キーワード抽出の整数計画問題としての定式化

乾 孝司 † 橋本 泰一 † 高村 大也 ‡ 内海 和夫 † 石川 正道 †

概要：本稿では文書からキーワードを自動抽出する手法を提案する．提案手法は教師ありデータを用いない手法に分類される．提案手法の特徴は，単語単体に関する特徴量と，二つの単語間の関係に関する特徴を同時に適切に考慮できる点にある．これを実現するために，キーワード抽出を，組み合わせ最適化問題の下位問題である施設配置問題の考えから捉え，キーワード抽出課題を整数計画問題として定式化する．評価実験を通して，単語単体に関する特徴量と，二つの単語間の関係に関する特徴を組み合わせることが性能向上に貢献すること，及び組み合わせを実現する方法論として提案手法が有効に機能することを示す．

キーワード：キーワード抽出，整数計画問題，施設配置問題

An Integer Programming Approach for Keyword Extraction

Takashi Inui Taiichi Hashimoto Hiroya Takamura
Kazuo Utsumi Masamichi Ishikawa

Abstract : In this paper, we propose an unsupervised method for keyword extraction. Based on the idea of the facility location problem, we formalize keyword extraction as an integer programming problem. Our proposed method is able to take in information about both the feature quantity based on single-words and the feature quantity based on word co-occurences. Experimental results showed that our proposed method contributes to increase in keyword extraction performance.

Keywords : keyword extraction, integer programming, facility location problem

1 はじめに

本稿では文書からキーワードを自動抽出する手法を提案する．ここで言うキーワードとは，文書の意味内容を端的に表現する語句のことであり，論文への自動キーワード付与等，文書にキーワードを付与することが直接的な応用となる．また，文書自動要約の前処理として要約に含めるべき語句を選択する際に利用する等の応用もある．

キーワード抽出の手法は，教師ありデータを用いた教師あり機械学習に基づく手法と，教師ありデータを用いない手法の２種類に大別できる．近年では教師あり機械学習に基づく手法の開発が活発であるが (e.g. [11, 12, 5, 15, 3])，当然のことながら，教師あり機械学習に基づく手法の場合，教師ありデータの存在，あるいは，教師ありデータを負荷なく構築する環境の存在が必須条件になる．上記条件を満たすことが困難な場合は教師ありデータを用いない手法を検討することになり，２種類のいずれの手法にも需要がある．

本稿では，後者の教師ありデータを用いない手法に分類される新しいキーワード抽出法を提

† 東京工業大学 統合研究院
‡ 東京工業大学 精密工学研究所
（代表連絡先：inui@iri.titech.ac.jp）

案する．教師ありデータを用いない手法では，古典的な tf_idf 等，単語単体に関する特徴量を用いる手法と，二つの単語間の関係に関する特徴量を用いる手法（e.g. [9, 10]）がある．しかし，これまで両特徴量を同時に適切に考慮する手法は提案されていない．本稿では，両特徴量を同時に適切に考慮できる手法を提案する．

提案手法では，単語単体に関する特徴量と二つの単語間の関係に関する特徴量を考慮するにあたり，キーワード抽出を，組み合わせ最適化問題の下位問題である施設配置問題の考えから捉え，キーワード抽出課題を整数計画問題として定式化する．

以下，2 節で関連研究について述べる．3 節で提案手法について述べ，4 節で，論文アブストラクトと新聞記事を評価データに用いた評価実験について述べる．最後に 5 節で本稿をまとめる．

2 関連研究

これまで文書からのキーワード抽出課題では，幾つかの設定が考案されており，例えば，複数の類似文書が与えられている状況において各文書からキーワードを抽出する設定 [14]（例えば，検索結果のスニペット集合からのキーワード抽出）等がある．しかし，本稿では，各文書は独立であると仮定したより一般的な設定において議論を進める．

文書間の独立性を仮定した設定では，前節でも述べたように，教師ありデータを用いた教師あり機械学習に基づくキーワード抽出手法と，教師ありデータを用いないキーワード抽出手法の２種類がある．

