**Research document**

***Automatically add missing data***

*8vance Matching Technologies BV*

*Venlo*

|  |
| --- |
| **Date : 09-03-2016** |
| **Version : 0.1** |
| **Status : Concept** |
| **Document name : Research document v0\_1** |
| **Author : Tim Hermens** |

**Version**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version** | **Date** | **Author(s)** | **Changes** | **Status** |
| 0.1 |  | Tim Hermens | First concept. | Concept |
|  |  |  |  |  |

**Distribution**

|  |  |  |
| --- | --- | --- |
| **Version** | **Date** | **Recipient** |
|  |  |  |

**Approval**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date of approval** | **Name** | **Signature** |
|  |  |  |  |

**Table of contents**

1. Introduction 4

1.1 Goal of this document 4

1.2 Approach 4

2. Research questions 5

2.1 Main research question 5

2.2 Sub-questions 5

3. Research results 6

3.1 Sub-question 1 - What data do the profiles consist of? 6

3.2 Sub-question 2: Which selection of profile data delivers the most comprehensive overview of all

the profiles to make predictions to supplement the missing KSC data per profile? 12

# Introduction

## Goal of this document

The company 8vance Matching Technologies BV uses scraping techniques on various social media websites like LinkedIn to collect a large amount of profile data of companies and people. The problem is that this data often misses interesting (parts of) information. The data especially lacks a complete list of skills of people (this is called the KSC data).

The goal of this document is to find solutions to solve this problem. Eventually, a solution must be found that is able to automatically add the missing KSC data to the profiles.

## Approach

This research document is divided in several chapters.

First, the main research question is specified. This question will ultimately yield an answer to the problem mentioned in chapter 1.1.

After this, the sub-questions are specified. These questions provide answers that help to determine a definitive answer on the main question.

And finally, the conclusion of this research can be read that discusses the answer on the main question. Depending on the answer, it may be possible that further actions need to be taken in the future. This will also be discussed in that chapter.

# Research questions

## Main research question

The main research question is as follows:

*What solutions make it possible to supplement the missing KSC data for profiles of persons as accurately as possible?*

## Sub-questions

To help answer the main research question, four sub-questions are specified which provide answers on different aspects to solve the problem. They're as follows:

1. What data do the profiles consist of?
2. Which selection of profile data delivers the most comprehensive overview of all the profiles to make predictions to supplement the missing KSC data per profile?
3. Which solution can be used with this selection of profile data to supplement the missing KSC data per profile?
4. Which solutions have the best quality and performance and is the client satisfied with?

# Research results

## Sub-question 1 - What data do the profiles consist of?

8vance has scraped millions of profiles from various social media websites. They currently scrape the profiles from the following websites:

* LinkedIn (https://www.linkedin.com/);
* Xing (https://www.xing.com/nl);
* Academia (https://www.academia.edu/);
* Researchgate (https://www.researchgate.net/);
* About.me (https://about.me/);
* Medium.com (https://medium.com/);
* Zoominfo (http://www.zoominfo.com/).

There're quite a few differences between the data scraped from these websites. Table 1 contains all the data that's retrieved from the profiles from all these websites. The second column contains a colour code indicating the usefulness of the data which could be used to determine the skills of a person.

