書誌検索における関連語表示法の検討

阿辺川 武　　高野 明彦
国立情報学研究所

1　はじめに

文書検索において，検索結果の文書集合ともに，その文書集合の中で特徴的なキーワード (関連語) を提示することは，利用者にとって有益であると思われる．例えばキーワード集合を眺めることで，どのような結果が得られたかを概観することができるし，キーワードを追加してさらに検索結果を絞り込むことも可能になる．後者はファセット検索とも呼ばれている．

本稿で対象としている検索は，ユーザが自由文で入力したクエリーに対して，関連度の高い本を検索するという書誌検索である．各書誌は表 1 に示すようなデータ構造を持っており，検索のインデックスには，タイトル・著者・出版・内容・目次のテキストから生成される．検索結果で得られた書誌集合に対して，その集合の内容を表すようなキーワード群を抽出したい．

現在，我々は書誌検索サイト「Webcat Plus[1]」を運営しており，現状のキーワード表示は形態素解析でわかち書きされた内容語をそのまま表示しているが，形態素 1 語では，本の内容を表現することは難しく，複合名詞のようなもう少し意味のある単位でのキーワード表示が望まれている．そこで本稿では書誌検索サービスにおいて，ユーザフレンドリーなキーワード表示に改善すべく，いくつかのキーワード抽出法を比較し，ユーザの評価実験により，どの手法が最適であるかを検証する．また抽出されたキーワードの表示の仕方にも目を向け，キーワードをクラスタリングし，各クラスタにラベルを付与する手法を提案する．

2　キーワード抽出法

文書からキーワードを抽出する手法はいままでに数多く提案されている．キーワード抽出には 2 つのフェーズがあり，1) キーワード候補の生成，2) キーワード候補のランキング，からなる．ここではそれぞれの代表な手法を概観する．

2.1　キーワード候補の生成

単純なものでは，キーワードは主に名詞から構成されていることから，名詞の連続をキーワード候補とするものである．その他には与えられた文書から語の連接情報を用いてキーワードを生成する手法 [2] や，医学分野などでは人手で生成された統制語彙を用意し，その語彙集合にあるものだけを候補とするものなどがある．一般にでは予め語彙を用意することは困難であることから動的に生成する手法が多数を占めている．

表 1: 書誌情報の例

タイトル	ボクサー：渡嘉敷勝男にみる「男」の研究
著者	勝目梓
出版	集英社 (集英社文庫), 1990.5
内容	渡嘉敷勝男―1960 年生まれのボクシング世界ジュニアフライ級の元チャンピオン。ボクシングに無縁だった 16 歳の少年が、たまたまみた具志堅用高の初防衛戦テレビ中継で、でっかい “夢” をみた―。“おれはこいつを倒してみせる”―。高校を中退し、苦しくきびしいトレーニングが始まる．そして夢の実現へ。ボクシングファンである著者が、渡嘉敷勝男の青春をとおして「男」の生き方を提示する。
目次	1 章 おれは普通の男じゃない 2 章 おれは具志堅用高を倒す 3 章 おれは主役だ 4 章 汚名はおれがそそぐ 5 章 夢がなくては生きていけない 「BOOK」データベースより

2.2　キーワード候補のランキング

キーワードは，与えられた文書内での重要度によりランキングされる．重要度の考え方は 2 通りあり，1) 単一文書内での重要度 [1, 3]，2) コーパス全体での重要度，が存在する．情報検索でよく使用される TF–IDF は，TF が文書内頻度であり，IDF が全文書中での出現頻度の逆数ということから，1) と 2) を組み合わしていると考えられる．全文書集合が定義されている場合には，TF–IDF でも IDF が大域的な情報を捉えることがきることから，単文書内の情報のみを使用した既存研究のランキング手法とほぼ同精度であることが報告されている [1].

3　キーワード抽出法の検討

現状の索引語によるキーワード表示を改善すべく，いくつかのキーワード抽出法を検討した．ここでは比較した抽出法を紹介し，複数のユーザによりどの抽出

[1] http://webcat-plus.nii.ac.jp/

表 2: クエリー「1979 年 1 月 7 日:具志堅用高がボクシング世界ジュニアフライ級チャンピオンを 7 連続防衛。」の書誌検索結果 10 件に対するキーワード