教師あり機械学習に基づくキーワード抽出手法の基本的な考えは以下の通りである．まず，文書内の単語からキーワード候補を生成する．そして，各候補がキーワードであるか否かを分類器を用いて判定する．分類器の学習アルゴリズムとしては，SVMs[15]，Naive Bayes[4, 12]，最大エントロピー法 [11]，決定木学習 [13, 3] 等，言語処理分野で適用される学習アルゴリズムが一通り適用されている．

一方，教師ありデータを用いない手法では，古典的には，tf_idf 等によって単語（あるいは何らかの方法で生成されたキーワード候補）に重みが付けられ，重みの高い上位の単語がキーワードとして抽出される．また，二つの単語（あるいはキーワード候補）間に関する特徴量を用いる手法がある．Palshikar [10] は， 文書内において（意味的に）中心的な単語をキーワードとするために，単語を頂点，単語間の共起情報を重み付きの辺にした単語ネットワークを構築し，ネットワーク内での単語間距離に基づいて単語の中心性を計測し，中心性の値に基づいてキーワードを抽出した．Mihalcea et al. [9] は，Palshikar と同様に文書から単語ネットワークを構築し，Web ページの重み付け法として提案された PageRank [2] を単語ネットワークに適用することで各単語に重みを付与し，重みの高い上位の単語をキーワードとして抽出した．

以上のように，教師ありデータを用いない手法では，単語単体に関する特徴量を用いる手法と，二つの単語間の関係に関する特徴量を用いる手法が提案されている．しかし，これまでに両特徴量を同時に適切に考慮する手法は提案されていない．そこで本稿では，両特徴量を同時に適切に考慮できる手法を提案する．

3 提案手法

まず，提案手法におけるキーワード抽出の流れを説明する．提案手法は以下のように，二段階の手続きからなる．

(1) 文書内の単語列からキーワード候補を生成する（キーワード候補生成）．
(2) キーワード候補から抽出すべきキーワードを選別し，出力する（キーワード選別）．

上記手続きは新規なものではなく，キーワード抽出課題における一般的な設定である．2 節で述べた関連研究では，二段階ある上記手続きのうち，手続き (2) のキーワード選別についての方法論に関する研究に主眼が置かれており，本稿でも以下の 3.2 節で述べるキーワード選別法

が新規性のある提案事項になる．

なお，一般には，出力すべきキーワードが抽出元文書に含まれていない状況も考えられる．しかし，上記手続きでは文書内の単語列からキーワード候補を生成，選別することでキーワードを抽出するため，文書内に含まれていないキーワードを出力することはできない．実際に，4 節で述べる評価実験では出力すべきキーワードが抽出元文書に含まれていないケースが含まれている．この件については表 1 を参照しながら後で再度述べる．

3.1 キーワード候補生成

本稿では，以下の手続きに従い，文書内からキーワード候補を生成する．まず，文書内の各文を ChaSen[8] で解析し形態素及び品詞情報を得る．この時，形態素が未知語であるか，品詞が以下のいずれかである形態素連続をまとめあげ，キーワード候補とする．

- 名詞（非自立を除く）
- 接頭詞 – 名詞接続
- 接頭詞 – 数接続
- 記号 – 一般
- 記号 – アルファベット

3.2 キーワード選別

キーワード候補の中から出力キーワードを選別する際に単語単体に関する特徴量と二つの単語間の関係に関する特徴量を考慮するにあたり，組み合わせ最適化問題の下位問題である施設配置問題の考え方を頼りに，キーワード選別を整数計画問題として定式化することを考える．

以下，まず施設配置問題を簡単に紹介する．

3.2.1 施設配置問題

施設配置問題とは，ある地域における工場や学校などの設立地計画をモデル化した問題である．最も基本的な設定である容量制約なし施設配置問題は次のように定義される [6].

