2004－DBS－134 (II) (77)
2004／7／14

産学官連携支援のための研究者情報システム

Chhy Huy[†]　谷口　伸一[†]

あらまし　本論文は大学の研究シーズと企業のニーズをマッチングするコーディネータのために研究開発した「研究者情報システム」について論じる．これまでにも地方自体や外郭団体により研究者データベースは構築されてきているが，それらの多くは人手による情報収集のため更新の困難さや網羅率の低さあるいは検索インターフェースが十分でなかった．本研究は滋賀県からの委託研究によるため滋賀県内の 10 大学を対象にしているが，情報収集の自動化，TF-IDF とベクトル空間モデルに基づく類似度計算およびシソーラス辞書を活用した 6 種類の検索方式を実現し有効かつ実用性の高いシステムとした．また，コーディネータが絞り込んだ研究者の大学窓口とワンクリックで連絡可能なテレビ電話会議システムを組み込み”face to face”のコミュニケーションを実現した．

The Researcher Information System for Industry-Academia and Government Cooperation (IAGC) Support

Huy CHHY[†]　Shinichi TANIGUCHI[†]

E-mail : m03217hc@econ.shiga-u.ac.jp , taniac@biwako.shiga-u.ac.jp

Abstract　This paper describes about “The Researcher Information System” which is developed for coordinator who tries to match the research seed and the need of enterprises. So far local authorities and the auxiliary organization have constructed the researcher database. However, because most information is gathered by hand, it has some difficulties in updating and has incomplete data coverage or imperfect search interface. As this research is sponsored by Shiga prefecture, only 10 universities in Shiga are targeted. But it is a useful and practical system, which was realized by automatic data gathering, using thesaurus dictionary, calculating similarity based on TF-IDF and vector space model, and 6 types of search methods. And this system also realized the “face to face” communication, which is the videoconference system where the coordinator can contact only in one click with university’s staff that is a representative of its university researcher.

[†]滋賀大学大学院経済学研究科,彦根市
[†]Graduate School of Economics ,
Shiga University , Hikone
E-mail : m03217hc@econ.shiga-u.ac.jp
taniac@biwako.shiga-u.ac.jp

1．まえがき

2002 年の施政方針で「知的財産の戦略的な保護・活用」が国家目標として表明された．これ以降，「知的財産立国」に向けた政府の取組みが矢継ぎ早に行われている[1]．その中で「産学官連携」

の取り組みは特に活発化している. 産学官連携は共同研究などの交流を通じて大学等研究機関の技術シーズを民間企業において事業化する営みである. 産学官連携を推進することにより, 資金, 設備, 研究開発などの資源が脆弱な企業においても, 外部資源を活用した製品開発が可能になる. 一方, 大学等研究機関においても産業界のニーズを的確に把握した研究遂行と外部資金獲得が可能になる. その仲介役の一つとして TLO (Technology Licensing Organization) が注目されている. TLO は研究成果を特許化し, その技術を企業に移転して事業化するとともに, その対価を大学等研究機関へ還元して, 新たな研究成果を生み出すことを推進する機関である（図１）. その仕組みの中で目利きとしてのコーディネータの役割が重要になってきている.

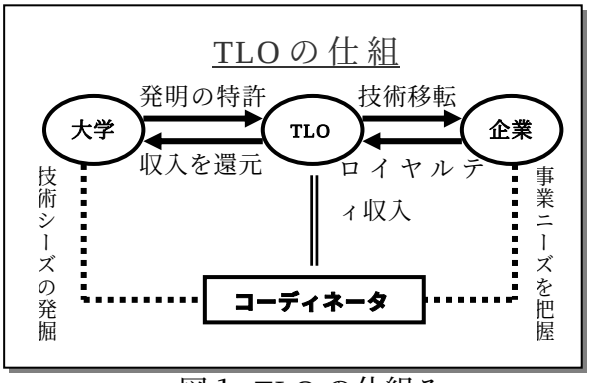

図１ TLO の仕組み

産学官連携の取り組みとして, 滋賀県は滋賀大学との共同研究により TLO の設立を含めた知的財産戦略に関する包括的な検討を行った. そのなかで国からの助成は 5 年間という制約があるため [2], それ以降の継続的な活動が問題視された. 事実, 2003 年 12 月時点で 41 ある TL0 のうち収支が黒字化しているのは数機関である. そこで, 共同研究報告書[3]ではハードとしての TLO の設立より, コーディネータや産学連携担当者らが, 大学のシーズと企業のニーズをマッチングする上で重要な情報源となる研究者情報データベースの一元化と産学官連携の環境整備というソフトの充実を提言した.