ユーザランキング	索引語 BM25	名詞連続 BM25	Wikipeida BM25	TermExtract
具志堅用高	用高	渡嘉敷勝男	具志堅用高	ボクシング
世界チャンピオン	具志堅	ボクシングファン	渡嘉敷勝男	世界チャンピオン
チャンピオン	ボクシング	ボクシング世界ジュニアフライ級	ボクシング	具志堅用高
渡嘉敷勝男	渡嘉敷	初防衛戦テレビ中継	ジュニアフライ級	チャンピオン
ボクシング	ジュニアフライ	渡嘉敷勝男 1960 年生まれ	WBA	渡嘉敷勝男
日本ボクシング	WBA	具志堅用高	ボクシングジム	ボクシング世界ジュニアフライ級
ボクサー	ボクサー	元チャンピオン	ボクサー	日本ボクシング
ジュニアフライ級	チャンピオン	具志堅用高伝	テレビ中継	日本
王者	勝男	拳児たち	タイトルマッチ	ボクシングファン
協栄ボクシングジム	ハメド	渡嘉敷	川島郭志	日本チャンピオン

法がよいかという評価実験を行なった．

3.1 検討した手法

キーワード候補の生成には，従来の索引語，名詞連続，TermExtract[2]，Wikipedia を統制語彙とした手法の 4 つを比較検討し，キーワードのランキングでは．TF，TF–IDF，BM25 を検討した．

3.1.1 キーワード候補の生成

- 索引語
  文書検索で使用している索引語をそのままキーワードとして表示する．このときの語彙集合は，形態素解析で用いる辞書 (とカタカナ語からなる未知語) に依存し，本実験では辞書に Unidic[2]を使用した．
- 名詞連続
  Unidic による形態素解析後，文節区切りを行ない，文節内で連続する名詞句をキーワードとして抽出した．正規表現では (名詞+接尾辞*) という形となる．
- TermExtract
  Web で公開されている TermExtract のモジュール[3]を利用し，キーワードを抽出した．このときの形態素解析辞書はモジュールの要求に合わせ IPA の辞書を用いた．
- Wikipedia
  Wikipedia 日本語版からページタイトルを取得し，タイトルの集合を統制語彙とした．リダイレクトや頻度 10 以上のパイプリンク (アンカーテキストとそのリンク先) もページタイトルの同意語として獲得している．与えられた文章に対し，最長一致でページタイトルとマッチングを行いキーワードとして抽出する．

[2]http://www.tokuteicorpus.jp/dist/
[3]http://gensen.dl.itc.u-tokyo.ac.jp/

3.1.2 キーワード候補のランキング

実験では，各書誌ごとに，それぞれの手法で抽出したキーワード集合を要素とするベクトルを作成し，ベクトル空間モデル (行:書誌，列:キーワード) として，クエリー (検索結果書誌集合) に対する類似度を計算し，類似度の高い順にキーワードをランキングする手法をとる．

- TF
  キーワードの重みを書誌内での頻度とし，類似度を Cosine 類似度で計算した値
- TF–IDF
  書誌内での頻度と出現書誌数の逆数をかけたものを各キーワードの重みとし，Cosine 類似度で計算した値
- BM25
  Okapi[4] というシステムで実装された TF–IDF を拡張した尺度．キーワードの重み付けと類似度の計算を 1 つの式でまとめて行なう
- TermExtract
  検索結果文書集合を 1 つの文書とみなし，TermExtract モジュールで算出した値をそのままランキングに用いる (TermExtract のみ)

各手法により，キーワード候補を抽出しランキングを行なった例を表 2 に掲載する．

3.2 評価実験

検索結果書誌集合を固定し，各手法により抽出されたキーワードの上位 10 件ずつを被験者が評価した．被験者には書誌集合と，各手法のキーワードをすべてマージして文字コード順にソートしたキーワードリストを提示し，人手で順位付けさせた．このときの指示は，「書誌集合の内容を表しているようなキーワードを順に並べ換えなさい．書誌集合の内容・目次などの情報

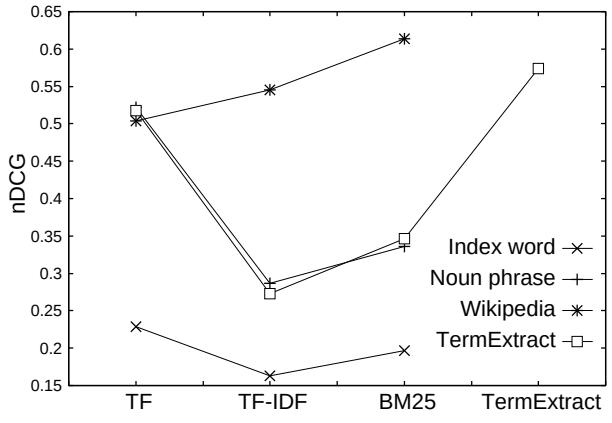

図 1: 評価実験の結果

は適時確認してもよい.」というものである．このときキーワードの定義とは何か，内容を表すとは何か，などという詳細なインストラクションはせずに，ユーザの主観にまかせる方針とした．

