University of Applied Sciences Northwestern Switzerland
School of Business

Identity resolution for fraud prevention

BY

Michael Studer

master thesis submitted in fulfilment of the requirements for the degree of

Master of Science in Business Information Systems

in the subject of

BUSINESS INFORMATION SYSTEMS

University of Applied Sciences Northwestern Switzerland FHNW

Supervisor: Dr. Hans Friedrich Witschel

January 2015

Author

Michael Studer

[redacted]

Supervisor

Dr. Hans Friedrich Witschel

Riggenbachstrasse 16

4600 Olten

hansfriedrich.witschel@fhnw.ch

Abstract

The purpose of this master thesis is to find the best solution for identity resolution between social media networks and bank customers. With the contacts of the identified social media user it should be possible to identify fraud based on circular references of money transactions.

The literature shows that identity resolution between multiple offline data sources is well researched. With the emerging social media networks the identity resolution between these networks was also deeply researched. Both problems can be solved by existing solutions. However, identity resolution between offline data and social media networks is not well researched, yet.

The thesis is based on design research. In the different phases knowledge was gained trough literature research, interviews and meetings. An artefact was developed to evaluate and optimise the different algorithm variations. The final algorithm was then evaluated with a set of test data.

In this work an algorithm is presented, which is able to identify social media users based on bank customer information with an accuracy of 80%. The key to a successful identity resolution lies within the similar data structure of the money transactions of a bank customer and the contacts of a social media profile. The best identity resolution was achieved with different weighting for the different attributes and by the normalisation of the transactions in addition to the normalisation of the name based on the name frequency.

The conclusion of this work is that the thesis statement is confirmed. It is possible to correctly identify a person within a social media network based on the information available from a bank customer.

Statement of Authenticity

I confirm that this master thesis research was performed autonomously by myself using only the sources, aids and assistance stated in the report, and that quotes are readily identifiable as such.

_____

Date, Prename Name

Acknowledgments

I would like to thank Dr. Hans Friedrich Witschel for his support and guidance for this master thesis. With his time and support it was very inspiring to work in the thesis.

In addition I would like to thank Trivadis for the possibility to write this thesis and specially Beno Leuenberger for his time and the constructive meetings.

Contents

1 Introduction 1

1.1 Background . . . 2
1.2 Thesis Statement . . . 3
1.3 Research Question . . . 4
1.4 Scope and Limitation . . . 4
1.5 Research Strategy . . . 5
1.6 Time Horizon Time Frame . . . 6

2 Literature Review 7

2.1 Identity Resolution . . . 7

3 Research Methodology 16

3.1 Research Strategy . . . 17
3.2 Research Philosophy / Paradigm . . . 22
3.3 Research Approach . . . 23
3.4 Research Choices . . . 24
3.5 Time Horizons . . . 24
3.6 Data Collection Methods . . . 24

4 Data Source 25

4.1 Banking Information . . . 25
4.2 Xing . . . 29

5 Identity Resolution 31
5.1 Graph . . . 31
5.2 Matching . . . 32
5.3 Rating Calculation . . . 38
5.4 Algorithm Variations . . . 38

6 Results 41
6.1 Experimental Setup . . . 42
6.2 Email Search . . . 42
6.3 Name Search . . . 43
6.4 Time zone Rating . . . 43
6.5 Contacts and Transaction Rating . . . 44
6.6 Birthday Rating . . . 44
6.7 Matching . . . 45
6.8 Identity Resolution Probability . . . 47
6.9 Performance . . . 49

7 Conclusion 51
7.1 Recommendations . . . 52
7.2 Future Work . . . 52

References 54

Appendix A Appendix 59
A.1 Weighted Attributes List . . . 59
A.2 Fuzzy search logic . . . 63

1 Introduction

The purpose of this study is to provide the best solution to identify an individual from within company data in a social media network, which is called identity resolution. It's goal is to prevent fraud in the financial industry. To achieve an accurate match, the solution should take use of all available and suitable information about the customer, which is available at the company, and the accessible information form a social media network. All this information has to be processed as efficient as possible to provide results in near real-time to make it usable for the company for fraud detection.

If we can identify a person in a social media network, which provides information about the persons relationships with friends, colleagues or relatives, this information can be used to identify circular references in financial transactions and therefore provide an approach to spot fraud in the financial industry.

According to the Annual Fraud Indicator the average fraud loss in the UK private sector in 2012 is estimated to be about 0.54 per cent of the countries turnover. If we assume, that

the fraud loss in Switzerland is similar to the one in the UK, the Swiss financial industry lost in 2012 about 911 million Swiss Francs due to fraud in 2012.

1.1 Background

Identity resolution is an extensively studied topic (Elmagarmid et al., 2007) in the field of business intelligence and has been studied for more than five decades (Knuth, 1959). It is used for the integration of different data sources and to increase the data quality (Bhattacharya & Getoor, 2004). This differs from the required matching between offline data and online data from the social media network.

1.1.1 Identity resolution

Identity Resolution is a semantic reconciliation activity, which is applied to people and organisations. To identify and/or match an identity two or more disparate data sets are searched and analysed. According to Jonas (2006) identity resolution is most frequently quantified in terms of accuracy (false positives and false negatives). Currently two fields are mainly studied – between offline data and between online data.

1.1.2 Social Media

In the year 2013 in Switzerland 86.7% of the population older than 14 used the Internet at least once a year and 81.1% more than once a week (BFS, 2014a). A study from "We Are Social Ltd" (2014) in February 2014 showed that Facebook had the highest social network penetration in Switzerland with 43% or 3.4 million users.

1.1.3 Fraud Detection

More than 99% of fraud perpetrated in Switzerland in 2012 was committed by people younger than 70 (BFS, 2014b). This is especially important, because the internet usage decreases heavily at the age of 70.

1.1.4 Big Data

Financial institutes possess a lot of information about their customers and partners stored in different applications. Master data about their customers are usually only changed during the process of the contract completion. The data generated during the daily business consisting of transaction data generate the vast majority of data. Some customer information is stored in a core system others spread over different other applications or data ware houses. Due to the amount of data, the available information cannot be analysed in the application, which holds the information, because they do not scale accordingly. To be able to perform an ad-hoc identity resolution with a social media network based on all available financial data a big data solution is required. The main advantage and identification of a big data solution lies in the scalability to analyse big data (Russom, 2011).

1.2 Thesis Statement

It is possible to design an identity resolution method that is able to recognise persons, based on company data, in publicly available Social Media Data with high accuracy in near real-time.

1.3 Research Question

Based on the context and purpose of this research study, the research question and objectives are:

Research Question

How does the best solution for identity resolution from banking data to social network data look like?

- What are the best available offline-online and offline-offline identity resolution solutions?
- What are differentiating characteristics of bank customer to social network identity resolution compared to online-online and offline-offline matching?
- What would be a better solution for attribute matching to identify a bank customer in social networks?
- What is the best solution to aggregate the attribute matching results to identify a bank customer in social networks?
- Which external data sources can support identity resolution and in which ways would it be possible?
- What are meaningful criteria to measure the accuracy of the identity resolution?

1.4 Scope and Limitation

This thesis is restricted on identity resolution for the purpose of fraud prevention in the retail and investment banking sector. National banks and clearing institutes are excluded due to the different customers and therefore also different data attributes and use cases. The target Social media network is XING, because it provides a social graph with all connected friends, which is the most important attribute for fraud prevention in this study. To increase the accuracy of the matching, it is possible to use additional directories, which can be done paradigmatic with

solutions limited to Switzerland. If the use of additional information, such as directories, increases the accuracy, the method should be adaptable with products in other countries as well.

1.5 Research Strategy

The research of this thesis is based on the design research paradigm described by Vaishnavi and Kuechler (2004). It consists of a cycle of the design research phases – awareness, suggestion, development, evaluation, and conclusion. The goal is to design and create an artefact for the identity resolution problem, described in the research question.

The model starts with the awareness phase where the problem is identified and defined in collaboration with the sponsor of the project in multiple meetings. With the common understanding of the problem the research target is then defined in the problem statement and research questions. In the following suggestions phase a preliminary solution for the identity resolution is created, based on the literature described in chapter 2. In the development phase the solution is redefined and the final artefact is developed. This is done in an iterative process with a constructivist methodology. When the development process is completed the artefact is evaluated according to the functional in the suggestion phase. The conclusion terminates the study.

A detailed research design and methodology for each phase is described in the chapter 3.

1.6 Time Horizon Time Frame

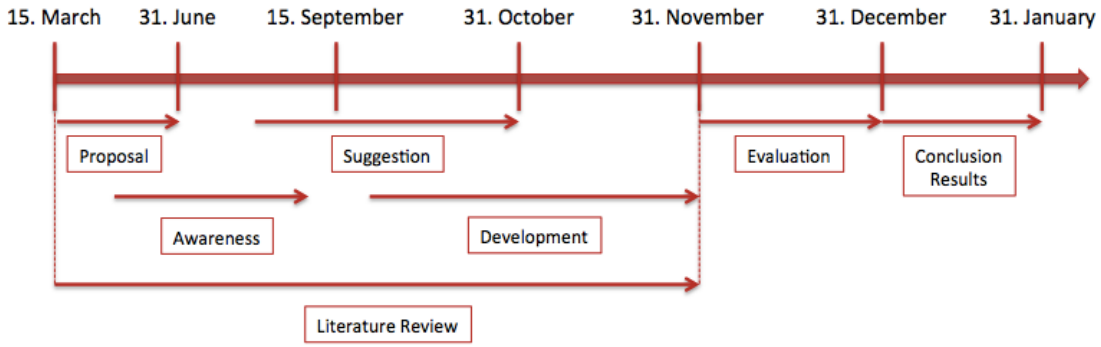

Figure 1.1: Research Timeline

2

Literature Review

This chapter provides insights on the existing knowledge in the research community about identity resolution and its background as well as fraud prevention. Both topics are explained based on literature review.

2.1 Identity Resolution

Today's IT-based economy depends on the accuracy of information. The required information is usually stored in databases. The quality of the stored data therefore has significant cost implications to a system that relies on information in order to function and conduct business. To gain a holistic data overview the linking of multiple tables according to their key fields is required. Unfortunately data usually does not contain unique global identifiers, therefore the matching or duplicate detection between different entities has to be achieved by field matching techniques. The results of this comparison are then aggregated and used

for duplicate detection.

According to Elmagarmid et al. (2007) a general problem is the data quality, because often it is not carefully controlled or defined in a consistent way. He suggest that there are different categories for data errors – data entry errors, such as spelling mistakes, missing integration constrains when fields are not validated, multiple conventions or abbreviations for an identical record.

The matching challenges can be categorised into two types of data heterogeneity – lexical and structural. Lexical heterogeneity deals with data that uses different representations to refer to the same real-world. This is the case when e.g. abbreviations are used (street = Winkelriedstr 62 compared to Winkelriestrasse 62). Structural heterogeneity deals with different structured database fields across different databases. These are challenges that have to be solved in online and offline identity resolution.

Talburt (2010) suggests that entity resolution generally consists of four basic techniques to define if entities are equal and should be linked – direct matching or transitive linking, linking by association, asserted linking.