- 利用者の有限集合 $\mathcal{D}$，（開設可能な候補）施設の有限集合 $\mathcal{F}$，各施設 $i \in \mathcal{F}$ を開設するための固定コスト $f_i \in \mathbb{R}_+$ および各利用者 $j \in \mathcal{D}$ が各施設 i を利用するための利用コスト $c_{ij} \in \mathbb{R}_+$ に対し，開設コストと利用コストの総和 $\sum_{i \in X} f_i + \sum_{j \in \mathcal{D}} c_{\sigma(j)j}$ が最小となるように，開設する施設の部分集合 $X \subseteq \mathcal{F}$ と利用者の開設施設への割当 $\sigma : \mathcal{D} \to X$ を求める．

以上の問題は次のように整数計画問題として定式化できる．

$$\min \quad \sum_{i \in \mathcal{F}} f_i y_i + \sum_{i \in \mathcal{F}} \sum_{j \in \mathcal{D}} c_{ij} x_{ij} \tag{1}$$

$$\begin{aligned}
s.t. \quad & x_{ij} \le y_i && (i \in \mathcal{F}, j \in \mathcal{D}) && (2)\\
& \sum_{i \in \mathcal{F}} x_{ij} = 1 && (j \in \mathcal{D}) && (3)\\
& x_{ij} \in \{0,1\} && (i \in \mathcal{F}, j \in \mathcal{D}) && (4)\\
& y_i \in \{0,1\} && (i \in \mathcal{F}) && (5)
\end{aligned}$$

ここで，変数 y_i は施設の開設状態を表し，$y_i = 1$ のとき施設 i が開設されることを意味する．また，x_{ij} は利用者の割当状態を表し，$x_{ij} = 1$ のとき利用者 j が施設 i に割当てられることを意味する．制約式 (2) は割当先の施設は開設されている必要があることを表しており，制約式 (3) は利用者 j は必ず割当先を一箇所もつことを表している．

3.2.2 施設配置問題とキーワード選別

先の施設配置問題では利用者と施設という 2 種類のプレーヤーが登場する．ここで，

- 利用者 = キーワード候補
- 施設 = 出力キーワード

という対応関係を考えると，開設施設と利用者の開設施設への割当を求める問題は，出力キーワードとキーワード候補の出力キーワードへの割当を求める問題と読み替えることができる．ここでの割当とはキーワード候補を（出力となる）キーワード候補で代表させることを意味し，読み替え後は同一のプレーヤーがキーワード候補（利用者）かつ出力キーワード（施設）になり得る．以上を整理すると，キーワード選別における各要素は，

- $\mathcal{D}$：キーワード候補の有限集合
- $i,\ j \in \mathcal{D}$：キーワード候補
- f_i：候補 i を出力するための固定コスト（出力コスト）
- c_{ij}：候補 j を候補 i で代表させるための固定コスト（代表コスト）

となる．

3.2.3 定式化

先の対応関係に基づいてキーワード選別を次の整数計画問題として定式化する．

$$\min \quad \sum_{i\in\mathcal{D}} f_i y_i + \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} c_{ij}x_{ij} \tag{6}$$

$$\begin{aligned}
s.t. \quad & x_{ij} \le y_i && (i,j\in\mathcal{D}) && (7)\\
& \sum_{i\in\mathcal{D}} x_{ij} = 1 && (j\in\mathcal{D}) && (8)\\
& \sum_{i\in\mathcal{D}} y_i = k && && (9)\\
& x_{ii} = y_i && (i\in\mathcal{D}) && (10)\\
& x_{ij} \in \{0,1\} && (i,j\in\mathcal{D}) && (11)\\
& y_i \in \{0,1\} && (i\in\mathcal{D}) && (12)
\end{aligned}$$

ここで，y_i は候補 i を出力キーワードとするか否かを表す変数であり，x_{ij} は候補 j を候補 i に割当てる（i で代表させる）か否かを表す変数である．この最適化問題を解いた後，$y_i = 1$ となる i が出力キーワードである．制約式 (9) 及び (10) はキーワード選別のための追加制約であり，制約式 (9) は出力するキーワード数は k 個（$1 \le k \le |\mathcal{D}|$）であることを表しており，制約式 (10) は，候補 i を出力するなら代表先は自身とし，出力しないなら代表先は自身以外とすることを表している．なお，出力キーワードと別に割当も同時に求まるが，キーワード選別では割当自体に関心はない．

