**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.

## Scope

The KSC data contains data about someone's experiences, educations and skills. In this document, finding a solution to add the missing skills has the highest priority. Finding solutions to add the other missing data is desirable, but doesn't have the focus. Therefore it may be possible no solutions for adding these missing data will be discussed in this document.

# 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. The third column contains a description of the data and a motivation why the colour code was picked. This table was created with help of data science expert Sabrina Ziebarth and a document containing information about all of the scraped data (Ziebarth, 2016).

|  |  |  |
| --- | --- | --- |
| **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 means it's 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 are only available in certain locations. |
| Industry |  | **Description:** The current industry the person is working in |
|  | | **Motivation:** The industry can say a lot about the skills a person is likely to own. For instance, someone who is in the IT industry likely owns IT-related skills. |
| Summary\_present |  | **Description:** A list of company names the person is working for at the moment |
|  | | **Motivation:** Just having the company's name someone is working for in most cases doesn't say enough about the skills someone could own. Companies often have a multitude of functions that require different skills. Since the function is unknown, these skills can't be extracted. However, some companies could have functions that all require a certain general skill. |
| Summary\_past |  | **Description:** A list of former companies the person has worked for |
|  | | **Motivation:** This data has the same problem as the 'Summary\_present' data. |
| Summary\_education |  | **Description:** A list of names of educational institutions where the person is studying at the moment |
|  | | **Motivation:** Just having the institution's name someone is studying at doesn't say enough about the skills someone could own. You have to know something about the major someone is following to be able to make a much more accurate prediction of the skills. However, there could be some general skills everyone acquires after the study at a institution, or is required to have when applying for a study at a institution. |
| Summary |  | **Description:** A free text description of the person |
|  | | **Motivation:** It's possible this data contains useful information regarding skills one could own. But since this data format is free text it's very difficult to extract any useful data. Additionally, you have to know the context of the information that is extracted from this data. For instance, if someone mentions skill A in his free text description, there's no telling if that person actually owns that skill without knowing the context (e.g. he could be working on acquiring that skill). |
| Slogan |  | **Description:** Description of the person's current state (typically function + company) |
|  | | **Motivation:** This data isn't really useful because it contains data that can be found more accurately elsewhere. The slogan typically contains the function and company where a person is working. This data and more data regarding jobs can already be found in the experiences data field. |
| Skills |  | **Description:** A list of skills the person has |
|  | | **Motivation:** Since we want to predict skills, it's essential to have a list of skills which can be used to find correlations with other data fields. |
| Languages |  | **Description:** A list of languages the person mastered and how proficient he is at them |
|  | | **Motivation:** This data could be useful to determine some skills. For instance, being proficient at an ancient language or many languages could say a lot about someone's educational background, skills and experiences. However, this would only be of use for a minority of people. |
| Experiences |  | **Description:** 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:** This data likely is very useful to find skills. After all, you need to put skills into practice when carrying out a function. So direct correlations between functions and skills are likely present. Furthermore, the starting and stopping date could be important. Similar functions that were carried out in the past and are carried out in the present likely require the person to have different skills. |
| Educations |  | **Description:** 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:** This data likely is very useful to find skills. Educations serve as a means to acquire skills. Especially the degree and major data fields are of interest here. The degree typically says something about the education level, like bachelor or master. The major typically says something about the discipline. However, sometimes the degree and major are combined in the degree or major data field. Similarly to experiences, the starting and stopping date could be important. Similar majors that were followed in the past and are followed in the present likely involve different skills. |
| Also\_viewed |  | **Description:** A list of profiles the person has looked at, including information about:   * The visited profile's url page * The visited profile's id |
|  | | **Motivation:** This data isn't useful. The profiles a person has looked at can be very diverse. They don't necessarily have to be profiles that have similarities to the person's profile. It might be interesting to retrieve the most viewed profiles as they have a stronger likelihood to have similarities with the person's background. But this is dangerous to assume. |
| ID |  | **Description:** The profile's id |
|  | | **Motivation:** The profile's id is unique for every profile, making it useless to find similarities between profiles. |
| Crawled\_at |  | **Description:** The date and time the profile has been scraped at |
|  | | **Motivation:** This data field could be interesting to use to check different versions of the same profile with each other. For instance, if a profile has more skills and 1 more education specified than an earlier version, these new skills are probably linked to this education. However, there're currently very few profiles present in the database that have been scraped multiple times. This makes this data field unusable for now. |
