄却することで無闇にクラスタ数が増えることを抑制する。また、LI に対する閾値は、クラスタ内に余計な要素が入ることを抑制する目的で設けている。

1. LF に於ける $f_j \in F$ の LI が閾値以上の $r \in R$ を集め、クラスタ $c(f_j)$を生成する
2. $c(f_j)$の要素全てが、既に生成されたクラスタの要素にもなっている場合は棄却する
3. $f_j \in F$ が人名名詞/組織名詞/地域名詞の場合は、$c(f_j)$をそれぞれ『人名』,『組織』,『地域』の子クラスタとする
4. $c(f_j)$の要素数が 1 となった場合、『その他』クラスタの子クラスタとする
5. クラスタ数が閾値(最大クラスタ数)以上になったら終了し、R の要素の内、まだどのクラスタの要素にもなっていないものを『その他』クラスタの要素とする

1 では、LF に LI が閾値以上の f_j を持つ $r \in R$ をクラスタ $C(f_j)$の要素としており、LF に LI が閾値以上の別の特徴語 f_k も含んでいれば、クラスタ $C(f_k)$の要素ともなり得る。2 では、閾値で $f_j \in F$ をフィルタリングしただけでは、クラスタ数が多くなってしまうので、新しく生成しようとするクラスタの要素が、全て他のクラスタにも含まれている場合は、新しく生成しようとしているクラスタを棄却している。3 では、「人名,組織

名,地域名は一つのクラスタにまとまっていた方が分かりやすい」という仮定の下に行ったプロセスである。4 は、「要素数が 1 のクラスタは意味がない」という仮定の下に行った。5 は、最大クラスタ数を設定することで、生成クラスタ数の上限を定めている。

3.5. 提案手法のまとめ

本手法は、検索結果から抽出された特徴語 $f_j \in F$ の、検索結果全体に於ける重要度の指標として GI を利用し、クラスタラベル f_j のクラスタ $c(f_j)$を生成する際の優先度を決めている。また、各検索結果 $r_i \in R$ に於ける f_j の重要度の指標として LI を定義し、r_i の $c(f_j)$への所属可否判定に利用している。所属可否判定に用いる LI の閾値を操作することで、$c(f_j)$の要素の精度を高めることが出来ると期待できる。また、最大クラスタ数を指定することにより、生成されるクラスタ数を任意の数以下に抑えることが可能である。

3.6. 実装

実装システムとして、NOCTURNE を開発した。ユーザインタフェースを図 2 に示す。NOCTURNE は C# にて実装し、.NET フレームワークを搭載したユーザの PC 上で動作する。検索には、複数の検索エンジンの検索結果を統合して利用するメタ検索エンジン方式を採用した。統合済みの検索結果に対し、非排他的クラスタリングを行い、クラスタリング結果を左側のペインにフォルダとして提示し、フォルダを選択するとクラスタに含まれる検索結果を右下のペインに表示する。右上のラジオボタンでは、GI として利用する指標を選択することが出来る。また、右上のペインには過去の検索履歴が表示される。

4. 実験と評価

4.1. 予備実験

本手法の有効性を確認するため、NOCTURNE・METAL・Lycos Japan[11]で検索を行い、比較を行った。NOCTURNE・METAL に於いて、メタ検索で利用した検索エンジンは、goo[12], Infoseek Japan[13], tocc[14] の 3 つである。検索語としては『雪国』を用いた。生成されたクラスタを表 1 に示す。ただし、経験則により作成されたクラスタは、METAL・NOCTURNE 共に省いている。また、NOCTURNE については、最大クラスタ数を 20 としている。NOCTURNE に於いて、重要度として SP, LP, DF, TF・IDF いずれを用いた場合も『雪』がトップに現れ、微妙に順位が違うものの、現れる語のうち 15 語はいずれにも共通しており、妥当な語が多く見られる。DF は、より多くの文書で出現した単語ほど重要度が高くなるので、検索結果全体に渡り、一般的な語が上位に来ていると言える。DF 以外の指標では、『トンネル』『出版』などは生成されず、『案内』が上位に出現しているのが特徴的である。一方 METAL の方は、非排他的クラスタリングを行っており、要素数 1 のクラスタは棄却する方式なので、NOCTURNE に比べクラスタ数が少なくなっている。また、Lycos Japan では、2 クラスタしか生成されず「ゆきぐ」という意味不明なクラスタが生成されている。