3.2.4 出力コスト

本節ではキーワード候補 i を出力するための出力コスト f_i について検討するが，便宜上，出力利得を $\acute{f}_i = -f_i$ と置き，出力利得 $\acute{f}_i$ について以下で検討する．

出力利得とは文字通りキーワード候補を出力として選別した際に得られる利得のことであり，単語単体に関する特徴量を用いる既存手法で検討されていた tf_idf 等の，単語への重みに相当すると考えられる．

出力利得として以下の 6 種類を検討する．

- tf
  - 文書内での候補 i の出現頻度．
- tf_idf
  - tf に候補 i の出現文書数を考慮したもの．具体的な定義は文献 [16] に従う．
- tf_pc
  - tf に候補 i の文書内での位置情報 $pc(i,d)$ を掛けたもの．

$$pc(i,d) = 1 - \frac{first(i,d)}{|d|} + \delta \tag{13}$$

ただし，$first(i,d)$ は文書 d 内で候補 i が初出する文の位置．例えば，第 1 文目に i が現れていれば $first(i,d) = 1$ である．$|d|$ は文書 d の文数を表す．δ は $pc(i,d) = 0$ となるのを防止するための微小な値（後述する評価実験では $\delta = 10^{-6}$）である．

- tf_idf_pc
  - tf_idf に $pc(i,d)$ を掛けたもの．
- tf_pi
  - 位置情報 $pi(i_l,d)$ を考慮した tf の変種．次式 (14) で定義される．

$$\sum_{i_l\in ins(i,d)} pi(i_l,d) \tag{14}$$

$$pi(i_l,d) = 1 - \frac{p(i_l,d)}{|d|} + \delta \tag{15}$$

ただし，$ins(i,d)$ は d 内における候補 i の実体の集合，$p(i_l,d)$ は文書 d 内での候補 i の l 個目の実体 i_l を含む文の位置である．なお，$pi(i_l,d) = 1$ で固定すると通常の tf となる．

- tf_idf_pi
  - tf の代わりに式 (14) を用いた tf_idf の変種．

いずれの出力利得 $\acute{f}_i$ においても，実際には $\sum_{i\in\mathcal{D}}|\acute{f}_i| = 1$ を満たすように正規化した値を用いる．

3.2.5 代表コスト

キーワード候補 j を（出力となる）キーワード候補 i で代表させるための代表コスト c_{ij} についても，先と同様，代表利得 $\acute{c_{ij}} = -c_{ij}$ として検討する．

代表利得の性質として，候補 j を候補 i で代表させた際に，候補 j の意味が失われないほど代表利得が高くなると仮定する．言い換えれば，候補 i と候補 j が強い連想関係であるほど代表利得が高くなるとする．

ここでは，二つの候補の共起度が強いほど，一方から他方が連想されやすくなると考え，共起度に基づく以下の代表利得を検討する．

- $P(i|j)$
  - 候補 j のもとでの i と j の共起度．具体的には次式で求める．

$$\frac{\text{文書集合の中で } i \text{ と } j \text{ の両方を含む文の数}}{\text{文書集合の中で } j \text{ を含む文の数}} \tag{16}$$

代表利得 $\acute{c_{ij}} = P(i|j)$ において，実際には $\sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}}|\acute{c_{ij}}| = 1$ を満たすように正規化した値を用いる．

4 評価実験

実験を通して提案手法の性能を評価した．

4.1 実験の設定

評価用データには論文データ（以下，Paper）と新聞記事データ（以下，News）を用いた．両データとも日本語データである．論文データについては，論文著者によって付与されたキーワードを正解キーワードとし，論文アブストラクトから正解キーワードを抽出する評価実験を行った．また，新聞記事のヘッドラインは本文の要約になっていると考えられる．そこで，ヘッドラインから 3.1 節で述べた方法で生成されるキーワード候補を正解キーワードとし，記事本文から正解キーワードを抽出する評価実験を行った．

