Statistical Regeneration and Scalable Clustering of Big Data using MapReduce in the Hadoop Ecosystem

A case study of Competence Management in the Computer Science Career

DISSERTATION
zur Erlangung des Grades eines Doktors
der Ingenieurwissenschaften

vorgelegt von
M.Sc. Mahdi Bohlouli

eingereicht bei der Naturwissenschaftlich-Technischen Fakultät
der Universität Siegen
Siegen 2016

Printed on non-aging wood- and acid-free paper.
Gedruckt auf alterungsbeständigem holz- und säurefreiem Papier.

1. Gutachter: Prof. Dr. Madjid Fathi, Universität Siegen
2. Gutachter: Prof. Dr. Roland Wismüller, Universität Siegen
Vorsitzender: Prof. Dr. Udo Kelter, Universität Siegen

Tag der mündlichen Prüfung: 28. October 2016

Acknowledgments

First of all, I would like to cite following sentence from Henry Ford: "Coming together is a beginning; keeping together is progress; working together is success." The foundation of any successful work is accompanied by a lot of positive and negative efforts from financial and emotional support, to mutual scientific cooperation and constructive feedback. In the meantime, the influential and positive supports should not be forgotten to be acknowledged meaning that a success cannot be reached without such strong accompanying delegation. My first and most thanks in the chain of supporters goes to Prof. Dr.-Ing. Madjid Fathi. He has supported me not only in my PhD and scientific career, but also a lot in my personal and private life. Special thanks goes for all of his supports. Prof. Dr. Roland Wismüller is one of the most kind people that I have ever seen in my life. He did his bests to help and support me in my PhD. I will never forget his careful and supportive comments, helps and kindness with me. I would also like to thank my committee members, Prof. Dr. Udo Kelter and Prof. Dr.-Ing. Roman Obermaisser for serving as my committee members even at hardship.

I am sure that everybody agrees with me that the main part of a successful career relies on a very relaxed private life. The first position in this regard is my wife, Sanaz, and my lovely daughter, Melissa. The only word which may describe a part of their important role in my life is just "Excellent". There are also some people who are not only a key for the success in your Phd, but also in allover your life from its beginning to the end. I, after being a father, can feel it better how important was my life and future to my parents. My father and my mother are the first and most reliable human beings that I got known of them in my life. So, a lot of thanks to my father and mother, Rahim and Zakiyeh, for all their efforts in tolerating a lot of difficulties because of me. Having a very kind sister and brother who understand you very well and are always on your side is the best gift that one can expect in the life. I am one of those lucky people in the world in this regard and would like to thank Parisa and Milad for everything.

Honestly, any success is not limited just to supervisors, family members and relatives. In this regard, I am grateful to Dr. Hamed Shariat Yazdi for being supportively beside of me in all of my difficulties. I express my special thanks to Prof. Dr. Udo Kelter and Johannes Zenkert for their supports to my responsibilities in the European funded COMALAT project while writing this dissertation. I would also like to thank Prof. Alexander Holland and Prof. Lefteris Angelis for their supports. My sincere thanks also goes to Tomi Ilijas as director of Arctur HPC centre in Slovenia for providing a virtual infrastructure and initial data to me for test and evaluation. Last but not least, I appreciate all contributions and supports of my master and bachelor students, all project colleagues specially European funded ComProFITS and COMALAT projects.

Thank you

– Mahdi Bohlouli

Abstract

Any adaptive analysis of domain specific data demands fully generic, sophisticated, and customizable methods. A mathematical representation and modeling of domain specific requirements ensure achieving this goal. In talent analytics and job knowledge management era, a mathematical model should resolve person-job-fit and skill mismatch problems as well as under-qualification concerns about workers and job-seekers. This issue becomes even greater challenge for large job centers and enterprises when they should process data intensive matching of talents and various job positions at the same time. In other words, it should result in the large scale assignment of best-fit (right) talents with right expertise to the right positions at the right time. The diversity in the domain of human resource management imposes large volumes of data. Hence, extending approaches towards speeding up analytical processes is essential.

The main focus of this dissertation is on efficient and scalable modeling, representation and analysis of career knowledge by proposing a hybrid approach based on big data technologies. In this regard three types of the data have been prepared through profiling, namely as talent profiles, job profiles and competence development profiles. The main focus is divided into three matching problems: (a) Scalable matching of talent profiles with job profiles towards person-job-fit using evolutionary MapReduce based K-Means (EMRKM) clustering and TOPSIS methods. (b) Matching of competence goals of under-qualified talents, prioritized using Arithmetic Hierarchy Processing (AHP), with competence development profiles towards improving competitiveness of job seekers using K-Means and TOPSIS algorithms. (c) Matching of competence development profiles with the job profiles. In order to evaluate the achievements of this work, the hybrid approach is applied in the computer science academic career.