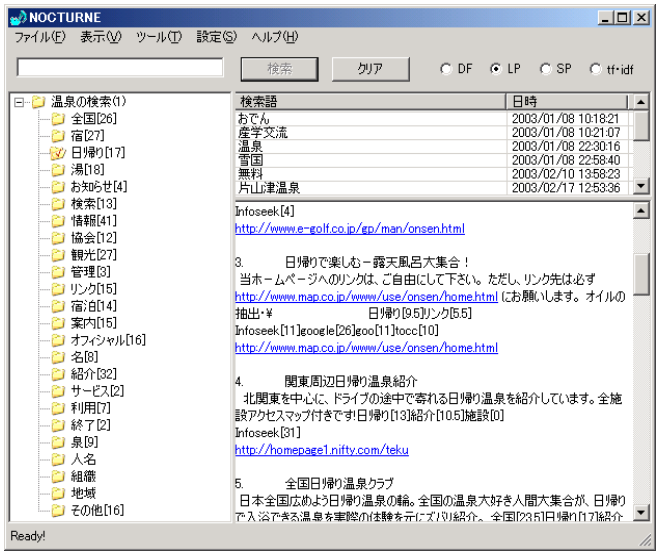

図2:ユーザインタフェース

表1:生成されたクラスタ

NOCTURNE				METAL	Lycos
DF	SP	LP	TF・IDF		
雪	雪	雪	雪	くらし	ゆきぐ
くらし	案内	案内	写真	はつらつ条例	雪国体験
情報	くらし	教科書	掲示板	世界	
生活	ゆき	ゆき	マガジン	新潟	
掲示板	社会	くらし	くらし	タクシー	
社会	写真	社会	研究	地酒	
写真	掲示板	掲示板	案内	ゆき	
更新	教科書	生活	社会	雪	
研究	更新	写真	情報	学研究センター	
ゆき	情報	更新	生活	FMゆき	
トンネル	生活	情報	ゆき	生活	
小学校	研究	研究	更新		
地酒	マガジン	温泉	教科書		
文化	地酒	マガジン	国境		
出版	文章	地酒	文化		
一覧	文化	文化	地酒		
案内	温泉	文章	開発		
世界	小学校	小学校	世界		
インターネット	世界	企画	小学校		
アマゾン	国境	世界	一覧		

4.2. 評価実験

4.2.1. 評価値の定義

我々は既に、文書のラベル付きクラスタリングの妥当性に対する評価尺度として、クラスタ再現率、クラスタ適合率、クラスタ値、クラスタリング率の定義を行った[1]。本稿では、クラスタ適合率の強調度合いに応じて評価値を変えられるように、情報検索の評価に一般的に利用されている F 値を組み込み、クラスタリング値として再定義した。

表2:クラスタ正解集合

人為的に作成した正解クラスタ集合		機械的に作成した正解集合
ラベル	URL	URL
ホテル	www.oyado-morimoto.com www.tabijozu.ne.jp/~ ryokan/chubu/kagakananoto...	www.oyado-morimoto.com www3.justnet.ne.jp/~ stsuda/KATAYAMADU.HTM www.tabijozu.ne.jp/~ ryokan/chubu/kagakananoto... www.cisnet.or.jp/~ ezo/oyado/20.ishikawa/04.kaga... www.jeims.co.jp/katayamazu/hmorimoto www.mitene.or.jp/~ koganoi www4.plala.or.jp/takefumi/etc/dousoukai-00.html
観光	www.katayamazu-spa.or.jp	www.katayamazu-spa.or.jp www.katayamazu-spa.or.jp/ryokan.html www.tabijozu.ne.jp/~ katayamazu www.cisnet.or.jp/~ ezo/oyado/20.ishikawa/04.kaga... www.incl.ne.jp/nakaya www.onsenfan.com/data/ishikawa/katayamadu.htm