本稿は, この提言を受けて滋賀県より委託された滋賀県研究者情報データベースと機能的な産学官連携を支援する情報システムについて述べ, 本分野の他の例と比較して本システムの特長と有効性を報告するものである.

2. 研究者情報データベースシステムの設計

2.1 情報収集と索引語の抽出

本システムは Web コンテンツ収集プログラム cURL[4]と形態素解析ツール茶筌[5]を実行させて, 情報収集と索引語の抽出およびそれによるデータベース更新を自動化している.

cURL で自動収集した研究者ごとの Web コンテンツは, 全文を茶筌で形態素解析して, 索引語が抽出される. ここで, Web コンテンツの意味を正しく反映するためには形態素の連続可能性を判断することが重要になる. 例えば,「青色発光ダイオード」が出現する文書中で, これが重要なキーワードである場合, 形態素解析に基づいて「青色」,「発行」,「ダイオード」という索引語を生成しただけでは, この文書にとって重要なキーワードを失うことになる. その解決法として茶筌では, 品詞間の連続可能性を統計学習して連続規則表を生成する辞書を組み込んでいる[6].

しかし,「青色発光ダイオード」などの専門用語は茶筌でも解決できないため, 本システムでは N グラム索引法 (N-gram indexing)を形態素単位に適用して索引語（これを組み合わせ索引語と呼ぶ）を生成している. また, 助詞「の」で連接する二つ以上の形態素も組み合

わせ索引語とするルールを組み込んでいる．また，英文に対しては Cornell 大学の SMART システム[7]で採用されている不要語リスト(stop words list)を用いて冠詞や前置詞等の不要語を索引語の候補から除外している．

2.2 TF-IDF およびベクトル空間モデルによる類似度計算

研究者 Web コンテンツより抽出される索引語は，研究者との関連性に応じて重要性が変化する．この重要性を数値化するのが索引語の重み付け(term weight)である．索引語の重み付けは，次の再現率または適合率を向上させる重み付け法である[8]．

- 再現率(recall)：完全性を評価するための尺度であり，検索対象となる文書集合の中の検索質問に適合する文書のうち，実際に検索された文書の割合を示す．つまり検索漏れの少なさを示す尺度である．
- 適合率(precision)：正確性を評価するための尺度であり，検索された文書集合の中で，検索質問に適合する文書の割合を示す．つまり検索ノイズの少なさを示す尺度である．

文書集合をD，Dから抽出された索引語集合をTとする．それぞれの要素を$d_i(i=1…n)$，$t_j(j=1…m)$と記す．文書d_iにおける索引語t_jの重みW_{ij}は，一般に以下のような三つの指標で数値化できる．

(1) 局所的重み(local weight)：文書d_iにおける索引語t_jの出現頻度f_{ij}により計算される重みである．この重みは再現率の向上に寄与する．局所的重みの計算式として，次の二つに着目する．
   a. 索引語頻度(Term Frequency) TF：$l_{ij}=f_{ij}$
   b. 対数化索引語頻度(logarithmic term frequency)LTF：$Ll_{ij}=\log(1+f_{ij})$

(2) 大域的重み(global weight)：文書集合Dにわたる索引語t_jの分布を考慮して決定される重みである．適合率向上を目的とした重みであり，特定の文書に集中して出現する索引語に対して大きな値が与えられる．nを全文書数，n_iを索引語t_jが出現する文書数とするとき，大域的重みの計算式として，次の二つに着目する．
   a. 文書頻度の逆数(inverse document frequency) IDF：$g_i=\log(n/n_i)$
   b. 確率的IDF(probabilistic IDF) PIDF：$Pg_i=\log((n-n_i)/n_i)$