表 1　評価用データの統計量

	文書数	平均キーワード数	上限値
Paper	100	3.9	0.917
News	100	3.8	0.787

評価尺度には F 値を用いた．表 1 に評価用データの統計量を示す．正解キーワードが抽出元文書に必ず含まれるとは限らないため，再現率の上限値が 1.0 とならず，評価値（F 値）の上限も 1.0 とはならない．

提案手法に含まれる整数計画問題は GLPK パッケージ [7] *を用いて解いた．幸いにして評価用データではすべての事例に対して厳密解を得ることができたが，これは評価用データに含まれる事例（記事）の規模が比較的小さいことに拠る．問題クラスとしては NP 困難であり，一般には何らかの近似手法も検討する必要がある．

抽出キーワード数は $k = 3$ と $k = 5$ を試した．

4.2 実験結果

まず，代表コストを無視し，3.2.4 節で述べた各種の出力コストを単独で使用することで，出力コストの優劣を評価した．具体的には，各種出力コストに基づいてキーワード候補にコストを付与し，コストの低いものから順に k 個の候補を出力キーワードとして抽出した．

結果を表 2 (Paper) 及び表 3 (News) に示す．参考として，Lead 法 [1] に即した手法，具体的には，文書の先頭から順に k 個のキーワード候補を抽出する手法の結果を同表の「*Lead*」に示す．まず，Paper と News に共通することとして，tf や tf_idf は F 値が著しく低く，キーワード候補の文書内での位置情報を考慮すべきであることがわかる．これは，結果全体に対して *Lead* が相対的に高い F 値であることからも示唆される．位置情報を考慮した中では，総じて，tf_pi，tf_idf_pi が tf_pc，tf_idf_pc より

* http://www.gnu.org/software/glpk/

表 2　実験結果（出力コスト単独，Paper）

$\acute{f}_i$	$k=3$	$k=5$
tf	0.056	0.070
tf_idf	0.114	0.170
tf_pc	0.076	0.120
tf_idf_pc	0.181	0.209
tf_pi	0.225	0.238
tf_idf_pi	**0.250**	**0.290**
Lead	0.234	0.259

表 3　実験結果（出力コスト単独，News）

$\acute{f}_i$	$k=3$	$k=5$
tf	0.094	0.103
tf_idf	0.103	0.142
tf_pc	0.142	0.185
tf_idf_pc	0.168	0.201
tf_pi	**0.221**	**0.256**
tf_idf_pi	0.195	0.192
Lead	0.121	0.144

も F 値が高くなっている．

上記結果を踏まえ，以降では最も高い性能を示す出力コストに固定して評価を進める．すなわち，Paper については出力コストを tf_idf_pi に固定し，News については出力コストを tf_pi に固定して評価を進める．

続いて，提案手法の結果を表 4（Paper）及び表 5（News）に示す．表中の「$P(i|j)$」が先で決定した出力コストと代表コスト（$=-P(i|j)$）を提案手法によって組み合わせた場合の結果である．出力コスト単独での最良の結果も表に再掲している．提案手法は代表コスト単独の結果よりも F 値が向上しており，出力コストと代表コストの両者の情報を組み合わせることが性能向上に貢献すること，及び出力コストと代表コストを組み合わせる手法として提案手法が有効に機能していることが伺える．各表の下段は，単語間の共起情報に関する特徴量を用いる従来手法 [9, 10] を本評価データを用いて再評価した結果であり，参考として掲載する．提案手法は従来手法よりも高い F 値を達成していることが確認できる．しかし，従来手法は英語データを対象として提案されている．また，キーワード候補生成の考え方が従来手法と提案手法では異なる等の理由から，各手法の優劣を決定するためには，より詳細な比較調査が必要であると考えられる．