To this aim, a generic Career Knowledge Representation (CKR) model is proposed in this research in order to cover all required competences in a wide variety of careers. The CKR model is the base of setting up profiles and has been evaluated by careful survey analysis through domain experts. The volume of collected data from the web is so large that any type of analytics demands for the use of big data technology. Accordingly, the original collected data of 200 employees from the web as well as through assessments have been statistically analyzed and rescaled to 15 million employee data using the uniform distribution. In order to find the best-fit employee which resolves skill mismatch challenge, the talent profiles are first clustered using EMRKM algorithm. The cluster with the closest Euclidean distance of its centroid with desired job profile is regarded as the talent cluster. Talents of this cluster are sorted on the basis of TOPSIS method

towards selecting the best-fit candidate in the cluster. Similar methods are used for the matching problem in recommending competence improvement programs such as Vocational Educational Training (VET) for under-qualified talents.

An analysis of achieved results shows that 78% domain experts believe that the proposed CKR model is beneficial for their industries and showed an interest to integrate the model in their workforce development strategies. The use of the uniform distribution in the regeneration of data showed a success rate of 94.27% at the significance level of 0.05 and 97.92% at the significance level of 0.01. The proposed EMRKM algorithm handles clustering of the large scale data 47 times faster than traditional K-Means clustering and 2.3 times faster than existing MapReduce based clustering methods such as the one provided in the Apache Mahout. Moreover, any investigation in developing further metrics for various domains such as nursing, politics and engineering based on proposed CKR model as well as discovering career data through web crawling methods will promote this work. In addition, novel text mining methods in order to discover job knowledge from large volumes of streamed social media data, web and digital sources and linked open data will improve the quality of data in talent profiles and enrich the proposed approach.

Zusammenfassung

Jede adaptive Analyse von domänenspezifischen Daten erfordert generische, weiterentwickelte und anpassbare Methoden. Eine mathematische Darstellung und Modellierung von domänenspezifischen Anforderungen hilft dabei, dieses Ziel zu erreichen. In der Analyse von Mitarbeiterfähigkeiten und in der Ära des Managements von Berufswissen kommt ein mathematisches Modell zum Einsatz, um das Problem des „Person-Job-Fits“ und Qualifikationsungleichgewichte zu lösen sowie eine mögliche Unterqualifizierung von Arbeitnehmern und Arbeitsuchenden zu berücksichtigen. Dieses Problem ist eine noch größere Herausforderung für große Job-Center und Unternehmen, die in datenintensiven Prozessen die Qualifikationen von Arbeitssuchenden mit den verschiedenen Anforderungen von Jobangeboten zur gleichen Zeit verarbeiten müssen. Mit anderen Worten kann hierbei die Zuordnung von Best-Fit (die richtigen) Talenten mit dem richtigen Know-how, den richtigen Positionen, zum richtigen Zeitpunkt gelingen. Die Anwendungsvielfalt im Bereich des Personalmanagements impliziert große Datenmengen. Daher sind weiterführende Ansätze zur Beschleunigung von analytischen Prozessen von wesentlicher Bedeutung.

Der Schwerpunkt dieser Arbeit liegt auf der effizienten und skalierbaren Modellierung, Darstellung und Analyse von Karrierewissen durch einen erstellten hybriden Ansatz basierend auf Big Data Technologien. In diesem Kontext werden Arten von Daten (Fähigkeitsprofile, Jobprofile und Profile der Kompetenzentwicklung) durch den Einsatz von Profilierung vorbereitet. Der Schwerpunkt ist in drei Abstimmungsprobleme aufgeteilt: (a) Skalierbare Abstimmung von Fähigkeitsprofilen mit den Jobprofilen zur Erreichung des „Person-Job-Fits“ durch Anwendung von evolutionärem MapReduce basierend auf k-Means (EMRKM) Clustering und TOPSIS Methoden. (b) Abstimmung von Kompetenzzielen von unterqualifizierten Talenten durch Priorisierung mittels Analytischem Hierarchieprozess (AHP) sowie der Entwicklung von Kompetenzprofilen, um die Wettbewerbsfähigkeit von Arbeitsuchenden unter Verwendung von k-Means und TOPSIS Algorithmen zu verbessern. (C) Abstimmung von Profilen der Kompetenzentwicklung mit den Jobprofilen. Um die Leistungen dieser Arbeit zu bewerten, wird der Hybrid-Ansatz in der Anwendungsdomäne der akademischen Laufbahn in der Informatik angewendet.