(3) 文書正規化係数n_j (document normalization factor)：文書が長くなるにつれて，その中に含まれる索引語の数も増えるため，長い文書に含まれる索引語の方が大きな重みを持つようになる．文書正規化係数は，文書の長さによる影響をなくす目的で導入されるものである．しかし，本システムが扱うWebコンテンツ間の文字数にはそれほど大きな差がないこと，また，長いコンテンツは自己ＰＲ(産学官連携)に積極的と見なすことができるため，この正規化を行わないことにする．

以上より，産学連携の機会公平性保証のために検索漏れを少なくするよう再現率を重視した索引語の重みw_{ij}を次式で求める．

$$w_{ij}=l_{ij} \times g_i \qquad (1)$$

次に，ベクトル空間モデル(vector space model) [9]による類似度計算について述べる．

文書d_i に出現する語t_jの重みをw_{ij}として，文書d_iの索引語表現d_iをM次元ベ

クトル$d_i=(w_{i1}, w_{i2}, \ldots, w_{iM})^T$として表す．ここで，$T$は転置行列を表す．同様に，検索質問qに出現する語$t_j$の重みを$w_{qj}$とすれば，検索質問qもM次元ベクトル$q=(w_{q1}, w_{q2}, \ldots, w_{qM})^T$として表現可能であり，検索質問qに対する文書$d_i$の類似度はベクトル空間中のベクトル$d_i$と$q$との類似度(similarity)として計算できる[10]．ベクトル間の類似度の尺度としてベクトルの内積(inner product)や角度の逆関数(inverse function of the angle)が提案されている[9]．文書検索において一般に用いられるのは内積や次式で定義されるコサイン尺度である[8]．

$$\cos(d_i, q) = \frac{d_i^T \cdot q}{\|d_i\| \cdot \|q\|} = \frac{\sum_{j=1}^{M} w_{ij} w_{qj}}{\sqrt{\sum_{j=1}^{M} w_{ij}^2} \sqrt{\sum_{j=1}^{M} w_{qj}^2}} \quad (2)$$

一方，内積は(2)式の分子の計算式で求められる．

2.3 検索インターフェース

検索インターフェースの善し悪しは，情報検索システムの評価基準の一つである．そこで本システムは利用者の習熟度に応じた 6 種類の検索インターフェースを考案した．

(1) Yahoo!/Google 方式：通常利用する検索エンジンのように検索キーワードを入力するだけの方式である．一般利用者は，まずこの検索方式を利用すると考える．

(2) Yahoo!/Google 方式＋シソーラス：シソーラスを利用することで検索キーワードの表記の揺れを補正し，さらにその意味概念を拡張することができる．

(3) キーワード偏重方式：二つ以上の検索キーワードが入力される場合，最初に入力するキーワードほど利用者にとって重要であることが考えられる．そこで，キーワードに重要度を与えることで検索質問ベクトルの索引語の重みを調整することを考案した．検索ベクトルの索引語の重みは重要度が「大」の時はそのまま，「中」の時は 2/3，「小」の時は 1/3 とする(図 2)．

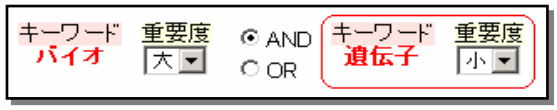

図 2　キーワード偏重方式

(4) キーワード偏重方式 ＋シソーラス：検索キーワード（見出し語）とそのシソーラスの重要度は同じとする．さらに，図 3 のように，表示されたシソーラスから新たなキーワードを連想することも考えられる．そこで，追加語の入力を可能にしている．コーディネータに限定されるが，この追加語は検索完了後にシソーラス辞書に登録することができる．これにより，シソーラス辞書が本システムにとって一層適切な辞書へと学習するとともに，「学術用語と産業用語の対応づけ」ができるようになる．

図 3 シソーラスとキーワード連想

(5) フリーワード検索：フリーワードの質問文は形態素解析および N グラム索引法により検索キーワードに展開される．展開されたキーワードから検索に利用する語を利用者が選択する方式とした．

(6) フリーワード検索＋シソーラス：

(5)で選択された語に対してシソーラスの利用を可能にする.