評価値を求めるにあたり、クラスタに対する正解集合を定める必要がある。検索キーワードを q としたときに、被評価検索エンジンでの検索結果集合を R(q)、評価用メタ検索エンジン M での検索結果集合を M(q)とする。このとき生成される、ラベルが f_j のクラスタを $c(f_j)$ とすると、$c(f_j)$ に含まれるべき要素は $M(q+f_j)\wedge R(q)$ であると仮定する。ここで、$M(q+f_j)$ は M に於いて q と f_j で AND 検索を行ったときの検索結果である。この仮定によれば、$c(f_j)$ の要素のうち、正しい要素は $M(q+f_j)\wedge R(q)\wedge c(f_j)$ と表現することが出来る。これらを踏まえて、次のように評価値を定義する。ただし、|X|は集合 X の要素数を表すものとする。

・ **クラスタ再現率**

[0,1]の値を取り、クラスタ $c(f_j)$ に含まれるべき要素の網羅性を示す。網羅されていれば 1 である。

$$recall_j = \frac{\left|c(f_j)\wedge M(q+f_j)\wedge R(q)\right|}{\left|M(q+f_j)\wedge R(q)\right|} \cdots ③$$

・ **クラスタ適合率**

[0,1]の値を取り、クラスタ $c(f_j)$ に含まれている要素の正当性を示す。全ての要素が正当ならば 1 である。

$$precision_j = \frac{\left|c(f_j)\wedge M(q+f_j)\wedge R(q)\right|}{\left|c(f_j)\right|} \cdots ④$$

・ **クラスタリング値 CV(Clustering Value)**

[0,1]の値を取り、クラスタリング全体の精度の指標として定義する。F は⑥に示す F 値であり、b をパラメータとした値で、|b|が 0 に近いほど precision が強調される性質がある。b=0 ならば、クラスタ適合率そのものである。これを、『その他』以外の各クラスタについて、クラスタの要素数で重み付けした値の和を取り、『その他』以外のクラスタの要素数の総和で除算し正規化した値となっている。

$$CV = \frac{\sum_j \left(\left|c(f_j)\right|\times F\right)}{\sum_j \left(\left|c(f_j)\right|\right) - \left|c(\text{その他})\right|} \cdots ⑤$$

$$F = \frac{\left(1+b^2\right)\times precision\times recall}{b^2\times precision + recall} \cdots ⑥$$

・ **クラスタリング率 CR(Clustering Ratio)**

『その他』以外のクラスタにクラスタリングされた要素の割合を示し、全ての検索結果がその他以外のクラスタにクラスタリングされれば 1 である。

$$CR = \frac{\left|R(q)\right| - \left|c(\text{その他})\right|}{\left|R(q)\right|} \cdots ⑦$$

4.2.2. クラスタ正解集合の考察

前節で、クラスタ $c(f_j)$ に含まれるべき要素は $M(q+f_j)\wedge R(q)$ であると仮定した。勿論、R(q)から人為的に正解を選び出して正解集合を作成することが望ましい。しかし、|R(q)|が大きくなると人為的に正解集合を作成することは相当な苦労を要する。そこで、機械的に妥当な正解集合が得られるのならば、その方が好ましい。今回、クラスタ $c(f_j)$ に対する正解集合として仮定した $M(q+f_j)\wedge R(q)$ について、妥当性を検証するために、R(q)から人為的に選び出した正解集合との比較を行った。表 2 は、『片山津温泉』を検索語として、NOCTURNE を利用して検索を行ったときに生成されたクラスタのうち、『ホテル』と『観光』について、人為的に作成した正解集合と、機械的に生成した正解集合をまとめたものである。検索を行うに当たり、正解集合作成の際に人的な作業を伴うため、労力を削減するために、各検索エンジンからの最大取得件数は 10 件とした。人為的に作成した正解集合は、『ホテル』は純粋にホテルの情報を掲載しているページ、『観光』は観光に役立つ情報を提供しているページを正解とし、リンクや一覧などは正解に含めなかった。『ホテル』について見てみると、人為的に作成した正解集合の 2 つの URL は機械的に生成した正解集合の方にも含まれている。この他に、ホテルなのか旅館なのかよく分からないものがあったがこれは排除した。機械的に生成した正解集合のうち、人為的な正解集合に含まれていないものは、ホテルの一覧やリンクとホームページ移転による転送ページであった。『観光』の方は、機械的に生成した正解集合にはリンク

