MIF37 - Mashup de donnée

Antoine Martin

13 juin 2015

Résumé

Pour ce projet, il a été demandé de réaliser un mashup de données sur différentes *API*(Application Programming Interface).

Le but d'un mashup est de fournir un accès uniforme à des sources de données hétérogènes stockées sur des *API* différentes.

Une entreprise peut faire du recrutement par elle-même ou demander à d'autres entreprises de s'en charger. Il arrive donc souvent que le nom des sociétés soit similaire et que l'on retrouve parfois les mêmes offres d'emplois sur des plateformes différentes.

Le but est de détecter ces offres d'emplois de manière unique et de réaliser un mashup de données à partir des données disponibles.

Le choix des *API* étant libre, il a été décidé de le réaliser sur le marché des offres d'emplois américain. Les deux *API* choisies sont donc Indeed et CarreerBuilder.

1 Contexte

Une fois la partie technique terminée, il a fallu faire plusieurs choix. Tout d'abord, ce mashup ne permet pas de donner les identifiants d'une offre d'emploi et de la récupérer. L'application va elle-même à partir de paramètres prédéterminés effectuer une recherche sur les API.

```
1 data: {
2     q: 'title:' + keywords,
3     l: location,
4     co: country,
5     jt: 'fulltime',
6 }
```

Listing 1 – Liste des paramètres pour chaque API

On voit ici que les paramètres permettent de restreindre les données que nous recevrons des API. Cette restriction est volontaire, car elle va permettre d'obtenir des résultats plus pertinents. Nous avons donc la localisation (*location*), le pays (*country*), la limitation à l'équivalent du Contrat à Durée Indéterminée (*it*) ainsi que des mots-clés (*keywords*) qui ne seront appliqués qu'aux titres des offres d'emplois et non à leurs contenus.

Le but ici est d'obtenir et d'identifier l'offre d'emploi émise par la même société sur les deux *API*. Le premier réflexe serait de trouver des résultats similaires.

2 Caractéristiques des sources

Les résultats vont dépendre du schéma des sources de données et ce schéma est donné par les API.

Les sources Indeed et CareerBuilder sont de type autonome et hétérogène. Chacune d'entre elles ne pense pas l'aspect géolocalisation d'une offre d'emploi de la même manière et les sources ne possèdent pas nécessairement les mêmes informations. L'aspect structurel est également différent avec des attributs pouvant signifier la même chose, mais avec des noms différents (ex : jobTitle et title).

3 Recherche des offres sur les API

L'application se déroule en trois parties, la première est la recherche des offres d'emplois basée sur des mots-clés, une ville et un pays.

À partir de cette recherche, l'application va remplir des collections d'objets correspondant aux schémas ci-dessous.

```
var jobSearchIndeedModel = Backbone.Model.extend({
    defaults: {
        apiUrl: 'http://api.indeed.com/ads/apisearch',
        api_key: '',
        jobtitle: '',
        company: '',
        city: '',
        state: '',
        country: '',
        formattedLocation: '',
        source: '',
        date: '',
        snippet: '',
        url: '',
        jobkey: '',
        sponsored: '',
        expired: '',
        totalResults: ''
    }
});
```

Listing 2 – Schéma d'une offre d'emploi chez Indeed

```
var jobSearchModel = Backbone.Model.extend({
    defaults: {
        apiUrl: 'http://api.careerbuilder.com/v2/jobsearch',
        employeeTypeApi: 'http://api.careerbuilder.com/v1/
            employeetypes',
        api_key: '',
        title: '',
        company: '',
        compagnyDetailUrl: '',
        did: '',
        distance: '',
        employmentType: '',
        jobDetailsURL: '',
        jobServiceURL: '',
        locationLatitude: '',
        locationLongitude: '',
        location: '',
        postedDate: '',
        pay: '',
    }
});
```

Listing 3 – Schéma d'une offre d'emploi de chez CarreerBuilder

4 Résultat de la recherche

Dans un second temps, nous pouvons vérifier les offres d'emplois trouvées et nous comparons chacune d'entre elles. Nous ne sauvegardons que les offres d'emplois qui possèdent exactement le même titre et la même société, ceci toujours dans le but de rester pertinent dans les résultats que nous fournirons.

