

WebAPI を用いた就職活動のための企業情報収集システム構築

著者 田中健太郎

指導教員 斉藤徹

1. はじめに

最近の就活難な状況の中、学生の企業のミスマッチによる早期離職になる場合が増加している。一方、学校に届く求人票では企業の実情などが分らないので、学生は求人票だけでなく、就活サイトをはじめとしたインターネットの情報も活用して企業調査を行うことが重要である。これの手助けを行うために、学校に求人票を出した企業について、WebAPI などを用いてインターネットで検索し、得られた口コミ情報をデータベースに蓄積していくという処理を自動で行うシステムを作る。そして将来的には、企業に合わせて情報を検索できるシステムとして学生に利用してもらうことを本研究の目標としており、企業情報の収集と蓄積を自動で行い、ユーザの入力に合わせて情報を表示するところまで完成している。

2. 研究背景

2.1 研究の概要

求人票は業務内容や賃金などの労働条件が明記されている。しかし、企業の評判、実績などは求人票からは分からない。また求人票の数も多く 1 つ 1 つじっくり調べていたらかなり時間がかかる。そこで、インターネットによる企業調査を自動で行ってくれるようなシステムがあれば調査の時間も短くでき、求人票では得られない口コミ情報なども得られると考えた。そのシステムの実現のためにはインターネットで検索を行い、検索結果を蓄積しておく必要がある。本研究では、データ蓄積のために MySQL を使い、検索には Google 社が提供する GoogleAPI という WebAPI を使用する。WebAPI を用いれば、自分の蓄積したデータだけでなく、インターネット上の膨大な情報を扱えるという利点がある。実際のプログラミングには PHP という、Web システムなどでよく使われるプログラミング言語を使用する。ここで WAMP (Windows/apache/MySQL/PHP) の Web システム構築環境が整っていることから、Web 上の情報を自動で蓄積し、それを閲覧するというシステムを選択した。

2.2 必要な技術

WebAPI には GoogleAPI を用いた。これは、地図情報を扱う Google Maps をはじめとした様々な機能

を扱うことができる。本研究で使用した機能は、Google 検索エンジンでの検索結果を取得できる Google Custom Search API を使用した。無料サービスでは、検索は 1 日 100 回まで、一回の検索で得られるデータは 10 件までという制限が設けられているので、求人企業の情報の取得を WebAPI で継続的に行い、データベース化することで、回数制限に影響されず、大量のデータを扱えるシステムを実現することを目標とする。

データベースには、大量のデータを高速に扱えるという特徴を持つ MySQL を、システムの自動化には、タスクスケジューラという、タスク処理を行う Windows の標準機能を用いた。

3. システム概要

本システムは、学校に来た求人票をソースとして、インターネットで求人企業について検索をかけ、雰囲気や評判などの情報を自動で収集してデータベースに蓄積し、そして蓄積した情報を後に閲覧できるようになっている。本システムの概要図を下に示す。

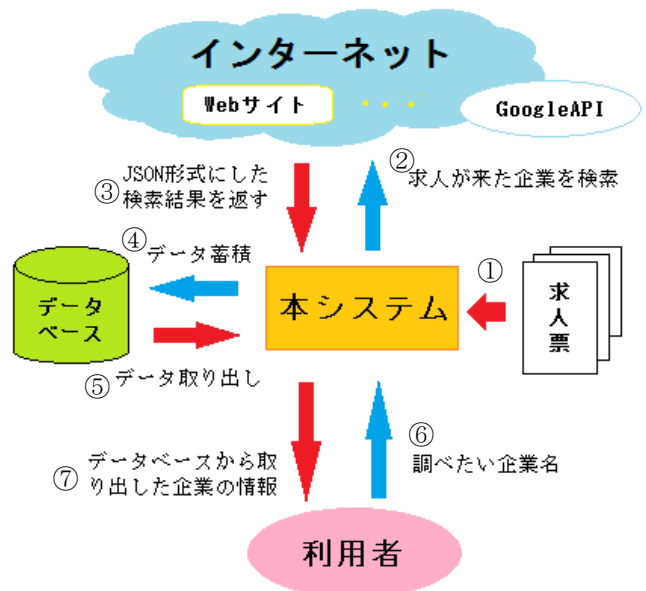


図 1. システム構成図

3.1 企業情報システムの構成

企業情報システムは大まかに、
 A. 企業情報収集ロボット
 B. 企業情報表示システム
 の 2 つからできている。

3.1.1 企業情報収集ロボット

企業情報収集ロボットとは、本校に寄せられた求人票を基に、(1) インターネットでの情報検索を行う検索機能、(2) 収集した情報から必要なものを抽出する解析機能、(3) 抽出した情報をデータベースに保存する蓄積機能の3つによって情報の収集と蓄積を自動で行うものである。