Zu diesem Zweck wird in dieser Thesis ein generisches Career Knowledge Representation (CKR) Modell vorgeschlagen, um alle erforderlichen Kompetenzen einer Vielzahl von Berufen abzudecken. Das CKR-Modell ist die Basis zum Erstellen von Profilen und wurde durch eine sorgfältige Umfrageanalyse durch Domain-Experten evaluiert. Das Volumen der gesammelten Daten aus dem Internet ist sehr umfassend, so dass jede Art von Analytik den Einsatz von Big Data

Technologien verlangt. Dementsprechend wurden die ursprünglich erhobenen Daten von 200 Mitarbeitern, die aus dem Internet sowie durch Mitarbeiterbewertung gewonnen wurden, statistisch analysiert und auf 15 Millionen Mitarbeiterdaten mithilfe der Stetigen Gleichverteilung neu skaliert. Um den am besten passenden Mitarbeiter zu finden, der das Qualifikationsungleichgewicht lösen kann, werden die Fähigkeitsprofile mithilfe des EMRKM Algorithmus gruppiert. Das Cluster mit dem kürzesten euklidischen Abstands des geometrischen Schwerpunkts des Clusters zu dem gewünschten Anforderungsprofil wird als Talent-Cluster betrachtet. Dieses Cluster wird anschließend auf der Grundlage des TOPSIS Verfahrens zur Auswahl des am besten passenden Kandidaten sortiert. Ähnliche Clustering Verfahren werden für das Abstimmungsproblem bei der Empfehlung zur Kompetenzverbesserung in Programmen der Berufsbildung für unterqualifizierte Talente eingesetzt.

Eine Analyse der erzielten Ergebnisse zeigt, dass 78% der Domain-Experten einschätzen, dass das vorgeschlagene CKR-Modell für ihre Industrie von Vorteil ist und zeigten ein Interesse, das Modell in ihren Entwicklungsstrategien für die Belegschaft zu integrieren. Die Verwendung der Stetigen Gleichverteilung in der Datenregeneration zeigt eine Erfolgsrate von 94,27% bei einem Signifikanzniveau von 0,05 und 97,92% bei einem Signifikanzniveau von 0,01. Der vorgeschlagene EMRKM Algorithmus erledigt das Clustering der Daten 47 mal schneller als das herkömmliche k-Means-Clustering und 2,3-mal schneller als bestehende MapReduce-basierende Clustering Verfahren, wie es beispielsweise in Apache Mahout integriert ist. Darüber hinaus kann die Entwicklung weiterer Metriken für verschiedene Bereiche wie Pflege, Politik und Ingenieurwesen auf dem vorgeschlagenen CKR-Modell basieren sowie die Sammlung von Karrieredaten über Web-Crawling Methoden die Ergebnisse der Arbeit weiter anreichern. Überdies können neuartige Text-Mining-Methoden zur Extrahierung von Job Wissen aus Social Media-Daten, Web, digitalen Quellen und Linked Open Data, dazu beitragen, die Qualität der Daten in den Fähigkeitsprofilen zu verbessern und das vorgeschlagene Konzept weiterzuentwickeln.

Publications of this Dissertation

1. Mahdi Bohlouli, Fazel Ansari, Yogesh Patel, Madjid Fathi, Miguel L. Cid, Lefteris Angelis, Towards Analytical Evaluation of Professional Competences in Human Resource Management, In the 39th Annual Conference of the IEEE Industrial Electronics Society (IECON), Vienna, Austria, November 2013.

2. Mahdi Bohlouli, Frank Schulz, Lefteris Angelis, David Pahor, Ivona Brandic, David Atlan, and Rosemary Tate, Towards an Integrated Platform for Big Data Analysis; In: Madjid Fathi (ed.), Integration of Practice-oriented Knowledge Technology: Trends and Prospective, pages 47–56. Springer Berlin Heidelberg, 2013.

3. Mahdi Bohlouli, Jens Dalter, Mareike Dornhoefer, Johannes Zenkert, and Madjid Fathi. Knowledge Discovery from Social Media using Big Data provided Sentiment Analysis (SoMABiT), *Journal of Information Science (IF=1.087)*, 41(6):779–798, December 2015.

4. Mahdi Bohlouli, Nikolaos Mittas, George Kakarontzas, Theodosios Theodosiou, Lefteris Angelis, and Madjid Fathi. Competence Assessment as an Expert System for Human Resource Management: A Mathematical Approach, *Expert Systems with Applications (IF=2.98)*, Accepted in October 2016 (In Press).