|  |  |  |
| --- | --- | --- |
| **Data** | **Usefulness** | **Description / motivation** |
| Source |  | **Description:** A keyword/flag that describes the source of the data (e.g. LINKEDIN). |
|  | | **Motivation:** A keyword/flag describing the source of the data doesn't tell anything about what skills a person might have. There're too many profiles that are in the same source. |
| URL |  | **Description:** The profile's URL |
|  | | **Motivation:** The profile's URL is unique for every profile, which is useless to find similarities in profiles. |
| Name |  | **Description:** The person's name |
|  | | **Motivation:** The name of a person is too random to use as a feature to predict skills. |
| Photo |  | **Description:** A list of urls pointing to the person's photos |
|  | | **Motivation:** Like profile URLs, there's no interesting information to be extracted from photo URLs. |
| Locality |  | **Description:** The person's current address |
|  | | **Motivation:** It may be interesting to use the locality of a person because certain functions or educations could only be available in certain locations. |
| Industry |  | The current industry the person is working in |
|  | | **Motivation:** The industry could be telling a lot of information about |
| Summary\_present |  | A list of companies the person is working for at the moment |
|  | | **Motivation:** |
| Summary\_past |  | A list of former companies the person has worked for |
|  | | **Motivation:** |
| Summary\_education |  | A list of educational institutions where the person is studying at the moment |
|  | | **Motivation:** |
| Summary |  | A free text description of the person |
|  | | **Motivation:** |
| Slogan |  | Description of the person's current state (typically function + company) |
|  | | **Motivation:** |
| Skills |  | A list of skills the person has |
|  | | **Motivation:** |
| Languages |  | A list of languages the person mastered and how proficient he is at them |
|  | | **Motivation:** |
| Experiences |  | A list of the person's working experiences, including information about:   * The function he carried out * The company he worked for * The company's website * The company's location * A description of the carried out function * The starting date * The stopping date |
|  | | **Motivation:** |
| Educations |  | A list of educations the person has studied for, including information about:   * The institution name * The institution's website * The degree he acquired * The major he studied for * The starting date * The stopping date |
|  | | **Motivation:** |
| Also\_viewed |  | A list of profiles the person has looked at, including information about:   * The visited profile's url page * The visited profile's id |
|  | | **Motivation:** |
| ID |  | The profile's id |
|  | | **Motivation:** |
| Crawled\_at |  | The date and time the profile has been scraped at |
|  | | **Motivation:** |
| Academic degree |  | The academic degree of the person |
|  | | **Motivation:** |
| Interests |  | A list of the person's interests |
|  | | **Motivation:** |
| Wants |  | A list of things the person is searching for (like topics) |
|  | | **Motivation:** |
| Group memberships |  | A list of group names the person is in |
|  | | **Motivation:** |
| Email |  | The person's email address |
|  | | **Motivation:** |
| Profiles |  | A list of profiles the person may or may not own of other social media websites, including information about:   * The profiles' source (Facebook, Twitter, etc.) * The profiles' url |
|  | | **Motivation:** |
| Publications |  | A list of the person's publications, including information about:   * The publication's title * The publication's content * A list of keywords describing the publication * A list of the publication's co-authors * The publication's url |
|  | | **Motivation:** |
| Followers |  | A list of profiles of people the person is being followed by |
|  | | **Motivation:** |
| Following |  | A list of profiles of people the person is following |
|  | | **Motivation:** |
| Co-authors |  | A list of profiles of people the person has co-operated with for publications |
|  | | **Motivation:** |
| Advisors |  | A list of profiles of people the person has been advised by |
|  | | **Motivation:** |
| CV |  | A file containing the CV (curriculum vitae) data |
|  | | **Motivation:** |
| Topics |  | A list of topics the person is interested in |
|  | | **Motivation:** |
| Disciplines |  | A list of disciplines the person has |
|  | | **Motivation:** |
| Keywords |  | A list of seemingly random keywords (tags) describing some key points about the person |
|  | | **Motivation:** |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Useful |  | Likely useful |  | May be useful |  | Likely useless |  | Highly likely useless |  | Useless |

Table 1 - A list of all the profile data.

Table 2 contains a view of the usefulness of the data per social media website, excluding the totally useless data indicated in Table 1.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Website** | **Locality** | **Industry** | **Summary** | **Slogan** | **Skills** | **Languages** | **Experiences** | **Educations** | **Academic degree** | **Interests / topcis** | **Wants** | **Profiles** | **Publications** | **CV** | **Keywords** | **Disciplines** |
| LinkedIn |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Xing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Academia |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Researchgate |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| About.me |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Medium.com |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Zoominfo |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Useful |  | Likely useful |  | May be useful |  | Likely useless |  | Highly likely useless |

Table 2 - A view of the usefulness of data per social media website, excluding the totally useless data.

It looks like LinkedIn and Xing have the most useful data. However, LinkedIn has an edge on Xing because of the likely useful industry data field. The academic degree data field of Xing could be useful as well. However, all of the degrees should already have been specified in the educations data field. Additionally, the academic degrees aren't linked to educations whereas degrees specified in the educations data field are. And finally, there're a lot more LinkedIn profiles available to be analyzed compared to Xing.

