産学連携のための教員検索システム

○田中猛彦[1]　平野貴弘[2]　中川優[1]
[1] 和歌山大学システム工学部, [2] 株式会社日本総合研究所

Searching System of Researcher Information for Business-Academia Collaboration

○ Takehiko TANAKA　Takayuki HIRANO　Masaru NAKAGAWA

We developed a searching system which takes a company profile as an input and retrieves the information of papers or researchers. For the system to extract the keywords from company profiles, papers and researcher information, we focus on complex noun phrases included in the documents. Actually, each keyword is weighted not only according to how often it appears, but according to how often the constitutive nouns appear. We took, furthermore, experiments for the comparison between our system and Namazu, a searching system. As a result, our system is 0.76 time as much as Namazu with regard to recall, while it is 3.4 times with regard to precision.

1　はじめに

近年，産学連携の推進が叫ばれており，その政治的・社会的・経済的基盤が整備されつつある．従来型の産学連携は，特定の研究者と特定の企業との間の密接な関係によって維持されてきた．しかしこれでは産業界の活性化にはつながらないため，産学連携の敷居を下げるためのさらなる体制整備が求められている．産学連携支援 Web サイトはその好例である．筆者らも，大阪府泉州地域 (阪南市・泉南市・岬町・忠岡町・熊取町) を対象とし，泉州地域の企業と泉州地域周辺の大学等研究機関 (和歌山大学・近畿職業能力開発大学校など) との産学連携支援システムに携わってきた [1].

このシステム構築に携わる中で，多くの企業が，産学連携を望みながらも，大学がどのような研究を行っているのか認識していないことがわかった．そのような状況でいきなり相談をもちかけるのは困難であり，相談を活発に行うには大学の教員情報を知ることが必要になってくる．現状の産学連携支援 Web サイトで教員情報を検索する際，「カテゴリ検索」や「キーワードによる全文検索」が主流である．これらの手法は企業が教員の詳細情報を獲得したり，相談を持ちかけられた大学窓口機関が相談に適した教員を探したりするのに用いられている．しかしこれらの検索に慣れていない者にとっては，知りたい情報を上手く見つけることができなかったり，逆に必要でない情報まで検出してしまったりするという問題点がある．企業の経営者に限らず，大学の産学連携支援担当の教職員にとっても，産学連携のためのニーズとシーズのマッチングが課題となっており，企業の相談にマッチする教員を探すのに手間がかかることが指摘されている．そこで，企業に関係が深いと思われる技術シーズ情報をシステム側で判断して提示する「教員・論文検索システム」を開発した．

2　システムの概要

産学連携支援システムのサーバには企業のプロフィールやニーズのデータ，研究者の技術シーズデータ（プロフィール・論文）が蓄えられている．本稿で提案する手法は，このニーズデータからキーワードを抽出し，シーズデータと適合度の高いものを選出する．そのため，検索に慣れていな