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Objectives

»One resists the invasion of armies; one does not resist the invasion of ideas. «

– Victor Hugo

An invention of new technologies requires new expertise. Enterprises put efforts for offering novel and innovative products and services, especially in today's competitive world. To this aim, professional and domain specific Career Knowledge (CK) is a key enabler. The CK is a sort of qualifications, knowledge, abilities, professionalisms that an expert gained through his[1] studies or career in specific field. In addition, CK supports identification and prioritization of all required qualifications for specific career. Enterprises[2] should efficiently match their available Human Resource (HR) CK to the requirements, so much the better. As a result, any efficient talent matching and analytics reduces an imposition of exorbitant costs through skill-fitness and improves the job performance of workers.

In the frame of this research, a "talent" refers to a person who has special profession, attitude or knowledge in order to successfully accomplish a job [Dries et al., 2014]. A talent could be a current employee of an enterprise, or a graduated and skilled person who applied for a job in an enterprise. Hereof, a proper analysis and use of CK fills competence gaps and refers to an efficient assignment of the right *personnel* with right expertise (*professionalism*) to right *positions* at the right time (*period*). This is defined as *4P rule* in this research. The main goal of this work is to model, measure CK of talents and classify them based on the level of their knowledge and match them with already defined competence gaps or open job positions (see Figure 1.1 in the page 2). In particular, this work focuses on an application of *Hadoop* and *MR* in scalable modeling, representation and analysis of large scale CK.

However, HR information systems lack computerization and integration of standards for representation and analysis of the CK in Human Resource Management (HRM) area [Mishra and Akman, 2010]. In order to prepare talents for fast and sudden political, strategic or any type of changes that may arise, enterprises monitor their HR activities and collect relevant data for further analyses. Mean-

[1] All examples and arguments in this work apply equally to all genders. For the sake of ease of reading and writing the male pronouns is only used overall the thesis.

[2] The word "enterprise" refers to any type of company, firm and organization overall this thesis.

while, a computerized Career Knowledge Management (CKM) is a key part of HR information systems that provides precious and semantic insights about CK analysis.

According to the practical technology review[3] in the frame of this work, HRM softwares such as Predict360, HRSG, SkillStation, SkillsXP, Competenzia support CM and CKM partially. Their main focus from CKM perspective is on collection and visual demonstration of employees' performance data and indicators. In addition, they lack scientific benchmarking and approval of their results. All evidences in this study pointed to the conclusion that they lack a scientific and analytical background in CKM.

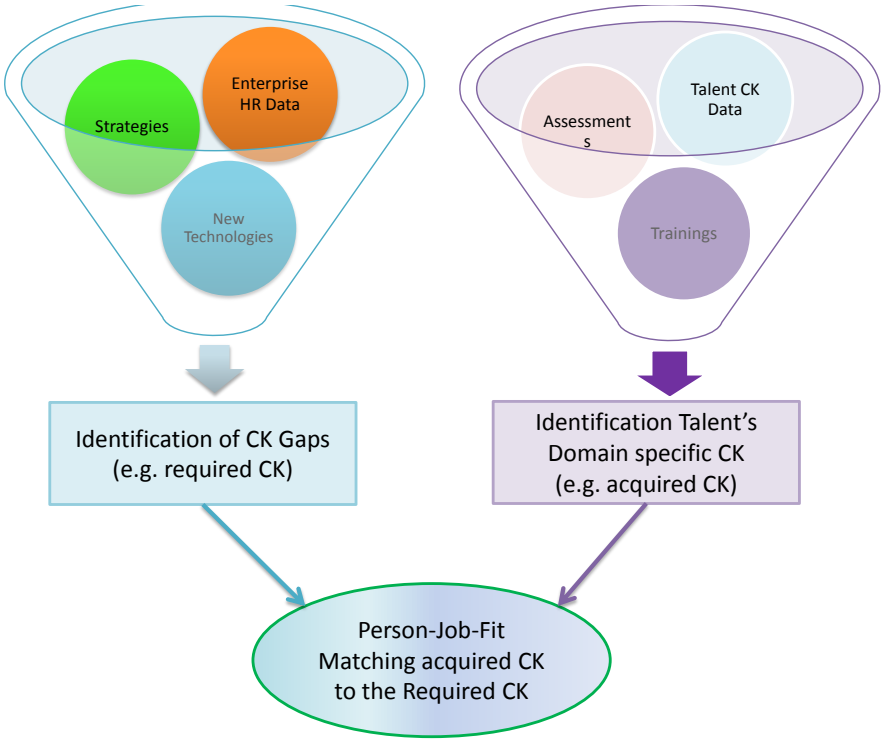


Figure 1.1: A high level conceptual overview of this research

As a result, a focus on the proper assignment and scheduling of HRs according to the needs and competence gap identification is missing from scientific and industrial points of view. This is known as skill mismatch challenge that can be solved through an efficient person-job-fit approach [Sloane et al., 2010]. Skill mismatch is referred by the Organisation for Economic Co-operation and Devel-

[3]Student project that has been supervised in the frame of this PhD in the Institute of Knowledge Based Systems (KBS) at the University of Siegen. [Eiler, 2015]

opment (OECD) as one of the most important challenges in the HRM area. Main issue here is that many of scientific solutions focus on tangible CK rather than intangible CK. Differences between tangible and intangible CK and associated difficulties are already described in section 1.2.1 in the page 9. The skill mismatch measure is divided by European Commission (EC) into subjective (e.g. talents' CK) and objective (e.g. jobs' CK) mismatch measures [Flisi et al., 2016].