2.4 テレビ電話会議システム

コーディネータの活動障害の一つに,大学研究者との連絡が困難であることが挙げられている.そこで,本システムは検索された研究者が所属する大学の産学連携窓口とワンクリックでテレビ電話会議システム[11]により連絡できる方式を実現している.

図４検索結果画面

図４のように検索された研究者名の右端が「*online*」と表示されているときは,産学連携窓口のテレビ電話会議システムが稼働中であることを示す.クリックするだけで先方とつながり対面会話ができる.資料等を提示しながら情報交換できるとともに,相互に信頼感が生まれ,円滑な連携が期待される（図５）.

図５V-talk REBIRTH の画面例

3．本システムの特長と有効性

3.1 Web 検索エンジンとの比較

Web 検索では Yahoo!や Google が広く利用され,その性能には定評がある.そこで,“滋賀県内の遺伝子やバイオ関係の大学研究者”を検索する場合について本システムと比較する(表１).

検索キーワード	Yahoo!	Google	本システム
滋賀 大学研究者 バイオ 遺伝子	0 名 /25 件	2(2)名 /690 件	15 名
滋賀 大学研究者 バイオ 遺伝子 バイオテクノロジー 生命工学	0 件	1(1)名 /132 件	31 名

表１( )内の数字は滋賀県内に大学がある研究者,該当検索結果は 5 画面

表１から商用検索エンジンでは適切な結果が得られないことが分かる.さらに,本システムが装備するシソーラス(thesaurus)のうち同義語を併用することで「バイオテクノロジー」や「生命工学」などがキーワードに付加され 31 名が検索できた.

このように,多種多様な Web コンテンツをデータベース化している検索エンジンを利用して研究者情報を見出すことは困難であり,本システムの有効性が認められる.

3.2 他の研究者データベースとの比較

地方自治体あるいは外郭団体等が作成している研究者データベース 3 例と科学技術振興機構(JST)の ReaD に対してデータ項目数,検索方法,表示方法,更新頻度に関して比較する(表２).

上記の研究者データベースは共通し

データベース名	作成団体	データ項目数	検索方法	表示方法	更新
岐阜県内研究者データベース	岐阜県研究開発財団	11 項目	キーワード，所属，分野，学位	共通テンプレート	不明
京都地域大学研究者データベース	(財)大学コンソーシアム京都	ReaD と各大学へのリンク			
長崎県研究者データベース	長崎県商工労働部産業振興課	16 項目	所属，研究者名，キーワード	共通テンプレート	不明
研究開発総合支援ディレクトリ	科学技術振興機構(ReaD)	12 項目	研究者，研究課題，研究資源	共通テンプレート	年 1 回

表 2 各研究者データベースと比較

て以下の特徴をもつ．

(1) データの収集は調査票にて行い，データベースへ入力している．データ内容は連携分野などのように産学連携を意識しているものと研究者の研究分野の紹介に止まるものに分かれる．しかし前者は少数であり，また，すべての研究者が産学連携の取り組みについて記入している状況でない

(2) 検索方式はキーワード，機関別，分野別がほとんどである．利用者の検索要求を十分満足させるものとはいえない．

(3) 表示方法は，それぞれのテンプレートにより表示している．

(4) 更新頻度は ReaD の 1 年を謳っているもの以外は不明もしくは更新が行われていない．上記 4 例以外も調査したところデータベース作成初年度のままになっているところが見受けられた．また，ReaD においても情報提供した機関の研究者がすべて登録されていないことがわかった．

以上のことから，独自の調査票に基づきデータベースを構築する方式では，データ更新の困難さが大きな問題点として指摘できる．

一方，これらに対する本システムの有効性および実用性を列挙すると以下となる．

(1) 滋賀県内の大学が発信する研究者 Web コンテンツを検索ロボットで自動収集し，これを形態素解析して索引語からなる研究者データベースを自動作成するため，常に鮮度の高い情報が提供できる．

(2) TF-IDF とベクトル空間モデルにより，検索キーワードと研究者情報との類似度に基づくランキング表示が可能で適合度の高い研究者から効率的に閲覧できる．

(3) 大学・学部別，キーワード，フリーワードによる六つの検索方式を提供している．これらは利用者の慣れに応じて利用できるようになっている．

(4)シソーラス辞書を搭載しており検索キーワードの意味概念を拡張して，広い観点から漏れの少ない検索できる．また，学術用語に対する産業用語を辞書に登録することができるため，コーディネータや一般利用者が産業用語で検索することができる．