| Academic degree |  | **Description:** The academic degree of the person |
|  | | **Motivation:** This data field is very similar to the degree data in the experiences data field. The only problem with this data field is that the major may not be specified. The experiences data field also has a lot more information about the educations someone has followed, making that data field superior to this one. |
| Interests |  | **Description:** A list of the person's interests |
|  | | **Motivation:** This data field is likely useless. There could be a correlation between few uncommon interests and skills, but in the majority of cases this won't be the case. |
| Wants |  | **Description:** A list of things the person is searching for (like topics) |
|  | | **Motivation:** Similarly to the interests data field, this data field likely is very diverse and might have very few correlations with skills or none at all. |
| Group memberships |  | **Description:** A list of group names the person is in |
|  | | **Motivation:** This data field is probably useless. Groups can have similar names but stand for different things. It's likely that people in a group are likeminded people and share some similarities. |
| Email |  | **Description:** The person's email address |
|  | | **Motivation:** An email address is likely to be unique for every person, which makes this data useless for finding similarities between the profiles. |
| Profiles |  | **Description:** 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:** Having profiles on other social media websites doesn't directly correlate to any background information of a person. Perhaps having a profile on almost every social media website could say something about the person, but that's highly likely nothing interesting. This data field would only be useful if there were social media websites for one |
| Publications |  | **Description:** 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:** Writing publications is a skill on its own. The keywords that describe the publication could say something about the nature of the publication. Perhaps the publication is about a research and this research could have a correlation with skills the person owns. However, to find any correlations with the publications, a lot of profiles with publications must be used. Writing publications is something only a minority of the people does, meaning this data could only be useful for a small portion of the profiles. And it's unlikely that there're strong correlations between publications and other data. |
| Followers |  | **Description:** A list of profiles of people the person is being followed by |
|  | | **Motivation:** This data has the same problems as the "also\_viewed" data field. |
| Following |  | **Description:** A list of profiles of people the person is following |
|  | | **Motivation:** This data has the same problems as the "also\_viewed" data field. |
| Co-authors |  | **Description:** A list of profiles of people the person has co-operated with for publications |
|  | | **Motivation:** It's dangerous to assume that people the person has co-operated with are active in the same field. Although, it's more likely that these people have similarities with the person than the people in the "also\_viewed" data field because they've worked on the same publication which require the persons to have some skills. |
| Advisors |  | **Description:** A list of profiles of people the person has been advised by |
|  | | **Motivation:** This data has the same problems as the "also\_viewed" data field. |
| CV |  | **Description:** A file containing the CV (curriculum vitae) data |
|  | | **Motivation:** This data field has similar problems and challenges as the "summary" data field. The main difference here's that when a skill is mentioned in a CV as opposed to a free text description, it's more likely the person actually owns that skill. |
| Topics |  | **Description:** A list of topics the person is interested in |
|  | | **Motivation:** This data field may be useful to find persons with similar topics of interest who may have some similarities in skills. However, it's dangerous to assume that people with similar topics of interest have any other similarities with each other. It's highly possible that people with different backgrounds share similar topics of interests. |
| Disciplines |  | **Description:** A list of disciplines the person has |
|  | | **Motivation:** This data is useful to find skills one could own. A person having a certain discipline directly correlates to the person having certain skills; otherwise he wouldn't have the discipline. If the correlations between the skills and disciplines are known, skills can be predicted for people who have disciplines. |
| Keywords |  | **Description:** A list of seemingly random keywords (tags) describing some key points about the person |
|  | | **Motivation:** This data field can contain some useful information about the person, like functions and educations he has done. However, there's a wide variety of other useless data like family- and hobby-related stuff. There's no telling what data is stored in this data field, making it highly likely useless. |

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

Table 1 - A list of all the profile data.

There're quite a few data fields that aren't useful. Table 2 contains a view of the most useful data fields that are available 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** | **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 usefulness of the 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 3 - 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 its most relevant data, which can be seen below.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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

It can be concluded that the *Industry*, *Skills*, *Experiences* and *Educations* data fields will help the most to create correlations that can help to predict the missing skills.

# Bibliography

Ziebarth, S. (2016). *people site for scrape2-1.* Venlo.