Il est possible que certaines sociétés puissent publier sous d'autres noms leurs offres d'emplois. L'une d'elles a particulièrement attiré mon attention, CyberCoders, qui est responsable du recrutement pour d'autres sociétés. Il est donc possible de retrouver la même annonce sous deux sociétés différentes.

5 Algorithme d'alignement

Lorsque les informations sont les mêmes, le choix s'effectue suivant le type de donnée. Par exemple, s' il s'agit d'une chaîne de caractères et que l'attribut est le même, alors la plus longue chaîne de caractères est choisie en partant du principe que plus la chaîne est longue, plus il y a d'informations dans la chaîne.

Pour les entiers, par exemple, lors du choix de la longitude et de la latitude, le choix est de faire une moyenne entre les deux mesures, car nous ne pouvons déterminer avec certitude laquelle des données est la plus correcte. Nous aurions pu faire appel à une *API* externe, comme celle de Google Maps Geocoding pour vérifier les informations géographiques.

Nous avons décidé de garder l'emplacement géographique (adresse) de l'API de Indeed, car elle permet de distinguer plus facilement les données. Nous n'avons aucun découpage ou travail à faire concernant la récupération des sous

informations contenues dans l'emplacement géographique donné par CarreerBuilder.

L'attribut distance n'existant pas du côté de l'*API* Indeed. Il s'agit d'un ajout d'information dans le mashup. Il en va de même pour l'attribut *pay*(rémunération) qui est unique à CareerBuilder.

```
alignmentSearchAction: function(e) {
    e.preventDefault();
    var longitude = (this.scb.get('locationLongitude') + this.sid.
        get('longitude'))/2;
    var latitude = (this.scb.get('locationLatitude') + this.sid.get
        ('latitude'))/2;
    var title =  this.compareString(this.scb.get('jobtitle'),this.
        sid.get('title'));
    var company =  this.compareString(this.scb.get('company'),this.
        sid.get('company'));
    var id =  this.idConversion(title,company);
    this.mashup.set({
        id:id,
        title: title,
        company: company,
        date: this.sid.get('date'),
        city: this.sid.get('city'),
        state: this.sid.get('state'),
        country: this.sid.get('country'),
        distance: this.scb.get('distance'),
        longitude: longitude,
        latitude: latitude,
        employmentType: this.scb.get('employmentType'),
        pay: this.scb.get('pay'),
        snippet: this.sid.get('snippet')
    });
}
```

Listing 4 – Algorithme d'alignement

6 Problèmes rencontrés et solutions

La majorité des problèmes rencontrés était d'ordre technique. Par exemple, CareerBuilder ne permet pas de faire des requêtes de type *CORS* (Cross-origin resource sharing) sur ses serveurs. Ce problème a été réglé avec l'installation d'un plug-in Chrome qui ajoute dans la réponse HTTP du serveur l'acceptation de toutes les requêtes ('Allow-Control-Allow-Origin : *).

Ensuite, il a été difficile de faire des choix concernant certains éléments, par exemple la date qui est formatée d'une manière précise chez Indeed. Cependant, la décision a été de privilégier l'aspect temporel par rapport à la précision de l'information. La date la plus vieille sera donc retenue pour la date de publication d'une annonce.

```
1  // CareerBuilder = '6/12/2015'
2  // Indeed =  'Tue, 26 May 2015 08:51:02 GMT'
3  // Retourne la date la plus vieille
4  compareDate : function(){
5      var dateI = Moment(this.sid.get('date'));
6      var dateC = Moment(this.scb.get('postedDate'));
7      if(dateI > dateC){
8          return  dateC;
9      } else {
10         return dateI;
11     }
12 },
```

Listing 5 – Format de la date

7 Conclusion

À partir des tests réalisés, nous avons pu obtenir un mashup avec de nouveaux attributs qui peuvent permettre à des candidats de mieux cibler certaines offres d'emplois. Nous avons pu découvrir les attributs distance et rémunération qui ne sont disponibles que chez CareerBuilder et donc les proposer dans le mashup. Nous avons également dû faire une conversion d'identifiant à partir du titre de l'offre.

Malheureusement, par manque de temps, nous n'avons pas réussi à implémenter la partie générique de l'application avec comparaison automatique du nom des attributs de données.