切れが 1 つ、観光協会が主催だが、観光情報とは関係無い公募のページが 1 つ、観光情報へのリンクを含むページなどが含まれていた。この結果から、機械的に作成した正解集合は、検索エンジンのインデックスを作成した時の情報の、極一部分のみを利用して作成されていることから、現在存在しないページであることはあるが、大それたものはそれほど無いと考えて良い。即ち、機械的な正解集合に対する再現率よりも適合率の方が重要な意味を持つと考えられる。そのため、先に定義したクラスタリング値を求める際の F 値のパラメータ b は、クラスタ適合率を重視するように、|b|は小さな値にする方が好ましいと思われる。

4.3. 実験方法

本提案手法を実装した NOCTURNE で、GI として SP, LP, DF, TF・IDF を用いた場合と、Lycos Japan・METAL それぞれについて検索を行い、生成されたクラスタに対して評価値を求め比較検討を行う。NOCTURNE のパラメータとして、最大クラスタ数：20, LI 閾値：6, GI 閾値:DF≧2 とした。また、METAL, NOCTURNE 共にメタ検索で利用する検索エンジンは、goo, Infoseek Japan, tocc の 3 つで、各検索エンジンからの最大取得件数は 50 件である。

検索キーワードは『無料』『壁紙』『アイドル』『ワールドカップ』『チケット』の 5 つとした。評価は、それぞれの検索キーワードで検索した場合ごとに、先に定義したクラスタ適合率・クラスタ再現率・クラスタリング値・クラスタリング率を求め、各検索キーワードで検索した場合に於ける評価値とし、5 つの検索キーワードに於ける評価値の平均を以てシステムの評価値とする。クラスタ正解集合を求める際の、判定用の検索エンジン M として、goo, Infoseek Japan, tocc の結果をマージして提示するメタ検索エンジン M を用いた。M は各検索エンジンから、最大 50 件の検索結果を取得する。

4.4. 実験結果

表 3,4 は、Lycos, METAL, NOCTURNE で『無料』を検索キーワードとして、検索を行った際の評価値をまとめたものである。表 3,4,6

表 3:『無料』での検索結果

Lycos				METAL			
ラベル	S	R	P	ラベル	S	R	P
Real com RealPlayer	2	1.000	0.500	掲示板	21	0.800	0.571
完全無料	3	0.300	1.000	レンタル	5	0.133	0.400
リンク集	7	0.423	0.714	インターネット	5	0.600	0.600
無料WEB	4	0.150	0.750	DOWNLOAD	4	0.800	1.000
無料レンタル	7	0.261	0.857	プロバイダ	4	-	-
インターネットお	3	0.500	0.333	アイドル	3	0.500	1.000
お小遣い	5	0.667	0.800	アクセスカウンター	3	0.231	1.000
その他	92	-	-	提供	2	-	-
平均	4.43	0.472	0.708	CGIレンタル	2	0.250	0.500
				歌詞検索	2	0.500	0.500
				フリー	2	0.167	1.000
				チャット	2	0.250	1.000
				20MB	2	0.667	1.000
				楽しめる	2	0.500	0.500
				作る	2	0.250	0.500
				その他	43	-	-
				平均	4.1	0.434	0.736