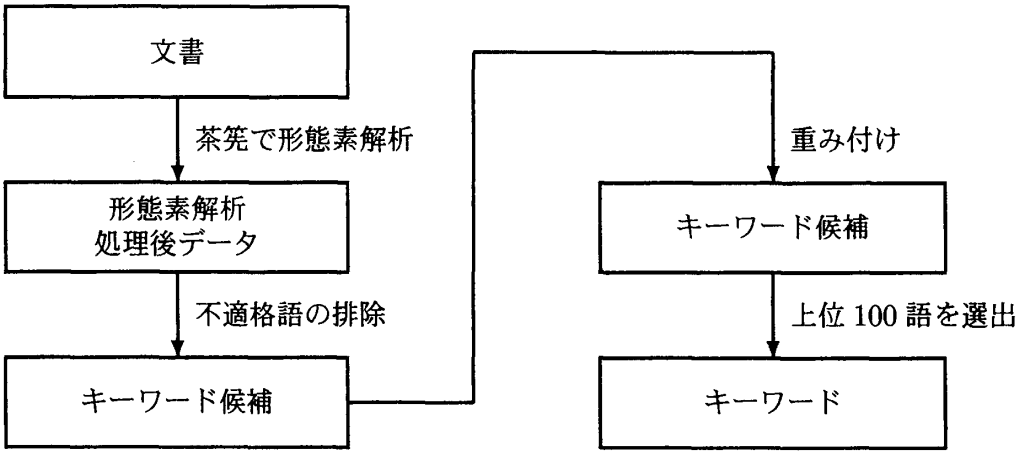

図 1: キーワード抽出の流れ

い者でも，企業プロフィールを用意するだけで検索が行えるというメリットがある．

ここで，キーワードの抽出方法について概略を述べる．ニーズデータでもシーズデータでも同様の方法でキーワードを抽出するので，それら一つ一つを「文書」と呼ぶことにし，文書は日本語で書かれているものとする．最初に，文書を単語ごとに分割するために，茶筅 [2] により形態素解析を行う．その出力を利用してキーワードを抽出する．キーワードになる可能性の高いものは名詞であり，さらに名詞が連続して構成される複合名詞は，重要なキーワードであることが多い．そこで，複合名詞の重要度を高くする (重み付けを行う) ことで，そのデータに対する的確なキーワード抽出を試みた (図 1).

重み付けの尺度には，(1) 複合名詞の長さ (文字数あるいは単名詞数)，(2) 個々の単語の出現頻度，(3) 名詞の文法的性質，およびこれらの組み合わせが考えられる．最近になり，複合名詞を構成する個々の単語がどれだけ多くの名詞と連接して複合語を作るかを尺度とした方法 (出現頻度と連接頻度による専門用語抽出)[3] が提案された．これは単名詞 “N” が対象分野の重要な概念を表しているなら，書き手は N を単独で頻繁に使うのみならず，新規な概念を表す表現として N を含む複合名詞を作りだすことも多いという特徴を利用したものであり，本研究でもこの方法をベースに重み付けを試みた．ただし，以下の語はキーワードになり得ないものとして，重み付けをする前に排除した．

1. 「名詞-一般」，「名詞-サ変接続」，「名詞-固有名詞-組織」，「名詞-固有名詞-地域-一般」，「名詞-固有名詞-一般」のいずれかの品詞分類を含まない，単名詞もしくは複合名詞 [5]
2. 2 文字以下の単名詞
3. 10 以上の (産学連携支援サーバにある) 文書で出現する語

その上で，[4] に記した重み付け方法により複合名詞ごとにスコア (正の数) を求め，スコアの高いものから順に最大で 100 語を，その文書のキーワードとした．

3 評価

研究室の Windows 計算機 1 台を産学連携支援サーバとして，論文データ 300 報 (経済系の論文と工学系の論文を中心に，インターネット上から入手できる論文を集めた) を格納し，企業データ

表 1: 論文データの再現率と適合率

企業	全適合文書数	再現率		適合率	
		提案手法	Namazu	提案手法	Namazu
A(電気機器)	6	0.83	1.00	0.46	0.12
B(電気機器)	3	0.67	1.00	0.18	0.08
C(ゴム)	2	1.00	1.00	0.40	0.08
D(一般機械)	4	0.75	1.00	0.43	0.17
E(家具)	1	1.00	1.00	0.25	0.04
F(化学工業)	6	0.67	1.00	0.27	0.15
G(金属)	2	0.50	1.00	1.00	0.10
H(木材)	3	0.67	1.00	0.29	0.09
I(精密機械)	3	0.67	1.00	0.33	0.15
J(その他)	6	0.33	1.00	0.31	0.16

から検索を行った.

検索エンジンの能力を評価するのは再現率と適合率の二つの尺度が使われており，本研究でもこれらに基づいて，前節で述べた方法 (提案手法) と，フリーソフトウェアによる検索エンジンとして有名な Namazu (`http://www.namazu.org/`) とで比較を行った．再現率と適合率の計算式は次の通りである.

$$\text{再現率} = \text{検索された適合文書数} \,/\, \text{文書集合中の全適合文書数}$$
$$\text{適合率} = \text{検索された適合文書数} \,/\, \text{検索結果の文書数}$$

インターネットで入手した企業データ 10 件のキーワードを用いて，論文データに対して検索を行った．それらの再現率および適合率を，表 1 に示す.