In addition to the skill mismatch problem and person-job-fit approach, the way to analyze a large volumes of big HR data is another challenge that enterprises face nowadays. Everyday, over hundreds of terabytes of HR data are being created by for instance over 232 million workers in Europe, 392 million in India and 144 million in the US[4]. Such large data volumes are produced by employees' administrative, managerial and behavioral data such as demographics, e-mobilities and professional job specific CK records. Bailey estimates a need for one million new Information Technology (IT) workers in 2018 and 1.4 million IT job openings by 2022 in the US [Bailey, 2014].

According to Bailey's estimations and considering an average of 100 candidates per job opening, each with about 2 gigabytes of an application data results in processing of over 280 petabytes of data only for IT jobs and just in the US. A dimension of this problem and its scale is not limited to the US, but also becomes more complex and challenging when all other sectors and regions are being considered as well. Can current solutions support such large volumes of disparate and unstructured data? Moreover, providing efficient storage services alone is not adequate to solve the data intensive CKM and HR problems. The substantial point is an efficient analysis and processing of the data for CK discovery (e.g. value creation) specially in regards of the person-job-fit approach.

In this regard, a key path to the value creation, according to [Manyika et al., 2011], is supporting human Decision Making (DM) with automated and computerized algorithms. Therefore, the gap between the cognitive abilities and the need for making sense of huge HR data is dramatically widening and calls for technological support. In addition, Decision Support (DS) methods have not kept pace with the massive increase in data available to Decision Support Systems (DSSs) with most providing either impracticable, fairly rudimentary theoretical models, or tools that are only applicable to specific application domains such as HRM, Internet of Things (IoT), smart factories and manufacturing [Bohlouli et al., 2013b]. As a result, another challenge is how to use modern DSSs in order to contribute to the resolution of problems associated with large scale data volumes in skill-mismatch and person-job-fit approaches in the HRM area.

Putting special emphasis on the HRM is because all DM tools and libraries that have been applied in the HR area, have shortcomings in scalability, near real-time analysis, integration with heterogeneous data or easy adaptations to other domains. In addition, as stated earlier, the data volume in HRM field is growing

[4]The statistics have been achieved from the Organisation for Economic Co-operation and Development (OECD) data source through `https://data.oecd.org` in February 2015.

exponentially. In fact, modern and state-of-the-art DM and *big data* algorithms have not been employed in CKM and Talent Analytics (TAs). Talent Analytic (TA) is a data intensive analysis of experts at work for recruiting, workforce development, leadership, HR performance improvement and job design goals [Davenport et al., 2010]. TA in this research refers to assessing people's CK level (identification of talents), matching them to the required CK and improving the quality of their work through recommending competence development solutions towards improving their competitiveness.

Providing any solution that complies with aforementioned key points eliminates today's HR trends and challenges that professionals foresee for the future of HRM. For instance, Chartered Institute of Personnel and Development (CIPD) as Europe's largest network of HR experts in cooperation with the Oracle human capital management defined the term competence based "Talent analytics" and issued the scalability and big data support as an important HR challenge [Chartered Institute of Personnel & Development, 2013]. SHL talent measurement solutions has similarly highlighted the role of big data in demonstration of business values for effective and objective assessment methods in order to get more productivity and success in enterprises [Fallaw and Kantrowitz, 2013].

This chapter covers a review and discussion on the overall scientific challenges and objectives of this PhD work. Fundamental problems focused in this research are presented in section 1.1. This section provides motivation of the research and a first impression and reasons for researching this topic and indicates importance of addressed problems. section 1.2 explains vision and objectives of this research. Formulating Research Questions (RQs) and Industrial Challenges (ICs) in this section clarifies a main focus of this work. The ways that this PhD work solves state-of-the-art challenges and the research approach are discussed in section 1.3. A conclusion of the chapter and an overall outlook on the structure of this dissertation is discussed in section 1.4.