表 4：『無料』での実験結果

NOCTURNE																
DF				SP				LP				TF・IDF				
ラベル	S	R	P	ラベル	S	R	P	ラベル	S	R	P	ラベル	S	R	P	
掲示板	25	0.765	0.520	ACROBAT	2	1.000	1.000	ACROBAT	2	1.000	1.000	掲示板	25	0.765	0.520	
レンタル	23	0.586	0.739	掲示板	25	0.765	0.520	掲示板	25	0.765	0.520	ゲーム	6	0.833	0.833	
サービス	18	0.412	0.389	レンタル	23	0.586	0.739	レンタル	23	0.586	0.739	サービス	18	0.412	0.389	
チャット	11	0.700	0.636	サービス	18	0.412	0.389	情報	7	0.417	0.714	レンタル	23	0.586	0.739	
情報	7	0.417	0.714	チャット	11	0.700	0.636	チャット	11	0.700	0.636	ACROBAT	2	1.000	1.000	
メール	5	0.500	0.800	ゲーム	6	0.833	0.833	作成	6	0.333	0.500	メール	5	0.500	0.800	
リンク	4	0.188	0.750	情報	7	0.417	0.714	サービス	18	0.412	0.389	チャット	11	0.700	0.636	
作成	6	0.333	0.500	作成	6	0.333	0.500	リンク	4	0.188	0.750	リンク	4	0.188	0.750	
アクセス	9	0.467	0.778	メール	5	0.429	0.600	メール	5	0.500	0.800	情報	7	0.417	0.714	
画像	4	0.273	0.750	リンク	4	0.188	0.750	ダウン	2	0.000	0.000	アクセス	9	0.467	0.778	
フリー	7	0.214	0.857	アクセス	9	0.467	0.778	ゲーム	6	0.833	0.833	検索	5	0.500	0.600	
登録	2	0.250	0.500	登録	2	0.250	0.500	得	2	0.667	1.000	フリー	7	0.214	0.857	
CGI	6	0.250	0.333	URL	2	0.222	1.000	アクセス	9	0.467	0.778	作成	6	0.333	0.500	
インターネット	6	0.385	0.833	得	2	0.667	1.000	URL	2	0.222	1.000	画像	4	0.273	0.750	
ゲーム	6	0.833	0.833	フリー	7	0.214	0.857	フリー	7	0.214	0.857	登録	2	0.250	0.500	
アイドル	5	0.625	1.000	検索	5	0.500	0.600	検索	5	0.500	0.600	アイドル	5	0.625	1.000	
検索	5	0.500	0.600	FIRJ	2	1.000	1.000	総合	4	0.333	0.500	ネット	5	0.375	0.600	
ネット	5	0.375	0.600	画像	4	0.273	0.750	ガイド	2	0.250	0.500	URL	2	0.222	1.000	
サーバー	3	0.500	0.667	ガイド	2	0.250	0.500	FIRJ	2	1.000	1.000	CGI	6	0.250	0.333	
素材	3	0.167	0.333	インターネッ	6	0.385	0.833	CGI	6	0.250	0.333	FIRJ	2	1.000	1.000	
その他	26			その他	29			その他	31			その他	27			
平均	8.0	0.437	0.657	平均	7.4	0.494	0.725	平均	7.4	0.482	0.673	平均	7.7	0.495	0.715	

表 5：5 つの検索語で検索を行った評価値の平均

システム名		S	R	P	$CV_{0.5}$	CV_1	CR
NOCTURNE	DF	5.390	0.355	0.708	0.542	0.460	0.719
	SP	6.235	0.381	0.728	0.553	0.472	0.737
	LP	6.168	0.362	0.700	0.509	0.454	0.722
	TF･IDF	6.217	0.370	0.734	0.536	0.468	0.734
Lycos		12.00*	0.431	0.724	0.594	0.523	0.420
METAL		3.897	0.344	0.766	0.529	0.430	0.649