Direct Matching

Direct matching is based on the similarity of two objects based on their values of the corresponding attributes. There are multiple methods to define if two objects should be linked:

- Deterministic matching: link if the values from all attributes are equal.
- Probabilistic matching: link if certain attributes are equal.
- Fuzzy matching: link if the values of attributes are similar. This is achieved by approximate string matching (ASM) as described in chapter 6.7

Transitive Linking

Transitive linking is based on a chain of intermediate links. This technique is also called transitive closure. An example would be if element A is linked to B and A is linked to C then A is linked to C.

Linking by Association

Linking by association is based on the references of an object. In an example this might look as follows: Michael Studer and Michael Studr have the same address and the same phone number. Therefore Michael Studer and Michael Studr can be linked by association.

Asserted Linking

This technique is based on additional knowledge from a reliable, external data source about the entities. This is also known as knowledge-based linking. An example would be a dictionary, which defines the American word analyze is equal to the British word analyse.

2.1.1 Field Matching Techniques

Misspelling and different conventions resulting in typographical variations within a field require techniques to measure the similarity of individual fields. Multiple methods for string similarities have been developed and each method works well for particular types of errors.

This chapter describes some techniques that are used for matching fields with string data in the duplicate record detection context. Because of the high importance for identity resolution this field matching techniques are covered in a dedicated chapter. According to Rehman & Esichaikul (2009) the techniques for detecting similarities of string-based data are character-based similarity metrics and token-based similarity metrics.

Character-Based Similarity Metrics

These metrics are designed to handle typographical errors, originated from variations of manually entered data, correct. The process of duplicate detection relies on approximate string matching techniques to handle such problems. Unfortunately there is very little research, which compares the effectiveness between the different metrics. According to Yancey (2005) the Jaro-Winkler metrics works well for name matching tasks for data coming from the US census. IT is also a widely used metric for the purpose of record linkage. The Jaro-Winkler metric is therefore further considered for the identity resolution.

Token-Based Similarity Metrics

Typographical errors are handled well by character-based similarity metrics, but if the words are rearranged due to a different convention, the metrics fail to capture the similarity. Bilenko et al. (2003) compared the effectiveness of character-based and token-based similarity metrics and discovered that the Monge-Elkan metric has the highest average performance across data sets. However SoftTF.IDF works better than any other metrics. These two metrics are therefore also further considered for the identity resolution and will be evaluated in this study.

In the research of Veldman (2009) these matching metrics are compared. Based on the result of experiments with names in social media the Monge-Elkan metric is however dropped from further consideration and the Jaro-Winkler metric will be used.

According to Bilenko et al. (2003) there is no single metric, which is suitable for all data sets. A metric may have a high performance for some data sets but will perform poorly on others. Therefore the most suitable metrics for identity resolution between banking data and the available social network data have to be evaluated.

2.1.2 Offline - Offline

In the traditional matching of offline data we can distinguish between two types:

- Entity Resolution is the process of finding non-identical duplicates in a relation and merging the duplicates into a single tuple (record), as described by Benjelloun et al. (2009).

- Record linkage is the process of finding related entries in one or more related relations in a database and creating links among them, as described by Malin & Sweeney (2005)

Identity resolution between offline and online data is closely related to the problem of record linkage. Therefore this topic is explained in further details. Across research communities this problem is known under multiple names. In the database community it is known as data deduplication (Sarawagi & Bhamidipaty, 2002), merge-purge (Hernández & Stolfo, 1998), and instance identification (Wang & Madnick, 1989) whereas the Artificial Intelligence (AI) community refers to it by the name Database Hardening (Cohen et al., 2000) and name

matching (Bilenko et al., 2003). Other commonly used names for the same problem are identity uncertainty and duplicate detection.

Data preparation

According to Elmagarmid et al. (2007) the process of identity resolution is a chain of two stages – data preparation and duplicate record detection. During the data preparation stage the entities are stored in a uniform manner in the database. This approach is also known as ETL, where at the structural heterogeneity problems are solved at least particularly but only on a very general level. The data preparation stage contains a parsing, a data transformation, and a standardisation step. Parsing locates, identifies and isolates individual data elements in the source data. It makes it therefore easier to correct, standardise, and match data based on the comparison of individual components.

Data transformation is the step where the field types of the different schemas are harmonised. It contains the type conversion and also renaming of a field from one name to another. This process is executed on one field at a time and does not require values from related fields.

Data standardisation is the last step of the data preparation. It refers to convert information represented in certain fields in a standardised, uniform representation. Without a standardised format many duplicate entries could be erroneously designated as unique object, because common identifying information cannot be compared. According to Elmagarmid et al. (2007) one of the most important standardisation is the address information, because there is no standardised way to store addresses, therefore the same address can be represented in many different ways. The result of this phase usually is a table, where all fields are stored in a comparable way.

Duplicate record detection

In chapter 6.7 the methods to compare individual fields based on string comparison will be introduced. To identify records across multiple databases usually multiple database fields are required, which makes the duplicate detection much more complicated. The approaches for duplicate detection can be divided into different categories. Supervised and semi supervised

techniques rely on training data. This approach dominated the field of duplicate detection for more than two decades. There are also unsupervised learning approaches, where only minimal labelling effort of the comparison vectors in the training set is required.

Distance-Based Techniques are based on a defined metric and threshold to match similar record. This does not require training data. A possible approach is to treat a record as a long string and compare it with the mean of a string comparison method. Another approach is to calculate the distance between individual fields and compute then the weighted distance between the records. The main challenge within this category is the definition of the threshold. The use of training data would mitigate the advantage to operate without training data. Chaudhuri et al. (2005) observed that duplicate entries to the same real word object have small distance to each other and that there is only a small number of other records within a small distance. Chaudhuri et al. calculated the threshold for each object and thereby outperformed a global threshold in terms of quality.

Rule-Based approaches are a special case of distance-base approaches, where experts define a set of attributes, which serve as a key and thereby should uniquely identify a record. This approach usually provides a very high accuracy, but the required tuning requires also a high manual effort from a human expert (Wang & Madnick, 1989).

The graph matching approach views the available data (schemas, catalogues, or other data structures) as graphs where the goal is to create a mapping between corresponding nodes of the graphs. To estimate the correlation of two instances, graph matching technique is used. Graph matching is a very versatile approach, which is used in graph matching of social network graphs (Cui et al., 2013), link resource files (Scharffe et al., 2009) or to identify handwriting recognition (Fischer et al., 2010). Cui et al. could increase the accuracy of the identity resolution from 30%, while only matching social network profile information, or 10% for graph information graph matching, to 50%. This was achieved by using a profile-based similarity search first and evaluate the results based on a graph-based similarity.

The goal of the clustering approach for entity resolution is to cluster the references according to their entities while taking the relationships into account so that equal entities end up in the same cluster (Bhattacharya et al., 2006). The similarity of multiple entities depends on the current cluster and the related references and is therefore affected if related references change. A relational clustering algorithm (RC-ER) for this purpose was also developed by

Bhattacharya et al., 2006).

The techniques can further be divided into ad-hoc techniques that work quickly on existing relational databases and methods based on probabilistic inference models. Ad-hoc techniques usually outperform probabilistic methods in terms of scalability and performance but on the other hand usually lack in accuracy (Elmagarmid et al., 2007).

2.1.3 Offline - Online

Identity resolution between offline an online data uses existing techniques that have been tested in offline identity resolution. Because of the different environments there are additional aspects that have to be taken into account, which will be explained in the following chapters.

Privacy

When working with online data and social networks, privacy is very important and it needs to be ensured, that no confidential information is exposed. This reduces the ability for entity resolution between offline and online data compared to offline data. Sproch & Jong (2010) avoided to expose data by downloading all publicly available data from a social media network. This is also similar to the approach used for record linkage of offline data, where the data is harmonised into a table, which is then analysed for duplicate entries.

Anonymity

Suler (2004) showed that people act differently while they are online compared to how they behave in person with each other. Therefore people share also potentially sensitive information and their relationships while they are online. One factor for this behaviour was that anonymity was expected. The research of Narayanan & Shmatikov (2009) proved them wrong, because they could re-identify a third of the users who have an account on both Twitter and Flickr, an online photo-sharing site.

Connectivity

The access to data within a public network is not as simple as the access to offline data. XING, as an example for a social media network, restricts the amount of queries on multiple levels. Additionally, the enormous mass of data do not make it possible to access and evaluate all the available data for the purpose of an identity resolution. Vesdapunt & Garcia-Molina (2014) used a caching mechanism to reduce the wait time because the rate limit was reached.

Accessibility

The accessible part of the online data is usually a tiny segment of the available data within the social media network. Therefore the ability to match different identities is limited to the accessible attributes. This was also one of the main challenges in the research of Vesdapunt & Garcia-Molina (2014). Most social media networks provide additional options for their users to further restrict the available information for the public.

2.1.4 Online - Online

Identity resolution between to online data sources are primarily researched in the field of identity resolution between multiple social media networks. The process can be divided into identity search methods and the process of identity matching methods.

The search method can be categorised into profile search, content search, self-mention search and network search. The used method depends mostly on the available and accessible information from the social media network.

Jain et al. (2013) identified two types of identity matching methods between social media networks. Syntactic matching is based on a string comparison metric, where the available attributes of the two identities are compared. Image matching tries to identify an identical user based on the image of the social media network. For an identity resolution between multiple social media networks this matching method achieved a higher accuracy compared to a syntactical method based on a name. How successful this method between profile picture and identification picture of a financial institute is, has yet to be analysed. Compared to similar or equal pictures within the social media networks the accuracy is clearly much lower.

Throughout the existing literature identity resolution does not provide a linked identity but proposes a ranking of possible matches and their probabilities.

2.1.5 Literature Gap

Offline-Offline identity resolution is an old field of study and therefore well researched. In some companies it is a business requirement to identify the same data within multiple data sources. This might be the case after a company merger or to integrate data from multiple applications.

Identity resolution between two online data sources, e.g. between Facebook and XING, is also well researched. Identity resolution between offline data and online data is however not well researched. There are very few studies about this topic and also no information available to how high the accuracy of such an identity resolution might be.

3

Research Methodology

This chapter provides information about the applied research design and methodology. The research design defines which specific methods and techniques are employed in the research. The research methodology is concerned with the philosophical perspective of the research in terms of ontological and epistemological assumptions.

The overarching methodology is based on the Research Onion described by Saunders et al. (2009), which depicts the entire spectrum of philosophical themes and methodological issues that have to be taken into account when contemplating a research project. It is a layered model, which supports the theoretical underpinning for the selection of the research methods. According to Saunders (2009), there are six layers – philosophies, approaches, strategies, choices, time horizons, techniques and procedures.