1.1 Motivation and Defining the Problem

As discussed earlier, the skill-mismatch problem is one of the main motivations of this research. One of its side effects is on increased unemployment rates specially for mis-allocated talents or youth who their professional directions are not yet clear to the industry. In order to improve unemployments, TAs should provide better person-job-fit results. According to Eurostats[5], the unemployment rate of the euro area in May 2016 for population aged between 20 - 64 was 10,1% whereas youth unemployment at the same time and region was 20,7%. The EC's plan is to reduce an unemployment ratio down to 7,5% by 2020. Such a goal needs an efficient and productive TAs and CK gap identification methods in order to support workforce development processes specially for youth and new graduates.

[5]Unemployment statistics provided by the Eurostats, accessed in June 2016 via `http://ec.europa.eu/eurostat/statistics-explained/index.php/Unemployment_statistics`

Furthermore, the Organisation for Economic Co-operation and Development (OECD)[6] reported an average European unemployment duration of 18,8 months in 2015. This rate was 13,3 months in 2009. Comparing both rates implies that the chance of getting hired becomes worth for job seekers or unemployed people. Such a fact indicates that: (1) Job seekers lack being competitive for market-oriented and therefore required professional CK and cannot convince employers in the recruitment process or (2) The unemployment salary and situation is enough suitable that unemployed people prefer to stay with jobless situation rather than getting employed. In both cases, not to have a proper competence development plan and market-oriented professional trainings during the study and education period, especially for the youth and job seekers, are sources of the problems.

The EC reported EUR 205 billion investment by member states in 2011 on labor market policy [Ronkowski, 2013]. The goal is to increase employability of individuals through supporting them in development of their professional CK. Such an investment highlights the importance of this challenge for governments from economic point of view that they are eager to make such an investment to improve the situation. For instance, unemployment costs EUR 329,5 billion per year for Belgium, Germany, France, Spain, Sweden and United Kingdom (UK). Any unemployed person costs yearly between EUR 18,008 (UK) and EUR 33,443 (Belgium). Rapid technological changes (e.g. skill based changes) cause CK gaps in enterprises and are defined as one of the major unemployment reasons in Europe [Gebel and Giesecke, 2011]. Hence, proper CK development plans facilitate adaptability and promotion of skilled workers to new required CK (competences).

Piirto et al. reported much higher unemployment rates for youth than other age groups. Additionally youth unemployment has risen in recent years. According to his opinion, the unchanged and low job vacancy rate (average 1.5% in EU-27 in 2011) reflects the unmet demand for labor and *"potential mismatches between the skills and availability of those who are unemployed and those sought to be employed"*. As it is being understood from his study, the main cause of this problem is non-efficient matching (mapping) algorithms between acquired and required CK assets [Piirto et al., 2013].

Lindgren et al. addressed the danger of misaligned HR systems which needs a great involvement of CK analysis and addressed a challenge of missing studies in Competence Management Technologies (CMTs). Many enterprises face serious difficulties in understanding their acquired and required CK which results inefficient use of HRs and lacks an integration of scientific analytics in HRM processes [Lindgren et al., 2004]. Mishra and Akman stated that HRM still lacks an application of the IT. According to their survey study with involvement of 206 domain experts, IT has a significant and positive improvements on all sectors in terms of HRM [Mishra and Akman, 2010]. Snell et al. estimated that HR-related issues like payroll, assessments, performance monitoring and career development planning of each employee cost around $1,500 annually for typical organizations

[6] Average duration of unemployment, Organisation for Economic Co-operation and Development, accessed in April 2016 via `https://stats.oecd.org/Index.aspx?DataSetCode=AVD_DUR`

which can be doubled and even tripled in less efficient organizations [Snell et al., 2001].

Mishra and Akman summarized that *"one of the impacts of IT is that it enables the creation of an IT- based workplace, which leads to what should be a manager's top priority-namely, strategic competence management"* [Mishra and Akman, 2010]. In this regard, computerized and scientific CK analysis identifies (strategic) competence gaps in enterprises and improves better allocation of available HRs. Today, it is being done traditionally in enterprises by HR experts or top-managers. Processes like paper-based Employee Development Reviews (EDR), face-to-face interviews or reviewing curriculum vitae of talents manually are current methods used in this regard. An important step is to collect talent data and cluster them in order to efficiently provide insights for key decision makers and top managers.

Competence management and TA, specially in job centers and large enterprises are data-intensive processes which need a collection and processing of huge HR data such as behavioral, professional, managerial competence data. Two sources of data can be distinguished: human-generated and machine-generated data, and both present huge challenges for data processing. Leveraging such data resource will generate tremendous value: for instance, $ 260 billion could be saved every year in the U.S. health care career by applying intelligent data analysis, according to a recent study by McKinsey Global Institute [Manyika et al., 2011]. Big data is the only choice of technology that could provide scalability for such data-intensive applications.