表 6：LI 閾値:10 の『無料』での検索結果

システム名		S	R	P	$CV_{0.5}$	CV_1	CR
NOCTURNE	DF	3.875	0.332	0.851	0.594	0.467	0.527
	SP	3.875	0.349	0.851	0.605	0.478	0.527
	LP	4.000	0.345	0.841	0.553	0.475	0.527
	TF･IDF	3.765	0.336	0.860	0.560	0.470	0.527

表 7：LI 閾値:-1 の『無料』での検索結果

システム名		S	R	P	$CV_{0.5}$	CV_1	CR
NOCTURNE	DF	13.2	0.567	0.546	0.517	0.537	0.955
	SP	12.8	0.577	0.571	0.516	0.536	0.936
	LP	12.6	0.552	0.542	0.485	0.530	0.927
	TF･IDF	12.8	0.584	0.528	0.481	0.534	0.964

の S,R,P はそれぞれ、クラスタ要素数,クラスタ再現率,クラスタ適合率を示している。表 5 には 5 つの検索キーワードで検索を行ったときに得られた評価値の平均をまとめた。R, P はクラスタ再現率、クラスタ適合率それぞれの平均、$CV_{0.5}$,CV_1 は F 値を求める際のパラメータ b の値をそれぞれ 0.5,1 とした際の、各検索キーワードに於けるクラスタリング値の平均である。表 6,7 は、それぞれ LI の閾値を 10, −1 としたときに『無料』で検索した際の評価値をまとめたものである。

4.5. 考察

表 3,4 の検索キーワード『無料』で検索を行った際の結果を見ると、Lycos がクラスタ数 7、METAL がクラスタ数 15、NOCTURNE がクラスタ数 20 である。検索結果の全体像の把握・関連語の連想・多様なユーザの趣向への対応などの観点からある程度のクラスタ数が必要だと思われる。この点で、Lycos のクラスタ数 7 というのは少ないように思われる。次にクラスタラベルについてみてみると、Lycos には『インターネットお』という意味不明なクラスタが出来ている。また、『無料』から連想が広がるようなラベルはあまりない。一方、METAL は『レンタル』『プロバイダ』『アクセスカウンター』...と連想が広がるクラスタが生成されている。NOCTUNE も、『ゲーム』『メール』....と連想の広がるクラスタが存在する。次に、クラスタ要素数を見てみると、Lycos は『その他』に要素が大きく偏り、要素数の大きなクラスタはあまり無い。METAL は『掲示板』という要素数 21 のクラスタ以外は、要素数 5 以下の小さなクラスタである。NOCTURNE は『掲示板』,『レンタル』,『サービス』,『チャット』という大きなクラスタがあり中程度の要素数のクラスタも多く存在する。絞り込みという観点で見ると、無駄に要素数が多く、クラスタ適合率が低いと意味をなさないが、NOCTURNE の実験結果に多く見られる、中程度(5〜10)の要素数を持つクラスタならば、十分絞り込みに貢献出来ると思われる。次に、クラスタ再現率を見てみると、クラスタ再現率の平均に於いて、それほど大きな違いは見られない。ただ、排他的クラスタリングを行っている Lycos, METAL では、クラスタ数が多くなるとクラスタ適合率が低下することは否めない。これに対し、非排他的クラスタリングを行っている NOCTURNE では、クラスタ数は一番多いが Lycos, METAL と同程度以上のクラスタ再現率を示している。次に、クラスタ適合率を見てみると、これも、各システムに於いてさほど大きな違いは見られない。また、NOCTURNE は、Lycos, METAL と同程度のクラスタ適合率を示しながら、Lycos, METAL の平均クラスタ要素数の 2 倍程度クラスタを生成することが出来ている。