すべての単語をインデックス化している Namazu と比べると，本システムが再現率において Namazu を超えることは不可能であるので，再現率においては Namazu に近づくほどよい成果であると言える．つまり，Namazu の値に近づくほど検索漏れが無かったと言える．この実験において，本システムにおける企業データ 10 件に対する再現率の平均は 0.76 であった．次に適合率についてであるが，論文データは大量の索引語を生成するので，Namazu による検索では不適合文書の検索をしてしまう．Namazu の平均適合率は 0.114，本システムの平均適合率は 0.392 となった.

今回は，検索対象となるデータ数が少なかったために，一つの重要なキーワードを逃すと結果に大きな影響を与えた．論文データ検索に関しては，本システムのようにキーワード抽出の段階で不適確なキーワードを排除することにより，検索漏れが 4 つに 1 つ程度で適合率を 3.4 倍にすることができた.

さらに，和歌山大学の教員データ 261 件についても，同様に企業データから検索を行った (表 2) が，適合数が非常に少なく，結果はよくなかった．この原因として，データ件数の不足とデータ内容の不足が考えられる．和歌山大学の教員は約 300 人であり (さらに，企業との研究開発実績が多い工学系学部については教員数 91 人であった)，これは 1,800〜1,900 人の教員を擁する大阪大学，京都大学など他の総合大学と比較して少ない．また，教員データの内容は，技術の詳細部まで解説しているようなプロフィールではなく，幅広い意味を持つ複合名詞が出現することが多いので適合数が少なかったと推測される.

表 2: 教員データの再現率と適合率

企業	全適合	再現率		適合率	
	文書数	提案手法	Namazu	提案手法	Namazu
A(電気機器)	3	0.67	1.00	1.00	0.33
B(電気機器)	0	-	-	-	-
C(ゴム)	1	1.00	1.00	0.50	0.50
D(一般機械)	1	0.75	1.00	1.00	1.00
E(家具)	0	-	-	-	-
F(化学工業)	1	0.00	1.00	-	0.33
G(金属)	1	1.00	1.00	1.00	0.25
H(木材)	1	0.00	1.00	-	0.08
I(精密機械)	0	-	-	-	-
J(その他)	0	-	-	-	-

4 おわりに

本稿では，企業のニーズと大学のシーズのマッチングが手軽に行えるような検索方式を提案し，その評価を述べた．この検索システムは，企業の検索意図をあらかじめサーバのデータベースに蓄えておくことで，システム側から検索を行う．そのため，キーワード検索に慣れていない企業でも，自社情報を登録するだけで適切な検索結果を獲得することが期待できる．また，企業は定期的に Web ページにアクセスして検索を行わなくても，新着情報に対してサーバ側で自動的にマッチするかどうかを判断し，マッチしていれば企業にシステム側から通知することが可能である．

一般に教員や企業のプロフィールは，論文データより文字数が少なく，それに起因する適合率の低さも見られた．特に数百字程度のプロフィールでは，排除しきれなかったキーワードとしてふさわしくない名詞や，逆に排除してしまった的確な名詞を見極めることでシステムの適合率向上につながると考えられる．本手法は自動でキーワードを抽出しているが，途中段階でキーワード候補の追加・削除・重み付け変更ができる「キーワード支援システム」というのも，教員や企業の自己主張の方法として有益な機能となるかもしれない．

参考文献

[1] 阪南ブロック商工会 (忠岡町商工会，熊取町商工会，泉南市商工会，阪南市商工会，岬町商工会): 広域連携産学交流事業.

[2] 形態素解析システム『茶筌』version2.2.9 使用説明書，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座 (松本研究室).

[3] 湯本紘彰; 森辰則; 中川裕志: 出現頻度と連接頻度に基づく専門用語抽出，情報処理学会研究報告「自然言語処理」，No.145，2001.

[4] 平野貴弘; 田中猛彦; 中川優: 産学連携のためのデータベースシステム，第九回社会情報システム学シンポジウム，2003.

[5] 今井直基: 電子掲示板におけるキーワード提示型記事検索システム，和歌山大学大学院システム工学研究科修士論文，2002.