The big data phenomenon cannot be defined by data volume alone. Additional layers of complexity arise from the speed of data production and the need for short-time or real-time data storage and processing, from heterogeneous data sources, from semi-structured or unstructured data items, and from dealing with incomplete or noisy data due to external factors. In addition to the data volume, it associates with the velocity and variety of data sources and analytics as well as creating the value through big data analytics. With the popularity of using social media in todays' analytic solutions, this point becomes even more important to the enterprises. Such a complexity and volume of the data is also concerned in HRM area, since enterprises have to monitor and evaluate their employees in order to improve the quality of their workforce, provided services as well as products.

The HR data is categorized by Bersin into (1) People (2) Program and (3) Performance data. Each category consists of hundreds of data types with over millions of elements such as demographic, job history, skills and capabilities, compensation, engagement and social data (see Figure 1.2) [Bersin, 2012]. By today's increasing popularity of using social media in a daily life by employees as well as customers and almost positive effects of using social media for marketing and CK analysis [Bohlouli et al., 2015b], the Bersin's classification of data sources becomes even larger, more complex and disparate in which needs efficient and modern data processing solutions. The complexity arises when those data should be stored, analyzed, visualized and accordingly proper recommendations provided to top managers in enterprises.

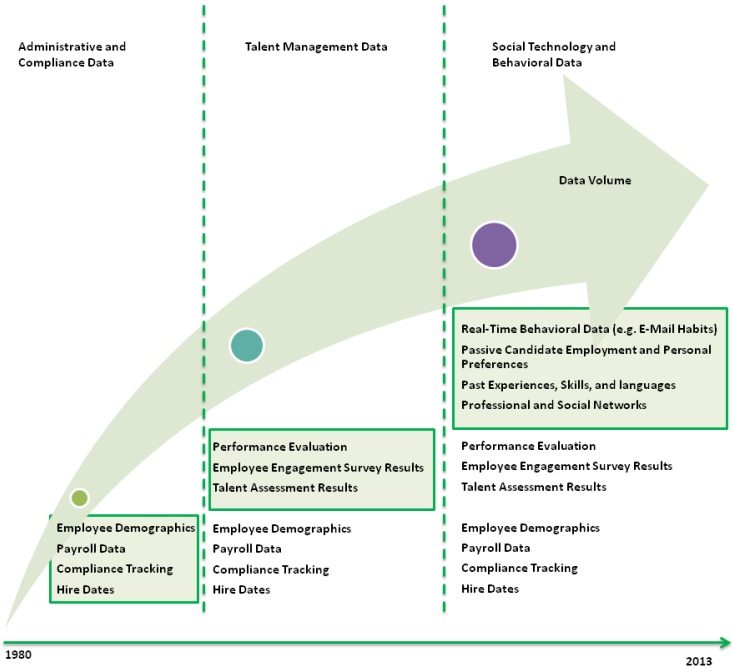


Figure 1.2: The exponential growth of Human Resource Data[7]

In order to understand, predict and improve the performance and efficiency of HRs, organizations analyze for instance talents' background, their spoken languages, education background, time-in-service, prior jobs, performance ratings, awards received, certifications, EDR, test and assessment results (e.g. 360-degree assessment [Hazucha et al., 1993]) and many more data sources in a daily basis. Such data intensive decision support and analytics imposes a demand for scalable solutions and higher processing demands. This will be even more challenging in utilization of traditional decision support algorithms for prioritizing candidates with respect to preferences of specific job position and retrieved HR data. They won't be efficient in this case to provide proper matching results.

Research projects and scientific literature that have been analyzed in the Chapter 2 show that adaptability and scalability are main missing issues in this area. Most of those projects are based on traditional IT infrastructure. They don't use emerging technologies like cloud computing [Bohlouli and Analoui], [Bohlouli et al., 2011] and [Bohlouli et al., 2012a] or big data analytics [Bohlouli et al., 2014] for providing modern CM solutions. In terms of the adaptability, a system should be able to be integrated and applied in a wide variety of case studies (sectors) without any need for further development efforts. Meanwhile,

[7]Image adopted from: The Analytics Era: Transforming HR's Impact on the Business, released by the CEB Corporate Leadership Council in `https://www.cebglobal.com/shl/images/uploads/linkau13-CLC-The-Analytics-Era.pdf`

such scalable system should facilitate an easy integration with current systems and data collection from external sources such as social media. Social media streaming scales the volume and variety of data that can not be handled with traditional data analysis methods.

In addition, most of current and recent researches use methodologies like ontologies and domain specific matrices (indicators) support only specific target sector. Since such studies based on ontologies are based on domain specific preferences, they cannot be applied in other fields easily (e.g. adaptability problem). They may work accurately in one specific field, because domain specifications should be studied very well in order to complete the domain ontology. But since this is domain specific study, the results achieved in one sector can not be exploited to new areas. That's why projects using ontology as their scientific methodology can only be applied in one specific sector.