テストセットの書誌検索結果を生成するクエリーは，Wikipedia の「1 月 7 日」のページ中の「今日は何の日」の項目からランダムに 5 項目抽出したものである．それぞれのクエリーに対し検索結果書誌集合を上位 10 件取得し，それらの書誌情報からキーワードを抽出した．なお被験者の人数は 5 人である．

キーワードごとに全被験者の順位を平均化し，全キーワードの順位を求める．この順位を元に各手法によってランキングされた上位 10 件のキーワードを評価する．評価尺度は，ランキング上位 K 件の正規化減損累積利得 (normalized Discounted Cumulative Gain, nDCG at K) を用いる．nDCG は以下の式で算出される．

$$nDCG_q = M_q \sum_{j=1}^{K} (2^{r(j)} - 1) / \log(1 + j)$$

M は，ランキングがユーザの順位と完全に同一ある場合に，そのときの $nDCG$ の値を 1 にする正規化定数で，j は順位，$r(j) \in \{0, 1\}$ はランキング j の適合度合で，今回はユーザによる順位の逆数とした．

3.3 結果・考察

各キーワード抽出手法を $nDCG$ で評価したときの結果を図 1 に示す．ユーザによるランキングと最も一致した手法は，Wikipedia 項目をキーワードとし，ランキングを BM25 とした手法であった．次は TermExtract の実装を利用したものである．

名詞連続と TermExtract では，TF–IDF や BM25 の評価値が TF に比べ大きく低い．その理由として，両手法では動的にキーワード候補を生成するので，極端に長い文字列も候補として抽出するが，そのような

表 3: 各手法のキーワード種類数

索引語	680,910
名詞連続	6,818,396
TermExtrct	5,647,478
Wikipedia	326,363

文字列は他の書誌では出現頻度が 0 に近く，IDF 値が極端に高くなることからランキングの上位に位置するためであると考えられる．実際に各キーワード候補抽出手法におけるキーワードの種類数を数えると表 3 になり，両手法の種類数が他の 2 手法の 10 倍近くも多いことがわかる．

一方で，Wikipedia を統制語彙として利用しキーワード種類数を抑えた場合，TF よりも TF–IDF，そして BM25 によるランキングがより良い評価値となっており，一般の文書検索の結果と一致する．Wikipedia の利用がうまくいくのは，書誌情報では，Wikipedia の項目として立ちやすい固有名詞などの具体概念が多用され，それがユーザの中でも内容を表現するキーワードとして認識されやすいからであろうと思われる．

4 キーワードのクラスタリングとラベル付与

ユーザインタフェースを考えたとき，得られたキーワード集合をそのまま列挙しただけでは，数が多くなると全体を把握することが困難になる．そのため，キーワード集合をクラスタリングし，さらに各クラスタに適切なラベルが付与できれば，ユーザにとって視認性が高まる．そこで本節では，3 節で検討した Wikipedia を用いたキーワード抽出の結果に対しクラスタリングとラベル付けをする手法を提案する．

4.1 提案手法

キーワードはそのまま Wikipedia 項目に対応していることから，各項目には Wikipedia のカテゴリが付与されている．カテゴリ間はシソーラスに似たグラフ構造を成しているため，複数の Wikipedia 項目があれば，共通する祖先カテゴリ求めることができ，そのカテゴリ名をラベルとして使用できる．ただし，あまりに祖先カテゴリが上位になると抽象度が増してしまい，クラスタを表す言葉としては適切ではなくなる．したがって，子孫にはなるべく多くのキーワードを含めたいが，できるだけシソーラスの下位にあるカテゴリを求めることが目的になる．この問題を解決するために我々は先にラベルとなるカテゴリを決定し，その後に子孫にあるキーワード集合を求めるという手法をとる．