次に、5 つの検索キーワードに於ける評価値の平均をまとめた表 5 を見てみると、Lycos のクラスタ要素数が抜きんでているが、これは『ワールドカップ』で検索した際に、検索語と同じラベルを持つ『ワールドカップ』という要素数 73 のクラスタと『FIFA World Cup』という要素数 9 の 2 クラスタしか生成されないという例外的な事例が原因である。これを除くと平均クラスタ要素数は 4.8 となり、NOCTURNE の方が、平均クラスタ要素数が大きくなることが分かる。クラスタ再現率を見てみても、Lycos が群を抜いている。これも、例外的事例の影響もあるが、全体的に多くのクラスタ数を生成しないことで、クラスタ再現率の低下を避けているという側面もある。また、METAL に対して NOCTURNE は、非排他的クラスタリングをすることで再現率の改善を試みたところ、GI に 4 種類の指標のうちいずれを用いた場合も若干の改善が見られた。適合率について見てみると、各システムともに大きく違いは見られない。つまり、NOCTURNE で生成されたクラスタは、Lycos, METAL で生成されたクラスタよりも要素数が大きく、数が多いながらも、クラスタの質は劣っていないことを示している。次に、CV について見てみると、b=1 のときは Lycos が抜きんでているが、要素数の大きなクラスタに於ける適合率が高くないなどの影響で、クラスタ適合率の重要度を上げて、b=0.5 とすると差が縮まっていることが分かる。次に CR を見てみると、NOCTURNE が 7 割を超えており、より多くの検索結果を『その他』以外のクラスタに振り分けることが出来ていると分かる。

次に、表 4,6,7 の結果を見てみと、LI 閾値を操作することで評価値が変化することが分かる。LI 閾値を大きくすれば、クラスタ適合率を高めることが出来るが、クラスタ適合率・クラスタリング率の低下を招く。一方、LI 閾値を小さくすればその逆になることが読み取れる。即ち、目的に応じてある程度自由にクラスタの要素数や精度を調節することが可能である。ただし、LI 閾値を−1 にした場合、クラスタ C(f)には LF に f を有する R の要素すべてが含まれることになるが、表 7 に示されている通りクラスタ再現率がそれほど 1 に近い値にはならない。この原因として、検索結果に表示されるサマリには、元のページの極一部しか表示されていないことが挙げられる。即ち、検索語 q で検索したときのサマリに表示されていない部分に、α に関する情報を含んでいれば、q+α で AND 検索したときに、ヒットする可能性があるからである。Google[15]の場合、『東京

Tokyo Metropolitan University
... 平成16年度入試からの第二部の募集停止について. 研究・技術シーズ のデータベース公開. 入学案内. 入試情報. 大学入学案内等・資料請求. Copyright 2001 Tokyo Metropolitan University. All rights reserved. **東京都立大学**広報委員会.
内容: 大学案内、学部案内、入試案内。その他、一般向け情報としてオープンカレッ...
カテゴリー: World > Japanese > ... > 関東 > 東京 > 教育・学校 > 教育
www.metro-u.ac.jp/ - 18k - 2003年2月10日 - キャッシュ - 関連ページ

図3:『東京都立大学』での検索結果

Tokyo Metropolitan University
... 学内委員会委員名簿 ・学外施設 (会津田島). **最新情報** トピック ス ビジネススクール開設のお知らせ. ... 入試情報. 大学入学案内等・資料請 求. Copyright 2001 Tokyo Metropolitan University. All rights reserved. **東京都立大学**広報委員会.
www.metro-u.ac.jp/ - 18k - 2003年2月10日 - キャッシュ - 関連ページ

図4:『東京都立大学』と『最新情報』の AND 検索結果

都立大学』のみでの検索では、図 3 に示す通り www.metro-u.ac.jp のタイトル・サマリに『最新情報』を含んでいないが、実際のページには最新情報についての記述があり、『東京都立大学』と『最新情報』で AND 検索をしたときに上位に表示された。このとき、サマリは図 4 に示す通り『東京都立大学』と『最新情報』の前後の文章がピックアップされており、『東京都立大学』だけで検索したときのサマリとは、明らかに異なっている。即ち、図 3 のサマリのみから、www.metro-u.ac.jp が『最新情報』というクラスタに適合するという判断は不可能である。一方で、『最新情報』という語を含んだ検索結果が、『最新情報』というクラスタに、不適合だと判断をすることは可能である。これは、今回用いた「文章の先頭に近い語ほど重要」という単純な法則が有効であったように、『最新情報』という語が出現した位置に相関があると言えるからである。