As described in chapter 1.5, this research is based on design research. Design research is a very popular research strategy for research in the field of information systems, because it facilitates research by building and evaluating artefacts designed to address, identify and solve

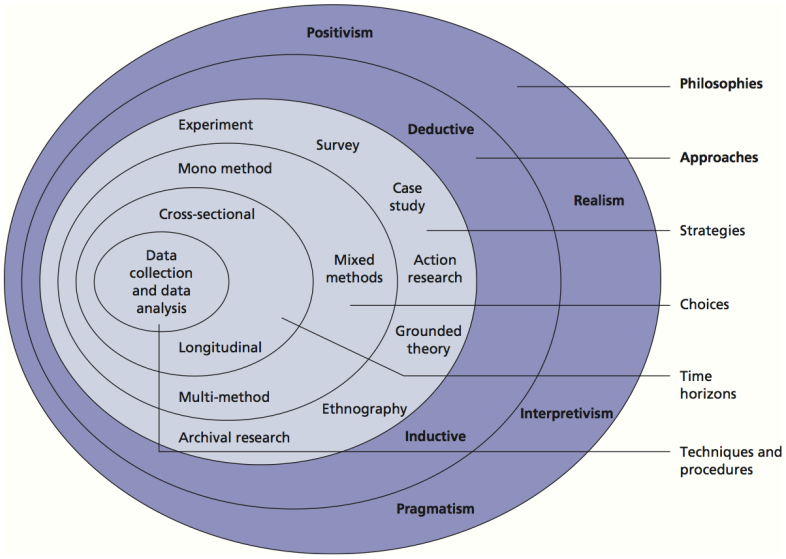

Figure 3.1: The research onion (Saunders et al., 2009)

problems. The main focus of the this chapter lies on design research, which is thus introduced first, despite being the third layer of the research onion, whereas the other layers of the research onion are then based on the outcome of the research strategy layer.

3.1 Research Strategy

Starting with the research strategy layer of the research onion, the following three layers to the inside of the onion are focusing on the process of research design. The research design will be the general plan of how the research question will be answered. Robson (2002) described it as turning the research question into a research project. A research strategy is an investigation approach that moves from the underlying philosophical beliefs to research design and data collection. It can be classified into two general categories – quantitative research and qualitative research (Olivier, 2009; Myers, 1997).

- Qualitative research is an interpretive approach to investigating subjects in their natural surroundings.
- Quantitative research attempts to answer questions about relationships among measured variables with the purpose of explaining, predicting and controlling phenomena.

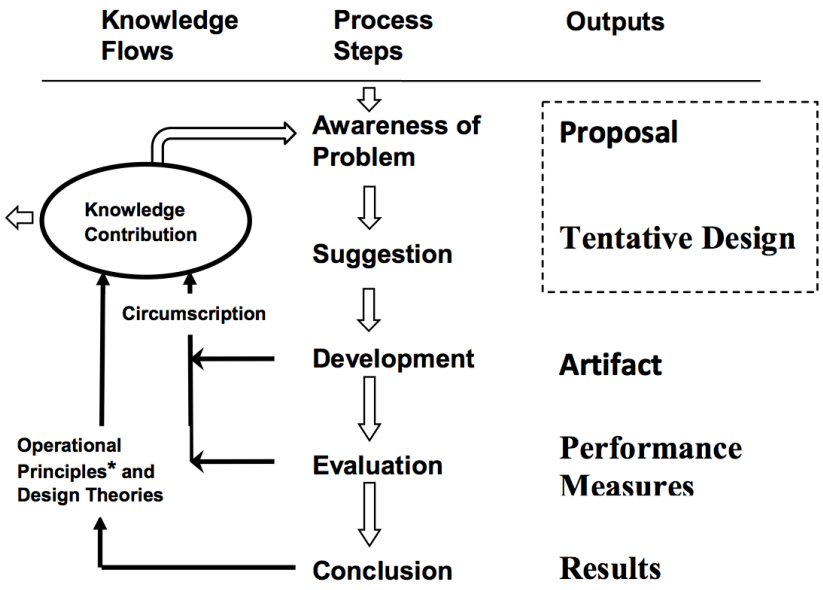

Figure 3.2: Phases of design research (Vaishnavi & Kuechler, 2004)

This study is based on design research, where we can identify both – qualitative and quantitative aspects, which is described in more details in the following. An interested reader may find additional information about other qualitative and quantitative research methods in Saunders et al. (2009) book "Research methods for business students".

3.1.1 Design Research

Design science facilitates research by building and evaluating artefacts designed to address, identify and solve problems. The result of design research in Information Systems (IS) is a purposeful artefact created to address an important organisational problem. The general design science research process described by Vaishnavi & Kuechler (2004) consists of iterations through five phases – awareness, suggestion, development, evaluation and conclusion. The model is an adaption of the computable design process model developed by Takeda et al. (1990), but the activity within the different phases are considerably different. One important difference is that knowledge contribution needs to be a key focus in design research. In Figure 3.2 the design cycle for the reasoning is illustrated.

Awareness

In this phase the design researcher becomes aware of the problem. The problem is identified and defined. The output of this phase is a formal proposal for a new research effort.

Social networks are becoming more popular and attain a high degree of market penetration. This leads to the possibility to use the available information in context with fraud prevention. Identity resolution, however, is one missing piece in fraud prevention of the financial industry in order to make use of this additional information. Because a private company provided the idea for this research topic, a common understanding of the problem is obtained during multiple meetings with the project sponsor. The meetings were held informal to gain a better overview of the problem and the challenges of fraud prevention in the financial industry. The problem definition was achieved throughout multiple iterative meetings, which led to a precise problem definition and research objectives.

Further insights are gained from the currently existing knowledge in the literature. The literature review is focused on the existing knowledge for identity resolution with offline and online solutions. Identity resolution of offline data is already extensively studied. The purpose can be distinguish between data cleansing, where the process is called entity resolution with the goal of finding non-identical duplicates and merge them into a single record and schema integration, where the process is the identification of related records across multiple database to create a link between them. Research of identity resolution on the other hand is rare. The goal is to identify the successful methods and apply them in this research.

Additional knowledge about the available data within banking institutes is gained from personal meetings with employees from a medium-sized Swiss bank and service providers from within the financial industry. The so gained information is then compared with the available information from the social network and categorised according to their suitability.

Suggestion

The suggestion phase is closely related to the awareness phase. Vaishnavi & Kuechler (2004) state that suggestion is an essentially creative step wherein new functionality is envisioned based on a novel configuration of either existing or a combination of new and existing elements. This phase results in a tentative design to solve the identity resolution problem.

The tentative design is based on the existing knowledge in offline-offline and online-online identity resolution adapted to the results of the awareness phase. It is to be analysed, how the existing approaches differ from required identity resolution between banking data and social networks data and which information and attributes can be used similar to the existing offline solutions.

The goal of this study is to create a method, which benefits form the existing knowledge in identity resolution and applies it to the new field of identity resolution between online and offline data. The outcome of this phase results in the proposal of the most promising attribute and a tentative matching algorithm for banking data with social network identities.

Development

In this phase the tentative design created in the previous phase is further developed and implemented. The required software development of the IT artefact is based on the requirements from the suggestion phase and on best practice. The idea is not to evaluate existing standard software and configure it, but to develop a novel and reusable piece of software, which might also be based on existing frameworks or open source projects.

Evaluation

The evaluation of the developed artefact checks whether the artefact achieved the objectives and analyses its strengths and weaknesses. It also assesses, if the artefact reached the expectation to achieve a suitable solution for identity resolution between offline banking data and online social network data.

A main focus is set on the comparison between the final solution with different additional data sources for the identity resolution. A baseline is defined by the use of only the name of the bank customer. The result is then compared to the results with additional data sources

like directory information and transactional data. The results are measured according to the mean reciprocal rank and the percentage of correct matches. These two measuring methods provide the ability to evaluate, whether the solution is able to identify the correct match or if it is able to propose the right match in a list of possible matches.

A sample dataset consisting of XING identities and banking data is provided by the project sponsor. This set is used as ground truth against which the method is measured. The data set is divided into a training set and independent testing dataset. According to Dobbin (2011), $\frac{2}{3}$ of the sample data are assigned to the training data and $\frac{1}{3}$ assigned to the test set. The generated data consists of bank customers with the corresponding XING profile id and the transactions. The probability that a money transaction corresponds with a contact of the XING profile is estimated at 3%.

A first-level evaluation is conducted in the final phase of the artefact development. This enables also an iterative development of the artefact and also the ability to measure how an additional data source or attribute influences the accuracy of the results. The evaluation includes the following comparison, including their combination:

- Simple Identity Resolution (Name only)
- With additional Information (Directory)
- With personal relations (Social Graph) from XING
- With money transfer information

The effectiveness of the best combination for the identity resolution is further evaluated with personal interviews with the corresponding partners. These interviews should asset how beneficial the new solution is for fraud prevention in a banking company. Assessments of the final artefact are:

- Does the artefact outperform existing solutions?
- Is the artefact usable in the financial industry in terms of absolute effectiveness and efficiency?
- Is it possible to integrate the new solution into the existing workflows?

Conclusion

The last phase concludes with the identification of the gained knowledge, unsuccessful approaches and lessons learned. It compiles the experiences and findings for the research question and each of the research objectives.

3.2 Research Philosophy / Paradigm

Based on design research as research strategy, the remaining layers are defined. The first layer of the research onion, depicted in Figure 3.1, is the research philosophy. The research philosophy, or also called research paradigm, represents a worldview and may be based on a set of basic assumptions and beliefs that are shared by members of a research community (Guba and Lincoln, 1994). The goal of this phase is not only to define or choose the philosophy, which suits best, but to improve the understanding of the way in which the study is approached.

In the Information System discipline the selection of the paradigms are not as clear as in other research disciplines. Emphasis has been placed on the paradigms of positivism and interpretivism by different researcher, such as Weber (2004) or Becker and Niehaves (2007). However, Hevner et al (2004) discuss another paradigm - the design science paradigm or also called design research, which supports the creation of new and innovative artefacts. Design Science is often used in the research field of information systems and technological advances are the result of innovative, creative design science processes. Vaishnavi (2004) suggest that Design science research is a complementary perspective to the positivist perspective for performing research in Information Systems (IS). Therefore the Research Philosophies chosen for this thesis are design science and positivist.

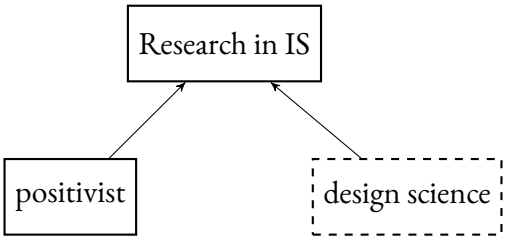

Figure 3.3: The research philosophy

Positivist: The positivist paradigm sees the world based on unchanging laws, and that everything that occurs around us can be explained by knowledge of these universal laws paradigm (Hughes, 2001). To understand these underlying laws the occurring events must be recorded and analysed systematically to work out the underlying principles.

Design science (-research): The goal of this philosophy is that new artificial constructs are designed and constructed. Compared to an existing construct, which remains valid, the new construct may be considered superior.

3.3 Research Approach

In research two types of reasoning exists – deductive and inductive reasoning. These types are used the gain a better understanding of the relationship between theory and research.

Deductive is where the researcher is concerned with developing a hypothesis based on existing theory and then designs a research strategy to test the hypothesis (Wilson, 2010).

Inductive reasoning is the opposite approach, where the theory is generated out of the research. This can be thought of as a "bottom up" approach to build knowledge. The researcher is required to use observations and data to find patterns and regularities to develop a tentative hypothesis, which will lead to a general conclusion of theory (Babbie & Earl, 2001).

In the application of design research in this thesis both approaches are used. During the awareness phase of design research, based on observation and data, a hypothesis is defined. The evaluation phase on the other hand is based on deductive reasoning, where we test our hypothesis.