(5) 複合キーワードによる検索では，それぞれのキーワードの重要度を考慮したキーワード偏重方式を考案している．また，キーワードの複合条件演算子(AND/OR)は，検索ステップの途中で柔軟に指定できるようにしている．

(6) 研究者情報を表示するテンプレートを持たず，適合度と研究内容を参考にして研究者が所属する大学の当該研究者のWebコンテンツへリンクして詳細情報を得る方式としている．このことはそれぞれの大学の産学官連携への取り組み姿勢を反映するもので，経済産業省が要求している「産学連携を考慮した大学のホームページ」作りを後押しすることに寄与する．

(7) 検索された研究者と円滑な交渉ができるように，各大学の産学連携窓口とテレビ電話会議で連絡調整できるようにした．

4. むすび

本論文では滋賀県から委託された『滋賀県研究者情報データベースシステム』について，理論的な考察とそれに基づいたシステムの設計・開発について述べた．そして，これまでに開発されている同様の研究者データベースと比較して以下のような有効性と実用性を示した．

(1) 各大学が公式に提供している研究者情報を検索ロボットで自動収集するため，人手による情報の収集と入力の必要がなく，常に鮮度の高い情報提供ができる．

(2) TF-IDFとベクトル空間モデルにより，検索質問に対する研究者情報の類似度を求めてランキング表示させ，利用者が効率よく研究者を検索できる．

(3) シソーラスを搭載することで検索質問の概念拡張を実現し，再現率を向上させている．また，シソーラス辞書の編集機能により，コーディネータからの強い要望である学術用語と産業用語の対応付けを実現している．

(4) 6種類の検索インターフェースを考案した．特に，キーワード偏重方式という検索質問に含まれる利用者がもつキーワードの重要性を反映した方式を提案した．

(5) ワンクリックでつながるテレビ電話会議システムという双方向コミュニケーション・ツールを提供した．

今後の課題としては，研究者のWebコンテンツからより多くの有効な索引語を生成するために単語の関連性を考慮した索引語抽出方式を考案する．また，検索システムの精度を客観的に検証するために，Webコンテンツに基づくテストコレクションを作成する．さらに，特許流通アドバイザやコーディネータが取得した情報を共有化し，効率的，効果的な情報活用ができる仕組みを実装する．

謝辞　滋賀県商工観光労働部ならびに(財)滋賀総合研究所諸氏，特に中後康氏ならびに奥野修氏には利用者の観点から貴重な意見を頂戴した．衷心より感謝する．また．数回のデモンストレーションにあたり貴重な意見を頂いたコー

ディネータ諸氏に感謝申し上げる.

参考文献

[1] “知的財産立国を目指して”, 内閣官房知的財産戦略推進事務局, http://www.ipr.go.jp/intro2.html
[2] “日本経済新聞”, 2004 年 6 月 1 日
[3] 大村和夫, 吉田慶志 : “滋賀県内中小企業知的財産権の創造・保護・活用策”, 滋賀県, 2002
[4] “cURL and libcurl”, cURL, http://curl.haxx.se/
[5] 松本裕治 : “茶筌”, http://chasen.aist-nara.ac.jp/hiki/ChaSen/
[6] 松本裕治 : “形態素解析システム「茶筌」”, 情報処理, Vol41,No.11, 2000, pp1208-121
[7] “SMART システムの不要語リスト”, ftp://ftp.cs.cornell.edu/pub/smart/ english.stop
[8] 北研二, 津田和彦, 獅子堀正幹 : “情報検索アルゴリズム”, 共立出版, 2002
[9] 岸田和明 : “情報検索の理論と技術”, 勁草書房, 1998
[10] Salton,G., Wong,A. and Yang,CS.: “A Vector Space Model for automatic indexing”, , Communication of the ACM, Vol.18, No.11, pp.613-620, 1975, http://trec.nist.gov/pubs/trec4/t4_proceedings.html
[11] “V-Talk REBIRTH”, Media Business Network Co.,Ltd , http://www.mbn.co.jp