8 Annexes

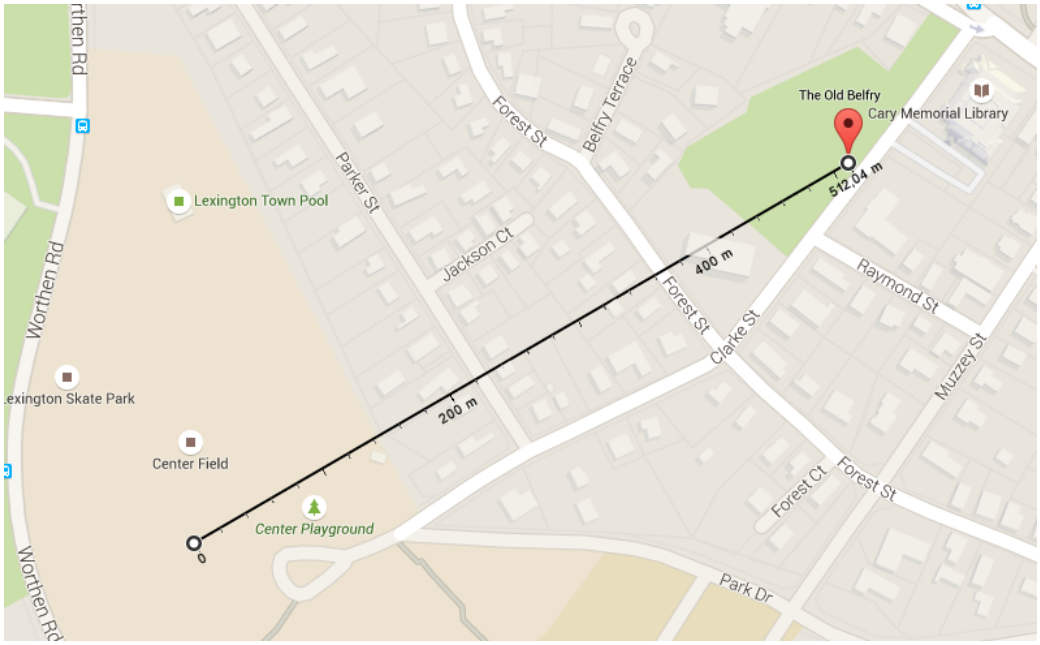

FIGURE 1 – Différence de géolocalisation

Source CarrerBuilder

- **titre**: Lead Software Engineer (PHP/LAMP)
- **company**: Resource Technology Partners
- **distance**: Nearby
- **city**: Boston
- **state**: MA
- **employmentType**: Full-Time
- **locationLatitude**: 42.3568
- **locationLongitude**: -71.0536
- **location**: MA - Boston
- **pay**: $135k - $175k/year
- **postedDate**: 6/10/2015
- **compagnyDetailUrl**:
- **education**: Not Specified
- **jobDetailsURL**: http://api.careerbuilder.com/v1/joblink?TrackingID=UNTRKD&HostSite=US&DID=J3K86J6684WS18FXBKH
- **jobServiceURL**: https://api.careerbuilder.com/v1/job?DID=J3K86J6684WS18FXBKH&HostSite=US&DeveloperKey=WDHS42Z6NVCBC17NTKWY

Source Indeed

- **titre**: Lead Software Engineer (PHP/LAMP)
- **company**: Resource Technology Partners
- **city**: Boston
- **state**: MA
- **country**: US
- **source**: ReSource Technology Partners
- **date**: Wed, 13 May 2015 20:24:45 GMT
- **snippet**: **PHP**. We are searching for a Senior **Software Engineer** to join our award winning engineering team. As a Senior **Software Engineer** on our team you will be tasked...
- **latitude**: 42.357143
- **longitude**: -71.05495

FIGURE 2 – Comparaison d'offres

Mashup


```
{
  "title": "Lead Software Engineer (PHP/LAMP)",
  "company": "Resource Technology Partners",
  "date": "Wed, 13 May 2015 20:24:45 GMT",
  "city": "Boston",
  "state": "MA",
  "country": "US",
  "longitude": "-71.0536",
  "latitude": "42.3568",
  "employmentType": "Full-Time",
  "pay": "$135k - $175k/year",
  "snippet": "PHP. We are searching for a Senior Software Engineer to join our award winning engineering team. As a Senior Software Engineer on our team you will be tasked..."
}
```


FIGURE 3 – Exemple de mashup