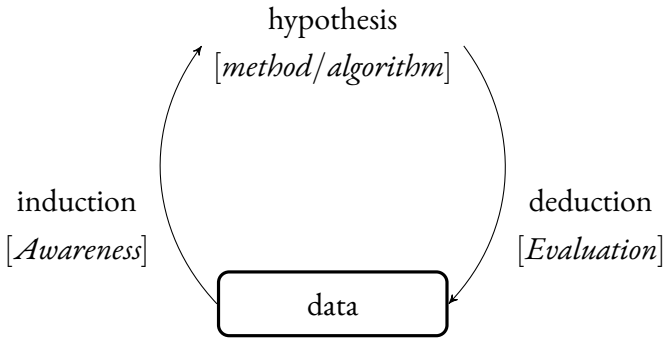

Figure 3.4: Research approach

3.4 Research Choices

In this layer of the onion, we differentiate between the data collection techniques and analysis procedures, which can be split into the terms quantitative and qualitative. Quantitative is typically used where large amounts of data are used to verify hypotheses or to find patterns. Qualitative research focuses on understanding the important characteristics of typically small samples of data consisting and also of non-numerical data.

For this thesis more than one data collection technique and analysis procedures are used. This is also called "Multiple methods" in contrast to "Mono method" where only one is used.

The different phases within the design research process require different analysis procedures. In the awareness phase the data is collected mainly from the existing literature, interviews and the findings from the development of the artefact. In the following phase, where we evaluate the artefact, we use quantitative data in form of the provided ground truth dataset to verify our hypotheses and our method.

3.5 Time Horizons

The goal of this master thesis was to create an artefact and evaluate how well objectives were reached. This study is therefore considered a cross-sectional study, because the artefact was created within a limited time. It can also be seen as a snapshot of an attempt to design and build an identity resolution artefact with the technology and data available at a certain time.

3.6 Data Collection Methods

The project sponsor provides a dataset, consisting of XING profiles and banking data. This set is used to train and test the developed method. Additionally, all public XING profiles are available for blind testing.

Qualitative data, which was used for the creation of the hypothesis and the development of the method are collected from interviews and existing literature.

4

Data Source

The identity resolution is based on the information available within a bank and the accessible information in a social media network. This chapter provides an overview of the available attributes and their mapping.

4.1 Banking Information

Within the bank exist a variety of different applications with their own information. Some of this information is connected, other is aggregated within a data warehouse for further analysis and some information is only accessible through specific applications, which is rarely the case.

The available information can be classified into three different categories – master data, transactional data and the remaining data.

4.1.1 Master Data

When opening a new account for a new bank customer, a set of data related to the new business partner is created. This set is also called master data or business partner.

The following attributes are available for natural persons and are useful in order to identify a person in a social media network. Other attributes might be also available depending on the banking institute and the used banking system.

Attributes available for all bank customers:

- Name
- Address
- Nationality
- Residence permit
- Language
- Date of birth
- Sex
- Marital status
- Education / occupation
- Account numbers

Attributes available for some bank customers:

- Married since
- Children
- Course of studies
- Date of the end of studies
- Email
- Mobile
- Phone

A persons can have an authorisation for other accounts. The authorisation can be split into having the right to use the account (Usufruct) or have a legal authority. This information is accessible on both sides of the relation. It links two or more business partners, where the previously listed attributes are available.

4.1.2 Money Transaction Data

Transaction data is the information, which is stored for a money transfer. Besides the usual attributes about the amount and the currency, it might also contain information required for transactions in a foreign currency or with other banking institutes.

The type of a money transaction can be divided into three types – Customer Transaction, Bank-to-Bank Transaction and Correspondence Transaction.

Customer Transaction

Besides the general attributes this transaction type contains the information to identify a customer by name, address and account number. Typical customer transactions are intra bank transaction or transaction within the same clearing system.

Bank-to-Bank Transaction

Bank-to-Bank Transaction do not contain a reference to the recipient. The reference is then used by the receiving bank to identify the bank account and the purpose of the transaction.

In Switzerland this type is widely used for billing purposes.

Correspondence Transaction

This is the rarest type of all three transactions. It is used if two banks do not maintain a business relation. This case is used for foreign transaction or currency transaction where specific restrictions apply.

In this case all related banks are listed including the customers and accounts.

General Attributes

All types of transactions share a common set of attributes. Depending on the combinations and banking systems the attributes may vary. The following attributes are generally stored:

- Order information
    - Source System Information (e-Baking, pay-in slip or counter)
    - Valuta
    - Booking date
    - Amount
    - Currency
    - Reference
- Remitted
    - Customer account
    - Information linking to the master data
- Remittee
    - Address
    - Account number
    - Reference
    - Bank info
- Settlement Instruction
    - Method of payment
    - Payment channels
    - Corresponding bank
- Calculation:
    - Costs
    - Fees
    - Exchange rate

Because customer information is required for this study and only customer transaction provide this information this is the only type of transaction, which will be used.

4.1.3 Other Information

Identification Picture

Every customer who is opening a business relationship with a bank has to identify himself with an identification card or passport. This document is then copied and archived as well as

scanned and stored electronically in the banking system. This information is directly accessible and can be used for identification purposes.

The quality of the image varies depending of the age of the business relationship. Old relationships, where the picture was taken a long time ago, have an image of inferior quality compared to the good quality of today's images.

Cash Machine Pictures

Automated teller machines (ATM) are often located in a semi-private area like a bank lobby or a niche in a shopping mall, which provides extensive security camera coverage. According to Peter Schumacher (Swisscom IT Services AG) this camera recordings are not linked to the transactions from the ATM machines.

Some cash machine machines like some machines Wincor have a built in camera to take a picture from the person who is using the ATM. This image is then stored on the machine itself and can be accessed for inquiries. The penetration of these machines in Switzerland however is very low and the images are not automatically linked to a transaction and accessible from the banking system. The information is therefore currently of no use for this thesis.

4.2 Xing

Each social media network has its own API, which provides access to individual attributes. This requires a customised approach to identify a person for every social media network.

For this thesis XING was chosen because it provides all the required attributes for identity resolution for fraud prevention. The two most important ones are the personal information and relationships. It is with 7.7 million active users behind Facebook the biggest social media network with the required information in the German speaking part of Europe. XING provides a relative open access to their customer information through their API, which is the biggest difference compared to Facebook, where most customer attributes are hidden.

4.2.1 Attributes

The following list represents a summary of the available XING attributes.

- ID
- First Name
- Last Name
- Address
- Gender
- Date of Birth
- Email Address
- Phone Number
- Company
- School
- Contacts

5 Identity Resolution

Identity Resolution is the first step to identify a person in a social media network to gain access to other people related to them. These related persons are then further used to identify fraud based on financial transactions.

5.1 Graph

The identity resolution is based on two graph subsets. One side represents the bank customer with all attributes and his bank transactions and their attributes. Based on this graph the goal is to identify a second graph in a social media network, which represents the same identity. This is depicted in figure 5.1. In the following work it is tested and evaluated, which attributes provide strong evidence that two identities belong to the same real world entity and which attributes are not useful for identity resolution. The goal is therefore to find attributes and information, which exist on both sides and can be matched.

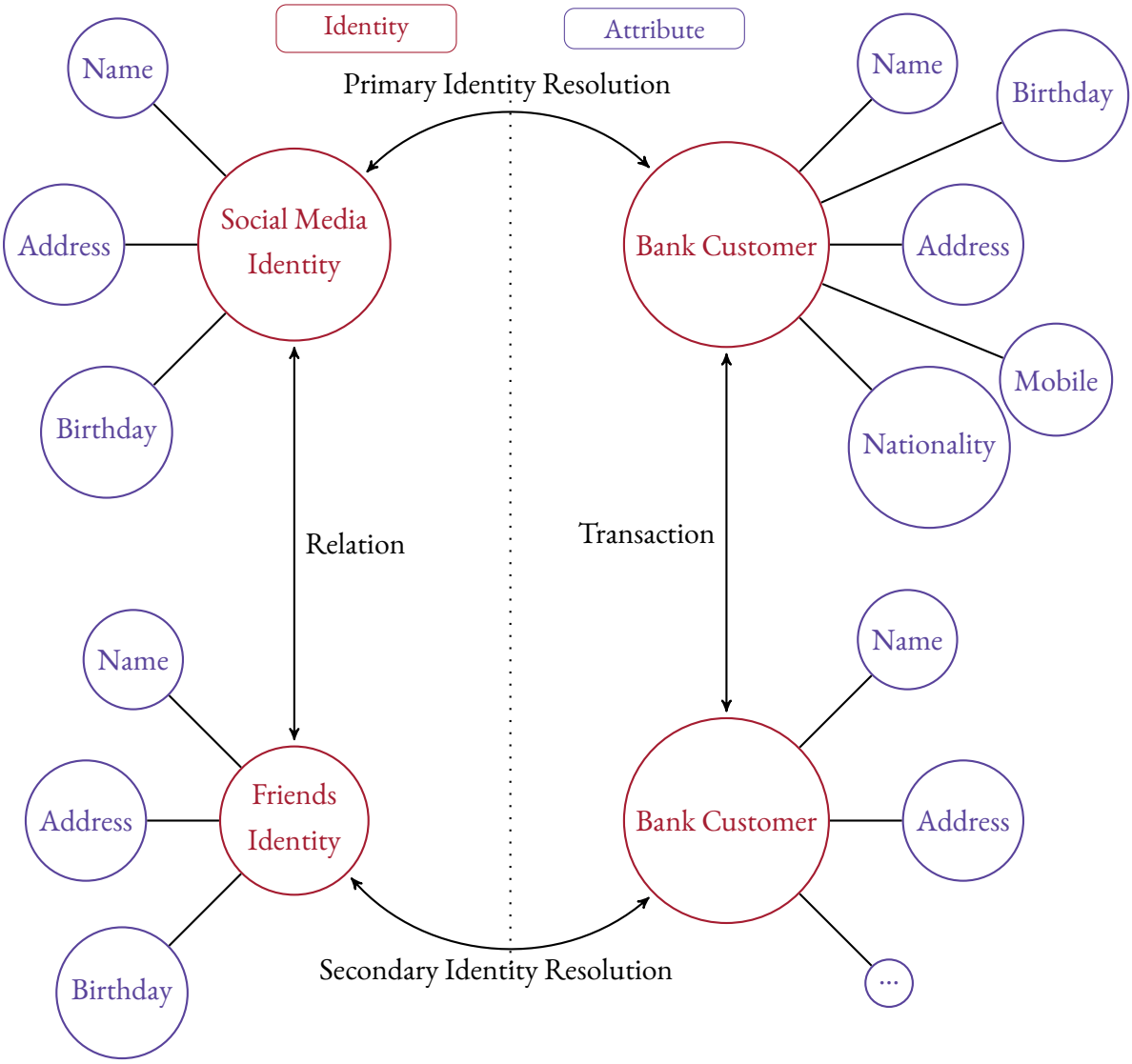

Figure 5.1: Identity Resolution Graph

5.2 Matching

The optimal method $\mathcal{M}(BankCustomer)$, which returns the correct social media profile to a bank customer, is the goal of this thesis.

5.2.1 Model

The model, how we approach the identity resolution can be split into multiple options. Weis et al (2006) described the following three options: pairwise vs. clustered, batch vs. search and effective vs. efficient.

