キーワード抽出を用いた就職活動支援システム

前山 侑平† 安留 誠吾‡

大阪工業大学 情報科学研究科† 大阪工業大学 情報科学部‡

1. はじめに

現在の就職活動では、学生はインターネットから収集した企業情報を、志望企業を選ぶ際の一つの情報源として用いている。しかし、企業情報は企業ホームページ内でも分散しており、大企業ともなれば 100 近いページから情報収集、精査を行わなくてはならない。そのため、志望動機を作成するための情報収集、精査に時間を費やす必要がある。また、収集した情報を同業他社間で比較するのが一般的な企業比較の方法だが、学生が知らない企業もかなりあるため、同業他社を探し出して比較するのに時間がかかり、学生の志望企業決定、ひいては就職活動そのものが遅れやすいという問題がある。

就職活動を支援するシステムとしてはスケジュール支援[1]、OB・OG とのパーソナルコネクションの活用[2]といった、就職活動中の学生支援を主な目的としている研究がある。本研究では就職活動の前段階である企業比較・志望動機作成に特化した Web コンテンツを提供することで、問題の解決を図る。

2. システムの概要

本研究では次の二つのコンテンツを作成した。

・ 同業他社比較ページ

学生の代わりに同業他社をシステム側で選択する機能を実装し、比較を行いやすい表形式で企業情報を表示することで、同業他社の比較作業を効率化することを目的としている。

・ キーワードグラフ

志望動機に活用できるキーワードを抽出し、ホームページの構造と対応したグラフを表示することで、志望動機の作成までの作業時間の短縮化を目的としている。

3. 実装

3.1. 対象企業の取得

就職活動の対象となる企業のキーワード抽出と企業情報の収集を行うため、Web ブラウザとその拡張機能を用いて、対象となる企業の企業名と URL を収集した。利用するブラウザは Mozilla Foundation が公開している Firefox、拡張機能は Greasemonkey である。この拡張機能は任意のユーザスクリプトを登録することで、指定ドメイン・URL で新たな動作を指定するものである。今回作成したユーザスクリプトは、特定の就職支援サイトにアクセスした際、HTML から企業名、企業の URL、就職支援サイトの URL を抽出し、データベースに保存する。

3.2. 同業他社比較ページ

収集した企業の URL を基に就職支援サイトから企業情報を収集し、比較を行いやすいように表形式で表示している(図 1)。表示するのは、事業内容、従業員数、資本金、職種、勤務地、福利厚生などの情報である。比較企業は学生が一覧から選択するほかに、特定企業の同業他社をシステム側が選択できる。これは Google AJAX Search API を用いて取得した類似ページ検索結果から抽出した企業の URL を活用している。これにより、同業他社発見までの時間を短縮し、素早く比較・分析することができる。

キーワード	• システムエンジニア/System Expert • EPMソリューションセンター • ソフトウェア開発エンジニア/Software Expert • マーケティングプロモーションセンタ • ソリューション・エキスパート集団	• シェートハック • リアルタイム • ソフトウェアデベロップメント • ビジネス推進 • 知名度
業種	[業種]情報処理（関連業種）ソフトウェア/インターネット関連	[業種]情報処理
従業員数	5,215名(2009年3月末日現在 連結)	5,283名(2009年3月末)
資本金	41億9,000万円(2009年3月末日現在)	341億8200万円
事業所	本社/東京(品川) 支社/中部(名古屋)、関西(大阪) 及びその他事業所	東京(本社)、札幌、弘前、横浜、金沢、名古屋、福岡
勤務地	東京(本社) 各支社(名古屋、大阪) 各事業所	主として本社地区(東京、神奈川)

図 1. 同業他社比較ページ

3.3. キーワード抽出

収集した企業の URL を基に、企業ホームページの HTML ファイルをローカルに収集する。この HTML ファイルから、各企業の志望動機作成に適したキーワード抽出を行い、そのキーワードの重要度をスコアとして算出する。抽出手法としては、HTML ファイルから抽出した文章を適切に区切り、形態素解析用ソフト CaboCha[3]を用いて解析してキーワードを抽出する。キーワードのスコアは、以下の項目から算出される。

・ HTML 構造的特徴、文字列的特徴

キーワード抽出時、それを囲んでいる HTML タグの種類によってスコアを変動させる。たとえば見出しに用いられる h1～h6 タグなどで囲まれているキーワードはスコアを上げる。また、キ