5. まとめ

本稿では、非排他的クラスタリングを行う METAL での問題点を踏まえ、次のことを行った。

- 出現文書の順位を利用した、語の重み付け指標として SP,LP を定義した。
- 文書内の語の重要度指標として、先頭の形態素から順に、線形的に値が減少する LI を定義した。
- 文書集合全体に於ける語の重要度 GI に基づきクラスタを生成し、LI の値を基に要素の所属可否判定を行うことで、非排他的クラスタリングを行った。
- 実装システムとして NOCTURNE を開発した。
- 評価指標として、F 値を利用したクラスタリング値を定義した。

また、本稿で定義したクラスタリング精度の評価指標を用いて行った実験により、次のことが確認できた。

- LI の閾値を操作することで、クラスタ適合率をある程度自由に操作できることから、LI がクラスタへの所属可否判定に有効に機能していることが確認できた。
- 非排他的クラスタリングを行ったことにより、METAL のクラスタ再現率に劣らない値を保ちつつ、クラスタ数を増やすことが出来た。

今回の実験結果から、文書内での語の出現位置と、その語をラベルとするクラスタへの所属可否に相関が見られると言えるが、LI として先頭の形態素から線形的に減少する値というだけでは、十分とは言えない。また、今回の実験では、生成されてクラスタの精度にのみこだわって評価を行ったが、生成されたクラスタの有用度に関しては何ら評価を行っていない。これらから、次の示す課題が挙げられる。

- 語の r∈R に於ける出現位置と、r がクラスタへ適合するか否かとの、より相関の強い法則の検討
- 生成されたクラスタ、ラベルの有用度に関する評価方法の検討

謝辞

本研究の一部は文部科学省科学研究費特定研究領域(2)「情報学:A02」(課題番号:14019075)、東京都立大学総長特別研究費(特別重点研究)による。

文 献

[1] 成田宏和,太田学,片山薫,石川博 “階層的クラスタリングを利用したメタサーチエンジンの提案” 研究報告 「データベースシステム」 アブストラクト No.128 - 050 ,pp.375-382,July.2002 http://www.ipsj.or.jp/members/SIGNotes/Jpn/04/2002/128/article050.html

[2] 福原知宏 “統計情報を用いた話題特定と文脈の再構築による複数テキスト要約” 修士論文 NAIST-IS-MT9751090,Feburary.1999

[3] 相澤彰子 “双対的クラスタリングによる情報空間のモデル化” The 16th Annual Conference of Japanese Society for Artificial Intelligence, 2002 1E4-03

[4] 茶筅 http://chasen.aist-nara.ac.jp/

[5] 高間康史,廣田薫 “情報検索作業を通じた話題分布構造可視化の提案” 第 1 回 AI 若手の集い MYCOM(人工知能学会主催) pp.58-61,May.2000

[6] 廣川佐千男 “専門検索サイトの動的統合による次世代検索システムの研究開発” http://daisen.cc.kyushu-u.ac.jp/aboutdaisen.html

[7] 長慎也,筧捷彦 “「つながり」を利用した検索語からの同義語の抽出” 日本ソフトウェア科学会第 19 回大会論文集 5E-1,Septempber.2002

[8] Oren Zamir, Oren Etzioni “Grouper: A Dynamic Clustering Interface to Web Search Results” Computer Networks,Amsterdam, Netherlands,1999

[9] Salton,G.Automatic Text Processing:The Transformation,Analysis, and Retrieval of Information by Computer. Addison-Wesley(1998)

[10] METAL http://133.86.42.149/metal/

[11] Lycos Japan http://www.lycos.co.jp

[12] goo http://goo.ne.jp

[13] Infoseek Japan http://infoseek.co.jp

[14] tocc http://www.tocc.co.jp/search/

[15] Google http://google.com