In a general entity resolution case it is unknown how many entities represent the same real world entity, therefore it is a necessity to use clustering in order to group multiple entities. This can be achieved by first comparing all records pairwise and calculating the transitive closure or by skipping this step and directly clustering all representatives of one entity into one cluster. In our case we assume that one bank customer has only one profile in the social media network (XING).

Because we do not have access to the entire database of any social media network and the identity resolution process is triggered as an ad-hoc-task, the process can be considered as a search problem. This allows us to reduce all available profiles to a few, which are then analysed for possible matches.

Efficient models aim to reduce the number of required comparison whereas effective models aim for a higher accuracy. We would like to find out if it is possible to achieve the best possible accuracy before focusing on efficiency.

This model results in an abstract approach, which is based on all information available about a bank customer. The identity resolution method will be performed and a set of possible social media identities will be returned ordered by their transitive closure.

5.2.2 Algorithm

XING provides an API with multiple methods to access information of their users. In this chapter algorithm is described in pseudo code. The algorithm makes use of available information XING provides to search for a user and filters the results based on the attributes provided.

```
Input:    Bank Customer I
Output:   Rated set of XING profiles p

p ← searchByEmail(I_email) /* search by email */

if(p != null)
    return p

/* search by Name */
p ← searchByName(I_firstname, I_lastname)

/* rate by Transactions */
p ← rateByTransactions(p, I_transactions)

/* rate by Time zone */
p ← rateByTimezone(p, I_country)

/* rate by Birthday */
p ← rateByBirthday(p, I_birthday)

return p
```

Figure 5.2: Matching Process

Search by email

Banks have more and more access to their customers e-mail addresses. The customer adds this information voluntarily out of convenience or to gain additional information, offers by newsletters or just for simple correspondence. Some banking institutes use it also in combination with their online portals. The availability of the email address at two interviewed

general retail banks is at about 30%. In combination with XING, this information is one of the most valuable attributes because the API provides a possibility to search by email address even though the email addresses are not visible for a viewer. A further advantage of email addresses is their uniqueness. If we can identify an identity by their verified email address we can be sure that we have found the right identity.

Search by Name

Searching by name is a possibility most social media networks provide. Because XING is mainly used for business and colleague contacts the accuracy of the name is higher in comparison to social media networks used for private purposes like Facebook.

```
Input:    Bank Customer I, XING profiles p
Output:   rated XING profiles p

/* XING API—access or cached */
p ← searchByName(I_firstname, I_lastname)

for each P ∈ p
    P ← calculateDistance(P)
    ignoreIfBelowThreshold(P)
endfor

return p
```

Figure 5.3: Search by Name

The XING-API same as other social media networks provides the possibility to search for users based on keywords. In this approach this is used to search for users. Within the XING-API an approximate string matching technique is already present. If the name of the bank customer differs only slightly from the correlating XING profile, the search will nevertheless

return it as a result. This reduces the need to search for names similar to the bank customer name.

For social media networks, which do not provide the implementation of a fuzzy search logic on their side, it is important, that profiles with slightly different spelling of the name is found nevertheless. This would be the case for LinkedIn. An implementation for a fuzzy search is therefore found in the appendix to extend this algorithm to social media networks without fuzzy search within the API.

Because of the fuzzy search logic within the search API, the names of the results have to be compared with the bank customer's name. Every XING profile is then rated with a value between 0 and 1. The used method is further described in chapter 5.2.3.

Rate by Contacts

Contacts are one of the few attributes available in both graphs. The bank customer has potentially the same information stored within his money transactions. Within the XING-API a very convenient interface is available to access the required attributes for all contacts of a XING-profile with one API-call.

All XING profiles are rated based on the correlation between contacts and transactions. Details for the rating of the different attributes can be found in chapter 5.3.

Rate by Time zone

All information regarding the personal address is not available through the public API. An attribute which partly reveals this information is the time zone. The field is a combination between the continent and the capital of a country or a state. The information is therefore more precise than only the time zone or the country. It however requires a matching table between address of the bank customer and the time zone of the XING profile.

```
Input:    Bank Customer I, XING profiles p
Output:   rated XING profiles p

for each P ∈ p
    /* XING API-Access or cached */
    contacts ← getContacts(P)
    for each contact ∈ contacts
        %P ← rateTransactions(P)
        for each transaction ∈ I_transactions
            match ← calculateMatch(transaction)
        endfor
    endfor
    P ← rateTransaction(match, contacts_size)
endfor
return p
```

Figure 5.4: Rate by Contacts

Rate by Birthday

The birthday attribute is present on all bank customers but only on very few XING profiles. The information is only available for direct contacts and is further more an attribute, which is not mandatory for all XING profiles.

If the birthday is visible in the XING profile, this attribute can be used to reduce the amount of false positive matches and increase the accuracy of the result.

5.2.3 Similarity Calculation

Based on the literature review in the chapter 6.7 the name of the bank customer and the XING profiles are compared using the Jaro-Winkler string matching metric. As evaluated by Veldman (2009) the optimal threshold for a positive name identification is 0.92.

In the used similarity functions an exact match results in a similarity $s_{Name} = 1$ opposed distance functions, where a smaller value indicates a greater similarity between two strings.

The same metric is also used to calculate the similarity between the contacts of a potential XING profile and the name of the remitee and remitted of the money transaction of the bank customer. The threshold is again 0.92 and a perfect match results in a similarity $s_{Transaction} = 1$.

The similarity calculation is executed between all transactions of a bank customer an all contacts of a potential XING profile. This results in multiple distances, which are added to the rating of a XING profile as described in chapter 5.3.

The time zone of the XING profile can only take a set of values and the country of the bank customer is verified. Therefore the similarity $s_{TimeZone}$ of the time zone and the country of the customer can either be 1 for a match or 0 for no match.

The birthday similarity $d_{Birthday}$ is also compared like a string with the Jaro-Winkler metrics. Based on the result of Veldman (2009) this provided the same accuracy as a custom implementation based on the individual fields – year, month and day of the month. The threshold is however reduced to 0.62 compared to name matching.

5.3 Rating Calculation

The rating of each profile is based on the sum of all calculated distances of the different attributes. To improve the accuracy of the algorithm multiple methods are therefore evaluated.

The baseline is provided by adding all results from the distance calculation to the rating of each profile $r_{Profile}$:

$$r_{Profile} = s_{Name} + sum(s_{Transaction}) + s_{TimeZone} + s_{Birthday}$$

5.4 Algorithm Variations

For the baseline implementation of the algorithm all available attributes were used – name, contacts, time zone and birthday. The attributes of the XING profile and the correlating attributes of the bank customer are compared based on the definition chapter 5.3 with equal weighting.

5.4.1 Name Frequency

The transactions of the bank customers are probably one of the most important attributes. To increase the accuracy of the matching between the transaction and the contacts of the XING profile, the names of the transaction are weighted according to the commonness of the combination of the first and last name. A unique name, which exists on both side – XING contacts and bank customer transactions – gets a higher importance than a very common name. In the Literature this is also known as inverse document frequency (idf) (Ji & Luo, 2001).

Based on 84'864 XING profiles the name frequency (nf) is locally calculated. This is the base line to calculate the globally valid inverse name frequency inf. The default formula to calculate inf is:

$inf = log(\frac{N}{nf}) = log(\frac{84'864}{nf})$

where N is the number of names and nf the number how often a name exists in the 84'864 names. If a name exists only once within the 84'864 name the inf would be 4.929 and go down to 4.230 if 5 name have a match within the available names. This results in a flat idf curve.

To increase the relevance of match from unique name of the transaction and XING profile the following nf - inf mapping is used:

nf	inf
1	1
2	0.5
>2	0.1

Table 5.1: Inverse Name Frequency

This is important to reduce the impact of false positive matches on common names. Furthermore, the weight of rare names is increased.

5.4.2 Transaction Normalisation

Based on the distance calculation of the transactions results, depending on the number of positive matches between transactions and XING contacts, the sum of all transaction dis-

tances $sum(s_{Transaction})$ might be overvalued compared to the other values. This results in a high importance of the transactions compared to the remaining attributes.

To adjust the transaction rating *transactionRating* the following formula is used

$transactionRating = \frac{log_2(matches)}{log_{10}(contacts)}$

where *matches* represent the number of contacts with a similarity higher than the threshold and *contacts* represent the total number of contacts of a XING profile.

Depending on the number of contacts of a XING profile and the positive matches between transactions and XING contacts, this function provided the best solution for bank customers, where we expect that about 3% of all transaction partners are present on XING.

The extended profile formula would therefore be:

$r_{Profile} = s_{Name} + \frac{log_2(matches)}{log_{10}(contacts)} + s_{TimeZone} + s_{Birthday}$

Together with the aspects from the name frequency this results in the following formula:

$r_{Profile} = s_{Name} + \frac{log_2(log(sum(\frac{84'864}{nf})))}{log_{10}(contacts)} + s_{TimeZone} + s_{Birthday}$

The variable *matches* is replaced by the sum of the calculated inf from each match between the transaction and the XING contact.

5.4.3 Attribute Weighting

On both sides of the identity resolution graph multiple attributes are available. To find an optimal combination of these attributes a set of weighted combinations was generated, where the weighting for all attributes sum up to one. The set consists of one weight for the following attributes – name, contacts, time zone and birthday:

$1 = w(Name) + w(Transaction) + w(Timezone) + w(Birthday)$

The different weights have a stepping of 0.1 between 0 and 1. This sums up to 286 possibilities possible attribute weightings.

If we apply these weighting to the formula from chapter 5.3 the following formula results:

$r_{Profile} = w(Name) * s_{Name} + w(Transaction) * sum(s_{Transaction}) + w(Timezone) * s_{TimeZone} + w(Birthday) * s_{Birthday}$

6 Results

The method modelled in chapter 5.2.2 was tested and improved with a training set of 26 bank account identities, consisting of 100 transaction each and a correlating XING profile. This training set was used to improve and verify the algorithm.

The test to evaluate the accuracy and the performance are based on a separate testing set of 10 bank identities. This set is used so evaluate the final algorithm with all learnings and improvements based on the training set. This chapter describes the impact of the different attributes and the performance of the proposed algorithm.

All tests were evaluated with a prototype and a set of automatically generated test and training data. To reflect the reality, all data are generated based on the findings in chapter 4 to represent the ground truth.

6.1 Experimental Setup

The data for all experiments is generated with an automated script, which generates a parametrisable number of bank account with their transactions.

The starting point of the generation is a list with possible names. A randomly chosen name is used to search for XING-profiles. If more than 2 XING profiles for the name exist, one of the results is used. Based on the available information from the profile the bank customer is generated.

The bank customer consists of the master data and the money transaction data. For 30% of the bank customers the email address is then manually added based on all the information available in the XING profile and further information from the internet.

The transaction data is generated based on two parameters – number of transaction and the probability that a transactions matches with a XING contact. The generated data has 100 transactions for each bank customer, where the probability that a transaction represents a XING contact is 3%.

The used setup of the parameter is based on the findings in chapter 4. To improve the algorithm 26 bank customers were used as trainings data. All tests and improvements with different variation of the algorithm were tested with this training set. The final algorithm was then evaluated based on an additional testing set consisting of 15 separate bank customers.

6.2 Email Search

The email address of the bank customer provides a unique identifier of a XING profile. Because the email addresses of XING profiles are not accessible, they could not be generated automatically. However it was possible to manually figure a correct email address for more than 90% of the training data bank identities.