表 4　実験結果（提案手法，Paper）

$\acute{c_{ij}}$	$k=3$	$k=5$
出力コスト単独	0.250	0.290
$P(i\|j)$	**0.278**	**0.330**
Mihalcea et al.[9]	0.128	0.161
Palshikar [10]	0.108	0.132

表 5　実験結果（提案手法，News）

$\acute{c_{ij}}$	$k=3$	$k=5$
出力コスト単独	0.221	0.256
$P(i\|j)$	**0.240**	**0.265**
Mihalcea et al. [9]	0.152	0.172
Palshikar [10]	0.064	0.070

次に，出力コストと代表コストのバランスについて考察する．このバランスの調査のために，式 (6) で示した目的関数を，出力コストと代表コストのバランスを制御するパラメータ λ ($0 \leq \lambda \leq 1$）を導入した式 (17) の目的関数に置き換えた最適化問題（その他の変更なし）を考え，パラメータ λ を 0.1 刻みで変化させた場合のパラメータ値と F 値の関係を調査した．

$$\min \quad (1-\lambda)\sum_{i\in\mathcal{D}} f_i y_i + \lambda \sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} c_{ij}x_{ij} \quad (17)$$

調査結果を図 1（Paper）及び図 2（News）に示す．表 4 及び表 5 で示した提案手法の結果は，図における $\lambda = 0.5$ の結果に対応する．また，$\lambda = 1.0$ の場合は出力コストを無視した設

定となっており，$\lambda = 0$ の場合は代表コストを無視した設定となる．図 1 から，出力コストと代表コストのバランスを制御することで性能向上の可能性があることがわかる．しかし一方で $\lambda = 1.0$ 近辺では出力コスト単独よりも性能が低下する可能性がある．図 1 ほど顕著ではないが，図 2 からも同様の傾向が観測できる．現時点ではバランス・パラメータ λ を最適点へ自動制御する方法が未知である．そのため，安定的な結果を得るには，バランス制御を行わず（すなわち $\lambda = 0.5$ に固定して）運用することが最善であることがわかる．