Job hunting support system using keyword extraction
†Yuuhei Maeyama – Osaka Institute of Technology
Graduate School of Information Science and Technology
‡Seigo YASUTOME – Osaka Institute of Technology
Faculty of Information Science and Technology

ーワードを構成する名詞の平均的な文字列長を割り出し、この値が高ければ高いほどスコアを上げる。これは、IT 系企業などでよく用いられる企業独自の造語をランキングの上位にするための計算である。

・ Google での検索結果件数

Google AJAX Search API を用いてキーワードの検索結果件数を問い合わせる。この検索件数が少なければ少ないほどスコアを上げる。これは、そのキーワードのオリジナリティが高いものをランキングの上位にするための計算である。

・ 単語の意味、特徴

キーワードに"システム"などの IT 用語に含まれやすい名詞を含む場合、英語を含む場合、キーワードの説明文中に"世界 No.1"、"日本一"などの文字列を含む場合、キーワードのスコアを上げる。

以上 3 つの項目で算出した値を基にキーワードの最終的なスコアを計算し、キーワードとスコアをデータベースに保存する。

3.4. キーワードグラフ

抽出したキーワードを用いて、志望動機の作成を支援するグラフを生成する。グラフの作成には jsViz[4]という SVG（Scalable Vector Graphics）形式でブラウザ上にツリーグラフを描画する JavaScript ライブラリを用い、描画用に生成した XML を読み込んでキーワードグラフを描画している(図 2)。

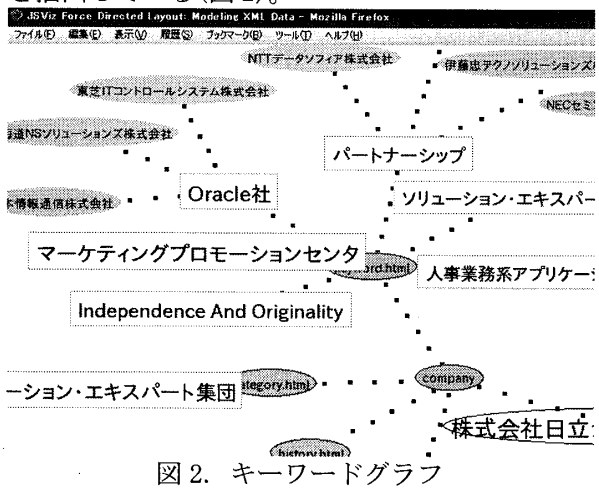

図 2. キーワードグラフ

グラフには 4 種類のノードが存在し、対象としている企業名、URL（階層ごとに分割したもの）、キーワード、同じキーワードを持つ企業名に分類される。また、キーワードノードはスコアに応じてサイズが変わる。各ノードは対象としている企業名ノードから次のようにリンクしている

・ 対象としている企業名 - URL

企業ホームページの階層構造を表現している。

・ URL - キーワード

キーワードがどのページに存在するか表現することで、重要なキーワードがどのページ集中しているかを見つけ出し、企業の傾向を分析することを可能とする。

・ キーワード - 同じキーワードを持つ企業名

同じキーワードを持つ企業を表示することで、同業他社の発見をサポートし、そのキーワードが対象としている企業独自のものかを見分けることも可能となる。

各ノードは情報収集の効率化のため、マウス操作に応じた機能を持つ。キーワードにマウスをあわせた場合はキーワードを含む文章など追加情報の小ウィンドウ表示、クリックした場合はキーワードが存在する企業ページへ移動、などの動作である。

4. まとめ

本研究で作成したシステムにより、同業他社の発見と企業比較が容易になり、志望企業決定までの作業の効率化が期待できる。キーワード抽出とグラフ化により企業分析を容易化し、志望動機の作成への作業時間の短縮化が期待できる。今後の課題としては、キーワード抽出の精度向上、本システムのコンテンツ追加・増強などが挙げられる。

参考文献

1. 山口 賢治、古井 陽之助、速水 治夫：“就職活動におけるスケジュール管理ソフトの提案”，情報処理学会研究報告. GN, [グループウェアとネットワークサービス] 2007 (32) pp.121-126, (2007-3).
2. 長谷川 忍、高橋 咲江、柏原 昭博：“就職活動に有用なパーソナルコネクション構築を促進する SNS の開発”，教育システム情報学会 2009 年度第 1 回研究会, pp.38-43 (2009-5).
3. CaboCha/ 南 瓜： Yet Another Japanese Dependency Structure Analyzer 工藤 拓著 (http://www.chasen.org/~taku/software/cabocha/).
4. jsViz.org :: blog (http://www.jsviz.org/blog/).