This email address is just one email address of the generated bank identity and might therefore be identical to the one the bank would have of their customer. As analysed in chapter 4.1.1, a retail bank in Switzerland possesses a valid email address of about 30% of their bank customers. With these email addresses it was possible to identify about 12% of the sample

identities in XING. With the email addresses of the bank customers it is therefore possible to exactly identify the following percentage of XING profiles:

$p(resolution) = p(email) * p(emailInXING) = 0.3 * 0.12 = 0.036 = 3.6\%$

6.3 Name Search

The Name of a bank customer is not as unique as the email address. Nevertheless a randomly chosen name, consisting of first name, potentially a middle name, and a last name, is in 57% only used by one XING profile. The name for this evaluation is based on a set of 84'864 name of XING profiles randomly downloaded from XING to represent the profiles at XING as precisely as possible. For the evaluation of the uniqueness of a name, one of the downloaded names was randomly chosen and the number of XING accounts with the same name evaluated.

The uniqueness of a name however decreases with the size of the social media network. This percentage can therefore vary between social media network and also change over time.

The remaining 43% of names on XING are not unique and thus the search returns multiple XING profiles. Depending on the commonness of the names the result set can be very large. Probably one of the most common name on XING is "Thomas Müller" with more than 2674 Profiles in XING.

Because the XING-API provides also results with similar names, the names are compared based on the Levenshtein approximate string matching technique. A positive match is based on the findings of Veldman (2009) achieved at a threshold of 0.8.

The results from the following chapters are therefore based on the remaining 43% of names, which cannot be uniquely identified.

6.4 Time zone Rating

The time zone attribute is important to reduce the amount of false positive results. Because no address or location attribute is accessible for the XING profiles they are substituted by the time zone attribute.

XING is mostly used in the DACH area (Germany, Austria and Switzerland). The time zone attribute is therefore not as important as it would be with a social media network where

the identities are broader distributed across different countries.

However, it turned out, that the time zone is based on the weighting the most important attribute to identify bank customers in social media networks and in this case in XING.

6.5 Contacts and Transaction Rating

Contacts are one of the most important identification attributes for linkage between social media networks. The results show, that it is one of the most important attributes to identify bank customers in a social media network too. Banking data have in this context one big advantage over other customer information because they also hold information about transactions and other personal relations, which can be found on the social media graph as well.

XING provides the advantage that the contacts are by default visible throughout the public API. For all evaluated XING-Profiles only about 11% hide their contacts. Other social media networks like Facebook recently changed the default behaviour to hide contacts from friends from 2^{nd} grade onwards through their API. Even though the access to contacts of XING-profiles are currently not widely restricted, but this may change in the near future.

By means of the profiles contacts it is possible to identify about 60 % of profiles, which are not uniquely identified by the name.

If the contacts are not visible at the correlation XING profile of the bank customer the chance for a correct match is reduced significantly. With only 11% of the XING users hiding their contacts it is not yet a problem but if this ratio increases the identity resolution might decrease massively and require some changes in the proposed identity resolution algorithm.

6.6 Birthday Rating

The birthday promised to be a strong identification attribute because of the hundreds of possibilities the attribute can take to identify or reject a false identification of a person in a social media network.

However, the attribute has the disadvantage, that it is only visible for contacts 1^{st} grade. This might be enough for other use cases, where only the birthday of known contacts is required, like a calendar based on XING contacts. For identity resolution, where a search for

contacts based on their name is required, and almost all results are not related to a certin person or organisation, which applied for the XING access token, the birthday attribute is almost always empty and therefore superfluous for identity resolution.

It was therefore also not possible to identify any profile within the training and test data based on the birthday.

6.7 Matching

The different variations of the algorithm were tested and improved with a trainings set consisting of 26 bank customers with 100 transactions each. The baseline implementation as described in chapter 5.2.2 already reached an accuracy of 20 correct identity resolution out of 26 bank customers, this corresponds to an identity resolution rate of 76.923%. This chapter describes the results from the algorithm variations described in chapter 5.4.

The following table represents the different possibilities and their effect on the accuracy of the identity resolution result:

Baseline	Name Frequency	Transaction Normalisation	Attribute Weighting	Matches	Total
X				20	26
X	X			20	26
X		X		21	26
X			X	20	26
X	X	X		21	26
X		X	X	21	26
X	X		X	21	26
X	X	X	X	22	26

Table 6.1: Matching Variations Results

Naming represents the optimisation based on the commonness of the combination of the first name and last name within the rating of contacts. If a name is more common the risk increases that a match between the bank transaction and the contacts of a social media profile,

which is based on the name, happened by chance and the weighting for this transaction is therefore reduced.

The optimisation for the transactions is based on a normalisation of their number. A match between a bank customer transaction and a profile with hundreds of contacts is not as significant as if the profile has only a few contacts. The usage of normalised transactions results in most combinations in one additional correct XING profile identity resolution.

The identity resolution was then performed for all bank customers within the trainings set with each combination for the four attribute groups.

The outcomes were multiple combinations, which resulted in the same identity resolution result, which is better than if all attributes are weighted equally. On the other hand some weighting combinations have a negative effect on the identity resolution ration and reduce it massively.

As already described in the previous chapter, the matching of the birthday has no effect on the identity resolution. Based on the results we can see, that the weighting is equally spread over all results.

All weighting combinations with the highest identity resolution rate share a high importance of the time zone followed by the weight of the contacts. The time zone is always higher rated than the contacts, which leads to the conclusion that the matching of contacts and transactions might result in false positive results, which are in these cases reduced by the weight of the time zone. The evaluation of the results confirmed this hypothesis.

If the weight of the time zone is further reduced this effect increased and resulted in three additional false positive matches if the weight of the time zone is ignored. The worst result is achieved if the weight is either put on the name or time zone.

For the identity resolution of a bank customer in a social media network we therefore do not have on optimal weighting of attributes but a set representing the same pattern of a high weighting of contacts and time zone where the time zone is usually more important than the contacts.

To find a general relation between the different attribute weightings, the median of the best result identity resolution weights was taken. This was then normalised such that the sum of all weights sums up to 1. The table 6.2 represents the attribute weights for the optimal identity resolution.

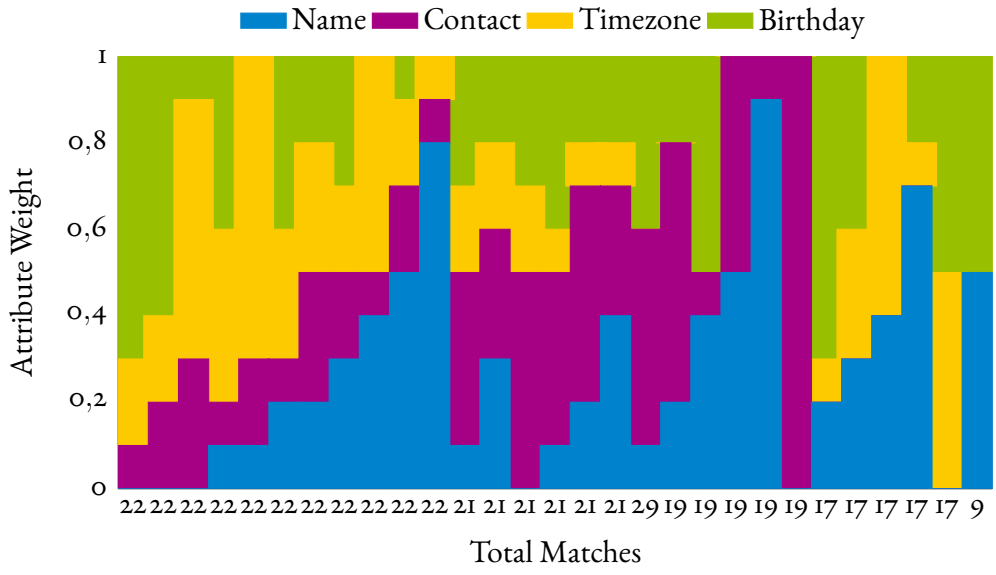

Figure 6.1: Weighted Attributes

Attribute Group	Weight
Time Zone	0.5
Contacts	0.25
Name	0.25
Birthday	0

Table 6.2: Attribute Weighting

6.8 Identity Resolution Probability

Based on all findings based on the training data, the algorithm was optimised to achieve the best possible identity resolution result. The accuracy of the identity resolution method is evaluated based on separate test data, which represents the banking information described in chapter 4 and consists out of 15 identities. The test data are an additional set of data, which is used to verify the method defined in chapter 6.7

From the 15 bank customers within the test data it was possible to identify 12 XING profile correctly, which is represents 80% correct identity resolution.

The second evaluation criteria is the mean reciprocal rank, which is slightly higher at 0.862%.

Based on all the available information, the following formula represents the universal calculation of the accuracy for the identity resolution:

$$p(correctResolution) = p(uniqueByName) + p(notUniqueByName) * p(correctResolution|notUniqueByName)$$
$$91.4\% = 57\% + 43\% * 80\%$$

Figure 6.2: Identity resolution probability

$p(correctResolution)$	The total probability that a bank customer is correctly identified in a social media network.
$p(uniqueByName)$	The percentage of bank customers, which can be assigned uniquely to the correct XING profile.
$p(notUniqueByName)$	The remaining bank customers, where multiple XING profiles exist for the name of a bank customer.
$p(correctResolution\|notUniqueByName)$	The probability that a bank customer is identified correctly in a social media network if his name is not a unique name in the social media network.

Table 6.3: Identity Resolution Parameters

The three bank customers of the testing set, which were matched to the wrong XING profiles, represent two different challenges. One bank customer hides his contacts in the correct XING profile. This reduced the attributes to match, and therefore multiple XING profiles had the same rating.

The second false positive match was due to a high correlation between the transaction of the bank customer and a random XING profile even though the profile did not have especially many contacts. The recipient's name of the transactions, which were identified in a

different XING profile had with ”Susanne Oehler” and ”Andreas Röösli” not very common names.

The third false identification was also associated with the wrong profile based on a higher correlation between the transactions and the XING contacts. The fact that other contacts have better matches between the contacts and the transactions was therefore the biggest problem.

These two challenges for the identity resolution were also present in the training data and it is not possible for the identity resolution algorithm to solve this problem without further attributes.

The 80% accuracy however are only possible if the assumption is true that each bank customer has a social media network profile. If this assumption is not true, the accuracy will decline because the algorithm tries to assign a social media profile to the bank customer and if the correct XING profile does not exists a false one will be assigned. The accuracy for the identity resolution, if the bank customer might not have a profile in the corresponding social media network, would therefore be

$$p(correctResolution) = p(correctResolution|ProfileExists) * p(ProfileExists)$$

where $p(ProfileExists)$ is the probability that the user has a social media profile.

Based on the penetration of XING in Europe, which might be below 10%, the accuracy of the identity resolution of a random European would decline below 8%.

6.9 Performance

The performance of the identity resolution is behind the accuracy of the result the secondary objective. For fraud prevention it is required to get the result in near real time.

The process is therefore a combination between accessing all information from XING and caching already retrieved information locally within a database. This has the advantage to gain a time advantage over accessing all information through the XING API and also provides access to information, which might no longer be available online. In this context the data protection act is not covered in this thesis and would have to be considered if it would be used on an enduring basis.