- (1) 検索は Google Custom Search API の URL に検索に必要なパラメータを加えてリクエストをかけることで結果を得る。検索クエリパラメータは、「企業名 AND (○○ OR ○○ OR … OR ○○)」とした。○○に入るキーワードは、就活サイトで行われた企業選びの基準アンケートという学生向けのアンケートで、順位が高かったワードを抜粋。(ブラック、強みなど)
- (2) 検索結果は JSON 形式のデータで返される。これを連想配列に変換することでデータの構造を簡単化し、情報の抽出を容易に行うことができた。抽出は API の制限限界の 10 件のサイト名、サイトの URL、Web ページの要約文(スニペット)にのみ行った。
- (3) 抽出した情報はデータベース内で企業名ごとのテーブルに蓄積するようにした。使用したデータベースの構造を下に示す。

表 1. データテーブルの構成

変数名	変数の型	内容	例
title	string	サイト名	○社の裏
link	string	サイトの URL	http://www.○.com
snippet	string	スニペット	○社とは近年力をつけてきた企業で…

これらの機能の PHP プログラムをタスクスケジューラによって起動させることで、自動で情報の収集を行う。WebAPI の制限を考え、ロボットは 1 日 (0~3 時) に 90 回動かす事を想定している。

3.1.2 企業情報表示システム

企業情報表示システムとは、ユーザーが入力した学科や企業名を基に、求人票とデータベースに蓄積された情報をブラウザで表示する機能である。ユーザから入力が行われると PHP プログラムにより、データベースから必要なデータを取り出し、全て表形式で表示する。

4. 研究成果

現段階では、自動で検索と蓄積を行い、ユーザの入力に合わせて、データを求人票と合わせて表示するところまで実装できた。

図 2 は閲覧画面である。機械工学科 (M) を選択した結果から一番上の「チームラボ株式会社」を選択した結果が図 3 である。

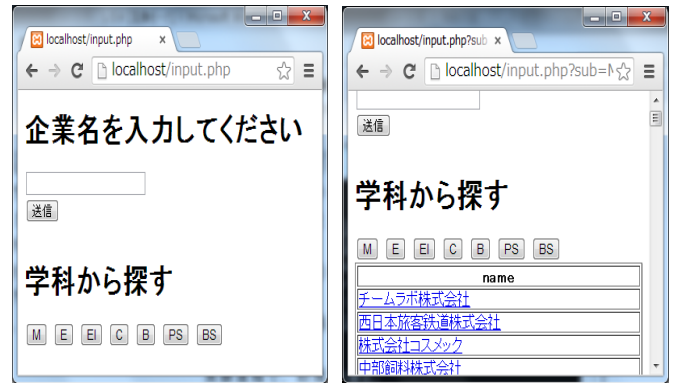


図 2. 閲覧機能の入力画面

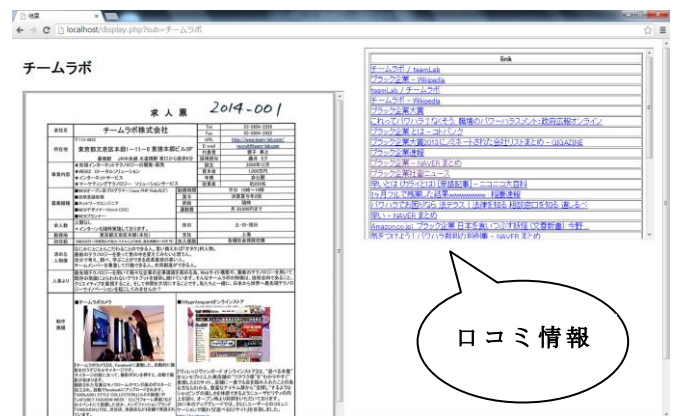


図 3. 検索結果の表示画面

検索結果にできるだけ不必要なデータが含まれないように、AND 検索、OR 検索、AND,OR 検索の三つで実験を行い、最も適切なデータを集めることができる AND,OR 検索で検索を行うようにした。

今後の展望

検索方法の変更だけではデータの絞り込みが不十分なので、蓄積の際に必要な不必要を判断するフィルタ機能を考えて、システムに実装する。

また、システムを実際に利用してもらい、必要な機能や改善点などを指摘してもらい、それらをシステムに反映させる。