Let's have a look at one of the profiles from LinkedIn.

|  |  |
| --- | --- |
| **Data field** | **Data** |
| Crawled\_at | 15-08-05 |
| Educations | Date\_start: 2003  Date\_stop: 2006  Degree: 2:ii  Degree\_major: None  Institution: University of Bradford  Major: None |
| Experiences | Company: Teradata Applications  Company\_url: http://www.linkedin.com/company/162487?trk=ppro\_cprof  Date\_start: March 2014  Date\_stop: None  Function: None  Location: Madrid Area, Spain  Company: eCircle  Company\_url: http://www.linkedin.com/company/17169?trk=ppro\_cprof  Date\_start: December 2010  Date\_stop: None  Function: None  Location: None  Company: NH Hotels  Company\_url: http://www.linkedin.com/company/10254?trk=ppro\_cprof  Date\_start: February 2010  Date\_stop: December 2010  Function: None  Location: None  Company: Emailvision  Company\_url: http://www.linkedin.com/company/19916?trk=ppro\_cprof  Date\_start: January 2008  Date\_stop: October 2009  Function: None  Location: None |
| Industry | Information Technology and Services |
| Languages | None |
| Locality | Madrid Area, Spain |
| Profile\_id | inpub-xavier-cabeira-a-80-79a |
| Skills | Email Marketing, Online Marketing, Online Advertising, Digital Marketing, Web Analytics, E-commerce, Management, Conversion Optimization, SEO, SEM, Google Adwords, Mobile Marketing, Digital Strategy, Project Managment, Direct Marketing, Affiliate Marketing, Google Analytics, PPC, Campaign Management, Multi-channel Marketing, Lead Generation, Web Marketing, Database Marketing, CRM, E-business, Marketing Automation, Analytics, Mobile Devices |
| Slogan | None |
| Summary | None |
| Summary\_education | University of Bradford |
| Summary\_past | NH Hoteles, Emailvision |
| Summary\_present | Teradata Applications, eCircle |

Table - Profile data example

This is an example of one of the decent LinkedIn profiles. Although, it's missing some very important data like the Major field in Educations and Function fields in Experiences.

A collection of a hundred thousand profiles was analyzed to check the syntax of the data fields. The syntax column in Table 4 indicates what formats occur for the data field. For instance, the Date\_stop data field in the Educations list can be specified as either None, yyyy (4 digit year), empty string, a string like 'Present' indicating it's ongoing, or MMMM yyyy (full month name and 4 digit year).

|  |  |  |
| --- | --- | --- |
| **Data field** | **Datatype** | **Syntax** |
| Crawled\_at | String | dd-MM-yy |
| Educations | List of objects |  |
| Educations.Date\_start | String | None  yyyy  Empty string  MMMM yyyy |
| Educations.Date\_stop | String | None  yyyy  Empty string  String like 'Present'  MMMM yyyy |
| Educations.Degree | String | None  Empty string  String |
| Educations.Degree\_major | String | None  Empty string  String |
| Educations.Institution | String | None  Empty string  String |
| Educations.Major | String | None  Empty string  String |
| Experiences | List of objects |  |
| Experiences.Company | String | Empty string  String |
| Experiences.Company\_url | String | None  Empty string  String |
| Experiences.Date\_start | String | None  Empty string  MMMM yyyy  yyyy |
| Experiences.Date\_stop | String | None  Empty string  MMMM yyyy  yyyy  String like 'Present' |
| Experiences.Function | String | None  String |
| Experiences.Location | String | None  Empty string  String |
| Industry | String | String |
| Languages | List of objects |  |
| Languages.Language | String | String |
| Languages.Proficiency | String | None  Empty string  String |
| Locality | String | None  String |
| Profile\_id | String | String |
| Skills | List of strings | String |
| Slogan | String | None  String |
| Summary | String | None  String |
| Summary\_education | List of strings | None (the list!)  String |
| Summary\_past | List of strings | None (the list!)  String |
| Summary\_present | List of strings | None (the list!)  String |

Table 4 - Syntax per data field

## Sub-question 2: Which selection of profile data delivers the most comprehensive overview of all the profiles to make predictions to supplement the missing KSC data per profile?

As discussed before, only the LinkedIn profiles will be used. Table 2 showed the usefulness of a collection of data of all the social media websites. Table 5 shows this only for the LinkedIn website and the relevant data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Website** | **Locality** | **Industry** | **Summary** | **Slogan** | **Skills** | **Languages** | **Experiences** | **Educations** |
| LinkedIn |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Useful |  | Likely useful |  | May be useful |  | Likely useless |  | Highly likely useless |

Table - LinkedIn data usefulness