The duration of the identity resolution depends mainly on the following factors:

- How many profiles exist for a specific name
- How many contacts the different profiles have
- If the results are already cached

Based on the results of the prototype, the resolution time can be estimated as follows:

- 1 second to search for the bank customer
- 0.7 second for each possible profile, to compare the transactions with the contacts of the profile

With the test and trainings data, this resulted in a resolution time between five and twelve seconds per bank customers. There are however outliers like a "Thomas Müller", which would take more than 30 minutes.

The scalability can substituted with the following formula:

$$tidentityResolution = tNameSearch * searchResultSize/100 + searchResultSize * tContacRating$$
$$tidentityResolution = 1s * searchResultSize/100 + searchResultSize * 0.7s$$

Figure 6.3: Identity Resolution Performance

Another limiting factor is the amount of API-calls allowed within a time period. It is however possible to optimise the calls to reduce the required number with aggregated call. To search for a bank customer requires one call per 100 search results. Another call is required for every 100 contacts, including their required attributes, for every possible XING profile.

XING has the following restrictions to prevent unlimited access to their data:

- 120 requests / 60 seconds
- 1200 requests / 60 minutes
- 15000 requests / 24 hours

The identity resolution for a name with 10 results from the search and an average number of 200 contacts results therefore in 21 API calls. It would therefore be possible to make almost 60 identity resolutions per hour if no results are caches.

7 Conclusion

From the existing literature, it was known that identity resolution between two social media networks achieves a high accuracy and also identity resolution between offline data source was extensively studied. The identity resolution between offline and online data was however not well studied.

With this study it was possible to present a model, which achieves with 91.4% a very high accuracy in identity resolution between bank customers and social media profiles. This proves the hypothesis right that is possible to identify a bank customer within a social media network.

The key to the high accuracy lies within the similar data between the contacts of the social media network and the bank transactions. However other offline data may provide similar models of relationships.

To achieve a high identity resolution accuracy it was important to weight the time zone higher than the similarity of the contacts. The contacts are however very important to find

possible social media profiles.

The accuracy of the identity resolution depends heavily on the penetration of the social media network. Because the contacts are increasingly hidden by Facebook users, which is the number one regarding social media users, XING was used as the social media network. In a real world use case, where not all bank customers exist in the social media network the bank customers with existing social media profile would still be identified with a high accuracy, the remaining bank customers would on the other hand be assigned to a false profile. This would therefore significantly decrease the accuracy.

7.1 Recommendations

Social media networks restrict the access to their data with limitations regarding the number of API-calls within a certain amount of time. This would therefore require some kind of caching to reduce the number of API-calls to increase the number of identity resolutions. Additionally it would be possible to keep information, which might no longer be available at the social media network through the API.

The penetration of the social media network is a key factor to achieve a high identity resolution accuracy. It could therefore be required to change the social media network once the penetration drops below a certain level or if another social media network provides a higher penetration.

7.2 Future Work

The model for XING achieves a high accuracy with the existing information. The history of other social media networks showed however that the required information might be no longer available in the future. This might happen due to changes within the API or also because of a higher sensitisation of the user.

One interesting information, which was not available for this thesis, is the identification picture that every bank has of their customers. Because of the quality and format of the profile picture within a business-oriented social media networking it would be interesting to see if the picture would increase the accuracy and if it would be a possible substitute for the profile contacts.

Caching of social media information about their users is critical. It would be required to increase the number of identity resolution, but provides challenges regarding the data protection right and the general terms and conditions of social media networks.

There are multiple social media networks with common and different users. It might therefore increase the accuracy if the information from multiple social media network is combined.

References

Babbie & Earl (2001). *The practice of Social Research, 9th edition*. Wadsworth Thomson Learning.

Becker, J. & Niehaves, B. (2007). Epistemological perspectives on is research: a framework for analysing and systematizing epistemological assumptions. *Information Systems Journal*.

Benjelloun, O., Garcia-Molina, H., Menestrina, D., Su, Q., Whang, S. E., & Widom, J. (2009). Swoosh: a generic approach to entity resolution. *VLDB J.*

BFS (2014a). Haushalte und Bevölkerung - Internetnutzung. Available at: <http://www.bfs.admin.ch/bfs/portal/de/index/themen/16/04/key/approche_globale.indicator.30106.301.html> [Accessed 18 May. 2014].

BFS (2014b). Verzeigungen nach StGB. Available at: <hhttp://www.bfs.admin.ch/bfs/portal/de/index/themen/19/03/02/key/02/05.html> [Accessed 20 May. 2014].

Bhattacharya, I. & Getoor, L. (2004). Iterative record linkage for cleaning and integration.

Bhattacharya, I., Getoor, L., & Licamele, L. (2006). Query-time entity resolution.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S. (2003). Adaptive name matching in information integration. *Intelligent Systems*.

Chaudhuri, S., Ganti, V., & Motwani, R. (2005). Robust identification of fuzzy duplicates. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on* (pp. 865–876).

Cohen, W. W., Kautz, H., & McAllester, D. (2000). Hardening soft information sources.

Cui, Y., Pei, J., Tang, G., Luk, W.-S., Jiang, D., & Hua, M. (2013). Finding email correspondents in online social networks.

Dobbin, K. K. & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. Available at: <http://www.biomedcentral.com/1755-8794/4/31> [Accessed 1 July 2014].

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Fischer, A., Riesen, K., & Bunke, H. (2010). : IEEE Computer Society.

Hernández, M. A. & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *DATA MINING AND KNOWLEDGE DISCOVERY*.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*.

Hughes, P. (2001). *Paradigms, methods and knowledge*. Allen & Unwin.

Jain, P., Kumaraguru, P., & Joshi, A. (2013). @i seek 'fb.me': identifying users across multiple online social networks.

Ji, H. & Luo, Z. (2001). A chinese name identifying system based on inverse name frequency model and rules. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 4 (pp. 2219–2225 vol.4).

Jonas, J. (2006). Identity resolution: 23 years of practical experience and observations at scale. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*: ACM.

Knuth, D. E. (1959). Automatic linkage of vital records. *Science*.

Malin, B. & Sweeney, L. (2005). Enres: A semantic framework for entity resolution modelling.

Myers, M. D. (1997). Qualitative research in information systems. *MIS Quarterly*.

Narayanan, A. & Shmatikov, V. (2009). : IEEE Computer Society.

Olivier, M. S. (2009). *Information Technology Research — A Practical Guide for Computer Science and Informatics*. Pretoria, South Africa: Van Schaik, 3rd edition.

Rehman, M. & Esichaikul, V. (2009). Duplicate record detection for database cleansing. *Machine Vision, International Conference on*.

Robson, C. (2002). *Real World Research (2nd edn)*. Oxford: Blackwell.

Russom, P. (2011). Big data analytics. *AI Magazine*. Available at: <http://public.dhe.ibm.com/common/ssi/ecm/en/iml14293usen/IML14293USEN.PDF> [Accessed 24 July 2014].

Sarawagi, S. & Bhamidipaty, A. (2002). Interactive deduplication using active learning.

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students*. Prentice Hall.

Scharffe, F., Liu, Y., & Zhou, C. (2009). : AAAI Press.

Sproch, J. & Jong, J. (2010). Resolving student entities in the facebook social graph.

Suler, J. (2004). The online disinhibition effect. *Cyberpsy., Behavior, and Soc. Networking*.

Takeda, H., Veerkamp, P., Tomiyama, T., & Yoshikawa, H. (1990). Modeling design processes. *AI Magazine*.

Talburt, J. R. (2010). *Entity Resolution and Information Quality*. Morgan Kaufmann Publishers Inc., 1st edition.

Vaishnavi, V. & Kuechler, W. (2004). Design science research in information systems.

Veldman, I. (2009). Matching profiles from social network sites : similarity calculations with social network support. Available at: <http://essay.utwente.nl/59436/> [Accessed 30 January 2015].

Vesdapunt, N. & Garcia-Molina, H. (2014). *Identifying Users in Social Networks with Limited Information*. Technical report.

Wang, Y. R. & Madnick, S. E. (1989). : (pp. 46–55).: IEEE Computer Society.

We Are Social Ltd (2014). Social, digital & mobile in europe in 2014. Available at: <http://wearesocial.net/blog/2014/02/social-digital-mobile-europe-2014> [Accessed 18 May. 2014].

Weber, R. (2004). The rhetoric of positivism versus interpretivism. *MIS Quarterly*.

Weis, M., Naumann, F., & Brosy, F. (2006). A Duplicate Detection Benchmark for XML (and Relational) Data. In *Workshop on Information Quality in Information Systems*.

Wilson, J. (2010). *Essentials of Business Research*. Sage.

Yancey, W. E. (2005). *Evaluating String Comparator Performance for Record Linkage*. Technical report, Statistical Research Division, U.S. Census Bureau.

Listing of figures

1.1 Research Timeline 6

3.1 The research onion (Saunders et al., 2009) 17
3.2 Phases of design research (Vaishnavi & Kuechler, 2004) 18
3.3 The research philosophy 22
3.4 Research approach 23

5.1 Identity Resolution Graph 32
5.2 Matching Process 34
5.3 Search by Name 35
5.4 Rate by Contacts 37

6.1 Weighted Attributes 47
6.2 Identity resolution probability 48
6.3 Identity Resolution Performance 50

List of Tables

5.1 Inverse Name Frequency 39

6.1 Matching Variations Results 45
6.2 Attribute Weighting 47
6.3 Identity Resolution Parameters 48

A Appendix

A.1 Weighted Attributes List

The following table contains the results from the evaluation of the weighting from the different attributes. The table consists of the four rows for the attributes – Name (N), Contacts (C), Time zone (T), Birthday (B) – followed by the number of matches (M) and the Accuracy (A). The tests were conducted with the training set consisting of 26 independent bank customers.

The weighting have in total 286 different combinations. The first 100 combinations, as described in chapter 6.7, all reach the highest identity resolution rate.