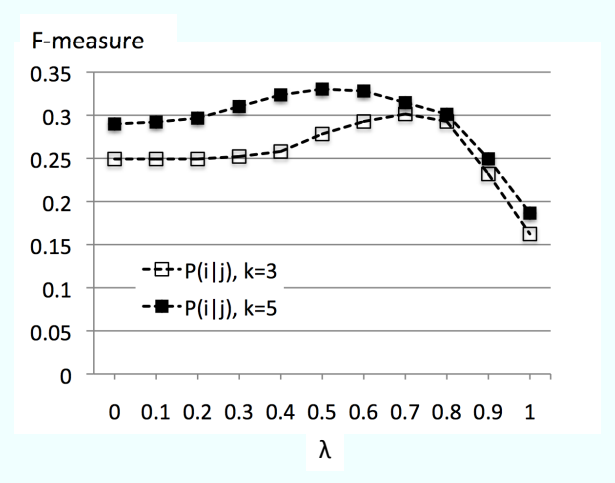

図 1　実験結果（λ と F 値の関係，Paper）

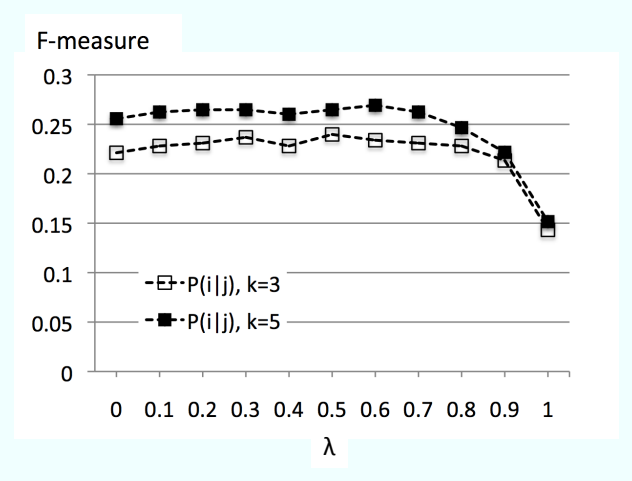

図 2　実験結果（λ と F 値の関係，News）

5　おわりに

本稿では文書からキーワードを自動抽出する手法を提案した．提案手法は教師ありデータを用いない手法に分類される．提案手法の特徴は，単語単体に関する特徴量と，二つの単語間の関係に関する特徴を同時に適切に考慮できる点にあり，これを実現するために，キーワード抽出を，組み合わせ最適化問題の下位問題である施設配置問題の考えから捉え，キーワード抽出課題を整数計画問題として定式化した．評価実験を通して，単語単体に関する特徴量と，二つの単語間の関係に関する特徴を組み合わせることが性能向上に貢献すること，及び組み合わせを実現する方法論として提案手法が有効に機能することを示した．

今後の課題として以下の項目が考えられる．

- 最適な出力キーワード数の自動推定法を検討する．つまり，制約式 (9) を除外する方法を検討する．
- キーワード候補生成をキーワード選別の前処理として位置づけることなく，最適なキーワード・ユニットを決定する方法を検討する．例えば，あらゆる単語 N グラムをキーワード候補とした場合のキーワード選別法を検討する．
- 教師あり機械学習に基づく手法との融合について検討する．例えば，代表コストを教師あり機械学習によって推定することを検討する．

謝辞

本研究の一部は，文部科学省科学技術振興調整費「戦略的研究拠点育成プログラム」の支援の下に実施した．

参考文献

[1] R. Brandowa, K. Mitzeb, and L. F. Rauc. Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, Vol. 31, No. 5,

pp. 675–685, 1995.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN systems*, Vol. 30, No. 1-7, 1998.

[3] G. Ercan and I. Cicekli. Using lexical chains for keyword extraction. *Information Processing & Management*, Vol. 43, No. 6, pp. 1705–1714, 2007.

[4] E. Frank, G. W. Paynter, and I. H. Witten. Domain-specific keyphrase extraction. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pp. 668–673, 1999.

[5] T. Jo, M. Lee, and T. M. Gatton. Keyword extraction from documents using a neural network model. In *Proceedings of the International Conference on Hybrid Information Technology*, pp. 194–197, 2006.

[6] B. Korte and J. Vygen. 組合せ最適化 - 理論とアルゴリズム. シュプリンガー・ジャパン, 2005.

[7] A. Makhorin. Reference manual of gnu linear programming kit,version 4.9. Technical report, 2006.

[8] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. *Japanese Morphological Analyzer ChaSen Users Manual version 2.0.* Technical Report NAIST-IS-TR990123, Nara Institute of Science and Technology Tec hnical Report, 1999.

[9] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing*, 2004.

[10] G. K. Palshikar. Keyword extraction from a single document using centrality measures. In *Proceedings of the 2nd International Conference on Pattern Recognition and Machine Intelligence(LNCS-4815)*, pp. 503–510, 2007.

[11] L. Sujian, W. Houfeng, Y. Shiwen, and X. Chengsheng. News-oriented keyword indexing with maximum entropy principle. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation*, pp. 277–281, 2003.

[12] J. Tang, J. Li, K. Wang, and Y. Cai. Loss minimization based keyword distillation. In *Proceedings of the 6th Asia Pacific Web Conference(LNCS-3007)*, pp. 572–577, 2004.

[13] P. D. Turney. Learning algorithms for keyphrase extraction. *Information Retrieval*, Vol. 2, No. 4, pp. 303–336, 2000.

[14] Xiaojun Wan and Jianguo Xiao. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 2008.

[15] K. Zhang, H. Xu, J. Tang, and J. Li. Keyword extraction using support vector machine. In *Proceedings of the 7th International Conference on Web-Age Information Management(LNCS-4106)*, pp. 85–96, 2006.

[16] 徳永健伸. 情報検索と言語処理. 東京大学出版会, 1999.