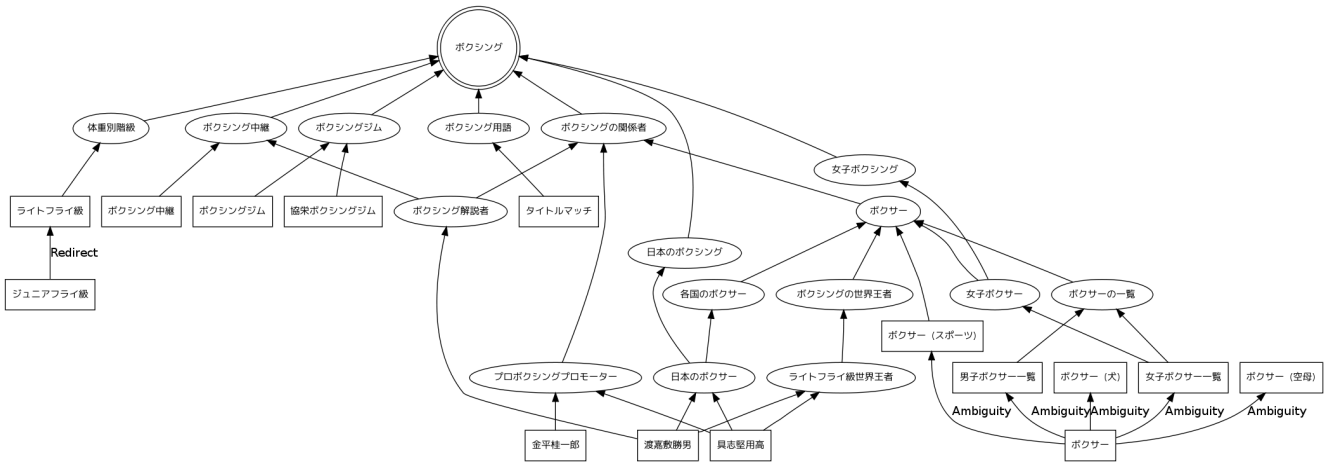

図 2: カテゴリーグラフの例

4.1.1 Wikipedia 項目・カテゴリ行列の作成

最初に各 Wikipedia 項目に対し，上位カテゴリを辿り，祖先にあるカテゴリを 5 階層上まで列挙する．たとえば項目「渡嘉敷勝男」に対しては，「日本のボクサー」「ライトフライ級世界王者」「ボクシング解説者」「太田プロ」「沖縄県出身の人物」などのカテゴリが付与されており，さらにその上位を 5 階層辿ると総計 79 カテゴリになる．このカテゴリ集合からカテゴリを要素とするベクトルを作成する．そしてすべての Wikipedia 項目に対しカテゴリベクトルを求めると，行に Wikipedia 項目，列をカテゴリとする行列が生成でき，この行列により 3.1.2 節と同様な類似度計算ができるようになる．

4.1.2 最適カテゴリーの発見

キーワード表示の視認性を考えるとすべてのキーワードをクラスタリングする必要はなく，大きなクラスタが 2,3 個あればよいと思われる．ここでは，まず最適なクラスタを 1 つ求めることにする．前節で作成した行列を用いて，検索結果書誌集合から抽出された上位 30 個の Wikipedia 項目からなる列ベクトルを作成し，この列ベクトルと類似する列ベクトルを検索する．その列ベクトルに対応するカテゴリが求めるカテゴリである．類似度の計算には BM25 を用いている．カテゴリが定まれば，その子孫にあるキーワード群がそのクラスタである．複数のクラスタが欲しいときは，すでにクラスタリングされたキーワードを除外し，残りのキーワードで同様な処理を行なう．

実際の例として，表 2 のクエリーを入力として，キーワードとして上位 30 個の Wikipedia 項目を獲得し，そこから最も類似度の高いカテゴリ (「ボクシング」) を得た．「ボクシング」をルートにして子孫に含まれるキーワードを描いたものが，表 2 のグラフであり，楕円がカテゴリ，四角がキーワードである．この例ではカテゴリ「ボクシング関係者」「ボクサー」も複数のキーワードを子孫に持つが，それよりも「ボクシング」の方がよりクラスタでまとまっていると考えられ，提案手法が有効に機能しているといえる．

5 おわりに

本稿では，書誌検索おん検索結果におけるユーザの理解の一助となるキーワード表示について，1) キーワード生成手法とランキング手法の検討，2) キーワードのクラスタ表示法の提案を行なった．ユーザによる評価実験より，キーワードの生成では Wikipedia の項目を統制語彙として使用した方法，そのときのランキングは BM25 の類似度が精度が高いものとなった．キーワードのクラスタ表示法については，今後ユーザ評価を行い，提案手法の良さを検証していきたい．

参考文献

[1] 松尾豊, 石塚満. 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム. 人工知能学会論文誌, Vol. 17, No. 3, pp. 217–223, 2002.

[2] 中川裕志, 森辰則, 湯本紘彰. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, Vol. 10, No. 1, pp. 27–45, 2003.

[3] 大澤幸生, ネルス E ベンソン, 谷内田正彦. Keygraph：語の共起グラフの分割・統合によるキーワード抽出. 電子情報通信学会誌, No. 2, pp. 391–400, 1999.

[4] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *Third Text REtrieval Conference (TREC 1994)*, pp. 109–126, 1994.