N	C	T	B	M	A
0	0.1	0.2	0.7	22	0.8814110256
0	0.1	0.3	0.6	22	0.8814110256
0	0.1	0.4	0.5	22	0.8814110256
0	0.1	0.5	0.4	22	0.8814110256
0	0.1	0.6	0.3	22	0.8814110256
0	0.1	0.7	0.2	22	0.8814110256
0	0.1	0.8	0.1	22	0.8814110256
0	0.1	0.9	0	22	0.8814110256
0	0.2	0.2	0.6	22	0.8814110256
0	0.2	0.3	0.5	22	0.8814110256
0	0.2	0.4	0.4	22	0.8814110256
0	0.2	0.5	0.3	22	0.8814110256
0	0.2	0.6	0.2	22	0.8814110256
0	0.2	0.7	0.1	22	0.8814110256
0	0.2	0.8	0	22	0.8814110256
0	0.3	0.3	0.4	22	0.8814110256
0	0.3	0.4	0.3	22	0.8814110256
0	0.3	0.5	0.2	22	0.8814110256
0	0.3	0.6	0.1	22	0.8814110256
0	0.3	0.7	0	22	0.8814110256
0	0.4	0.3	0.3	22	0.8814110256
0	0.4	0.4	0.2	22	0.8814110256
0	0.4	0.5	0.1	22	0.8814110256
0	0.4	0.6	0	22	0.8814110256
0	0.5	0.5	0	22	0.8814110256
0	0.1	0.1	0.8	22	0.8814110256
0.1	0.1	0.1	0.7	22	0.8814110256
0.1	0.1	0.2	0.6	22	0.8814110256
0.1	0.1	0.3	0.5	22	0.8814110256
0.1	0.1	0.4	0.4	22	0.8814110256
0.1	0.1	0.5	0.3	22	0.8814110256
0.1	0.1	0.6	0.2	22	0.8814110256
0.1	0.1	0.7	0.1	22	0.8814110256
0.1	0.1	0.8	0	22	0.8814110256

N	C	T	B	M	A
0.1	0.2	0.2	0.5	22	0.8814110256
0.1	0.2	0.3	0.4	22	0.8814110256
0.1	0.2	0.4	0.3	22	0.8814110256
0.1	0.2	0.5	0.2	22	0.8814110256
0.1	0.2	0.6	0.1	22	0.8814110256
0.1	0.2	0.7	0	22	0.8814110256
0.1	0.3	0.3	0.3	22	0.8814110256
0.1	0.3	0.4	0.2	22	0.8814110256
0.1	0.3	0.5	0.1	22	0.8814110256
0.1	0.3	0.6	0	22	0.8814110256
0.1	0.4	0.3	0.2	22	0.8814110256
0.1	0.4	0.4	0.1	22	0.8814110256
0.1	0.4	0.5	0	22	0.8814110256
0.2	0.1	0.1	0.6	22	0.8814110256
0.2	0.1	0.2	0.5	22	0.8814110256
0.2	0.1	0.3	0.4	22	0.8814110256
0.2	0.1	0.4	0.3	22	0.8814110256
0.2	0.1	0.5	0.2	22	0.8814110256
0.2	0.1	0.6	0.1	22	0.8814110256
0.2	0.1	0.7	0	22	0.8814110256
0.2	0.2	0.2	0.4	22	0.8814110256
0.2	0.2	0.3	0.3	22	0.8814110256
0.2	0.2	0.4	0.2	22	0.8814110256
0.2	0.2	0.5	0.1	22	0.8814110256
0.2	0.2	0.6	0	22	0.8814110256
0.2	0.3	0.3	0.2	22	0.8814110256
0.2	0.3	0.4	0.1	22	0.8814110256
0.2	0.3	0.5	0	22	0.8814110256
0.2	0.4	0.4	0	22	0.8814110256
0.3	0.1	0.1	0.5	22	0.8814110256
0.3	0.1	0.2	0.4	22	0.8814110256
0.3	0.1	0.3	0.3	22	0.8814110256
0.3	0.1	0.4	0.2	22	0.8814110256
0.3	0.1	0.5	0.1	22	0.8814110256

N	C	T	B	M	A
0.3	0.1	0.6	0	22	0.8814102 56
0.3	0.2	0.2	0.3	22	0.8814102 56
0.3	0.2	0.3	0.2	22	0.8814102 56
0.3	0.2	0.4	0.1	22	0.8814102 56
0.3	0.2	0.5	0	22	0.8814102 56
0.3	0.3	0.3	0.1	22	0.8814102 56
0.3	0.3	0.4	0	22	0.8814102 56
0.4	0.1	0.1	0.4	22	0.8814102 56
0.4	0.1	0.2	0.3	22	0.8814102 56
0.4	0.1	0.3	0.2	22	0.8814102 56
0.4	0.1	0.4	0.1	22	0.8814102 56
0.4	0.1	0.5	0	22	0.8814102 56
0.4	0.2	0.1	0.3	22	0.8814102 56
0.4	0.2	0.2	0.2	22	0.8814102 56
0.4	0.2	0.3	0.1	22	0.8814102 56
0.4	0.2	0.4	0	22	0.8814102 56
0.4	0.3	0.3	0	22	0.8814102 56
0.5	0.1	0.1	0.3	22	0.8814102 56
0.5	0.1	0.2	0.2	22	0.8814102 56
0.5	0.1	0.3	0.1	22	0.8814102 56
0.5	0.1	0.4	0	22	0.8814102 56
0.5	0.2	0.2	0.1	22	0.8814102 56
0.5	0.2	0.3	0	22	0.8814102 56
0.6	0.1	0.1	0.2	22	0.8814102 56
0.6	0.1	0.2	0.1	22	0.8814102 56
0.6	0.1	0.3	0	22	0.8814102 56
0.6	0.2	0.1	0.1	22	0.8814102 56
0.6	0.2	0.2	0	22	0.8814102 56
0.7	0.1	0.1	0.1	22	0.8814102 56
0.7	0.1	0.2	0	22	0.8814102 56
0.7	0.2	0.1	0	22	0.8814102 56
0.8	0.1	0.1	0	22	0.8814102 56
0	0.2	0.1	0.7	21	0.8621794 87
0	0.3	0.2	0.5	21	0.8621794 87
0	0.4	0.2	0.4	21	0.8621794 87
0	0.5	0.3	0.2	21	0.8621794 87
0	0.5	0.4	0.1	21	0.8621794 87
0	0.6	0.3	0.1	21	0.8621794 87
0	0.6	0.4	0	21	0.8621794 87
0.1	0.2	0.1	0.6	21	0.8621794 87
0.1	0.3	0.2	0.4	21	0.8621794 87
0.1	0.4	0.2	0.3	21	0.8621794 87
0.1	0.5	0.3	0.1	21	0.8621794 87
0.1	0.5	0.4	0	21	0.8621794 87
0.1	0.6	0.3	0	21	0.8621794 87
0.2	0.2	0.1	0.5	21	0.8621794 87
0.2	0.3	0.2	0.3	21	0.8621794 87
0.2	0.4	0.2	0.2	21	0.8621794 87
0.2	0.4	0.3	0.1	21	0.8621794 87
0.2	0.5	0.3	0	21	0.8621794 87
0.3	0.2	0.1	0.4	21	0.8621794 87
0.3	0.3	0.2	0.2	21	0.8621794 87
0.3	0.4	0.2	0.1	21	0.8621794 87
0.3	0.4	0.3	0	21	0.8621794 87
0.4	0.3	0.2	0.1	21	0.8621794 87
0.4	0.4	0.2	0	21	0.8621794 87
0.5	0.2	0.1	0.2	21	0.8621794 87
0.5	0.3	0.2	0	21	0.8621794 87
0	0.3	0.1	0.6	21	0.8621794 87
0	0.4	0.1	0.5	21	0.8621794 87
0	0.5	0.1	0.4	21	0.8621794 87
0	0.5	0.2	0.3	21	0.8621794 87
0	0.6	0.1	0.3	21	0.8621794 87
0	0.6	0.2	0.2	21	0.8621794 87
0	0.7	0.1	0.2	21	0.8621794 87
0	0.7	0.2	0.1	21	0.8621794 87
0	0.7	0.3	0	21	0.8621794 87
0	0.8	0.1	0.1	21	0.8621794 87

N	C	T	B	M	A
0	0.8	0.2	0	21	0.862179487
0	0.9	0.1	0	21	0.862179487
0.1	0.3	0.1	0.5	21	0.862179487
0.1	0.4	0.1	0.4	21	0.862179487
0.1	0.5	0.1	0.3	21	0.862179487
0.1	0.5	0.2	0.2	21	0.862179487
0.1	0.6	0.1	0.2	21	0.862179487
0.1	0.6	0.2	0.1	21	0.862179487
0.1	0.7	0.1	0.1	21	0.862179487
0.1	0.7	0.2	0	21	0.862179487
0.1	0.8	0.1	0	21	0.862179487
0.2	0.3	0.1	0.4	21	0.862179487
0.2	0.4	0.1	0.3	21	0.862179487
0.2	0.5	0.1	0.2	21	0.862179487
0.2	0.5	0.2	0.1	21	0.862179487
0.2	0.6	0.1	0.1	21	0.862179487
0.2	0.6	0.2	0	21	0.862179487
0.2	0.7	0.1	0	21	0.862179487
0.3	0.3	0.1	0.3	21	0.862179487
0.3	0.4	0.1	0.2	21	0.862179487
0.3	0.5	0.1	0.1	21	0.862179487
0.3	0.5	0.2	0	21	0.862179487
0.3	0.6	0.1	0	21	0.862179487
0.4	0.3	0.1	0.2	21	0.862179487
0.4	0.4	0.1	0.1	21	0.862179487
0.4	0.5	0.1	0	21	0.862179487
0.5	0.3	0.1	0.1	21	0.862179487
0.5	0.4	0.1	0	21	0.862179487
0.6	0.3	0.1	0	21	0.862179487
0.1	0.1	0	0.8	19	0.798898751
0.1	0.2	0	0.7	19	0.798898751
0.1	0.3	0	0.6	19	0.798898751
0.1	0.4	0	0.5	19	0.798898751
0.1	0.5	0	0.4	19	0.798898751
0.1	0.6	0	0.3	19	0.798898751
0.1	0.7	0	0.2	19	0.798898751
0.1	0.8	0	0.1	19	0.798898751
0.1	0.9	0	0	19	0.798898751
0.2	0.1	0	0.7	19	0.798898751
0.2	0.2	0	0.6	19	0.798898751
0.2	0.3	0	0.5	19	0.798898751
0.2	0.4	0	0.4	19	0.798898751
0.2	0.5	0	0.3	19	0.798898751
0.2	0.6	0	0.2	19	0.798898751
0.2	0.7	0	0.1	19	0.798898751
0.2	0.8	0	0	19	0.798898751
0.3	0.1	0	0.6	19	0.798898751
0.3	0.2	0	0.5	19	0.798898751
0.3	0.3	0	0.4	19	0.798898751
0.3	0.4	0	0.3	19	0.798898751
0.3	0.5	0	0.2	19	0.798898751
0.3	0.6	0	0.1	19	0.798898751
0.3	0.7	0	0	19	0.798898751
0.4	0.1	0	0.5	19	0.798898751
0.4	0.2	0	0.4	19	0.798898751
0.4	0.3	0	0.3	19	0.798898751
0.4	0.4	0	0.2	19	0.798898751
0.4	0.5	0	0.1	19	0.798898751
0.4	0.6	0	0	19	0.798898751
0.5	0.1	0	0.4	19	0.798898751
0.5	0.2	0	0.3	19	0.798898751
0.5	0.3	0	0.2	19	0.798898751
0.5	0.4	0	0.1	19	0.798898751
0.5	0.5	0	0	19	0.798898751

A.2 Fuzzy search logic

Depending on the API of the social media network a fuzzy search logic has to be implemented in the identity resolution algorithm. Because the XING API already has an integrated fuzzy search logic this was not required within this thesis.

The implementation of a fuzzy search logic within the identity resolution algorithm can be done with two different approaches – iterative or complete. With the iterative approach the search for potential social media profile is extended if no fitting profile was found e.g. with a threshold. The second approach searches for all potential profiles by already extending the search criteria in the first place. This extended search might result in more potential profiles, but the chance that the correct found profile may be found increases.

The following list represents a set of possible name variations

- Search with the given name
- Search only for the first name and last name
- Switch first and last name
- Split or combine compound names

The use of such a fuzzy search increases the chance that a person is found, where the name is miss spelled, but it also increases the chance that a wrong profile is regarded as a match.

Depending on the API functionality some optimisations might be necessary to increase the performance and reduce the API-calls required.