This research does not focus on technical HRM processes and how they can be integrated in a Competence Management System (CMS). In addition, it does not provide any perspective on the behavioral analysis of HRs from psychological point of view. The findings and research results of this work do not cover issues like what are the conditions of convenient workforce for different employees. Additionally, the research does not concentrate on the cultural, regional and country or company specific issues of the CKM. Therefore, the main focus is in mathematical and computational point of view rather than HRM perspective. Selection of HR area as a case study provides very good evidence for application and usefulness of this work in real world problems and its innovation to the industry in addition to the science.

As a summary, motivational problems and challenges are: (1) skill-mismatch challenge that requires person-job-fit approach, (2) a generalization problem that needs CKM methods and models to be applied in a wide variety of case studies (sectors), (3) extreme pace expansion of HR data which lacks scalable computerized DS algorithms, (4) lack of e-recruiting and modern approaches in analyzing and classification of talents' data and (5) importance of competence gap identification and prediction in enterprises based on rapid and fast changes in technologies. These problems are reflected by defining research questions in the next section.

1.2 Vision and Objectives

The vision of this work is to *"liberate"* information that is currently hidden in the big HR data (e.g. CK data) so that they can be used by decision makers in many different domains and to move towards a new era of DSSs in the HRM. There are three main focuses in this multi-disciplinary research: (1) discovering career and job specific knowledge, (2) preparation and regeneration of large scale HR datasets for big data algorithms and (3) assignment of proper scalable algorithms for matching available talents and HRs to required expertise in the workforce development. The motivational challenge of this work which is also addressed in

the literature such as [Sloane et al., 2010] is the skill-mismatch problem.

In order to clearly describe the main focus of this dissertation, Research Questions (RQs) and Industrial Challenges (ICs) have been defined. Separating RQs and ICs supports in clear identification of scientific and industrial novelties of this work. The RQs represent scientific highlights of the work. In contrast, the ICs indicate the practical side and cover potential contribution of this research to the real word challenges. The ICs are defined through close collaboration with industrial partners in the frame of different practical research projects. These RQs and ICs are referred overall the thesis. The solution approach to any of them is also given in later chapters. In fact, they construct the core structure of thesis and will be referred repeatedly.

In order to analyze CK, it should be represented in a computer understandable format. One of the most effective methods is mathematical representation of CK by means of for instance matrices and accordingly statistical analysis methods for processing of the data. The process of modeling CK using methods like profiling is the first objective of the work. As a next and according to the exponential growth of HR data, utilization of big data technology should provide scalable DS algorithms. In the case of using big data as technological enabler, traditional DS algorithms and methodologies should be adopted and work correctly with the goal of providing scalability to the system. Additionally, the test data volume should be large enough to test the efficiency of proposed algorithms. The conception and assignment of proper DS algorithms is one side of the problem and ensuring their scalability is another side of the problem.

1.2.1 Mathematical Profiling and Clustering of CK

The mathematical representation of CK is useful when it can be efficiently used in further analyses and DM situations. One of important decisions that enterprises always deal with is how to assign people to the most relevant positions based on their knowledge and expertise. They should find out first what competences do they need to fill specific needs in an enterprise. Based on those identified needs, they should also find out who is the best matching person to these requirements. This is known in the literature as skill mismatch challenge [Sloane et al., 2010]. To this aim, a proper identification of required competences for those positions (i.e. JPs) is one side of the problem and proper clustering and assignment of individuals to such JPs is another side of the problem.

This challenge deals with methods to evaluate, measure or assess available CK of talents. The main difficulty arises when both *tangible* and *intangible* CK should be discovered. *Tangible CK* refers to qualifications like degrees, certificates and all others that are issued by certificates. In contrast, *intangible CK* refers to personal expertise, skills, behavioral issues and competences that are not (and cannot be) certified. The CK is not only about competence level of individuals, but also career gaps and required competences in enterprises. Those data can be processed, analyzed and visualized in order to provide better insights in complex

decision situations such as skill matching. In addition to modeling of acquired CK, enterprises should identify and define competence gaps (i.e. required CK).

A computerization of required CK needs workforce analysis and awareness in order to identify competence gaps. As soon as required CK is identified, talents should be matched to the gaps in enterprises. In this case, they should be first assessed by themselves through self-assessment or by others through multi-assessment methods like 360-degree feedback [Hazucha et al., 1993]. Since talents will be assigned exactly to their fitting expertise, any solution to these problems improves their job performance in enterprises. A solution approach to this challenge is referred as person-job-fit approach. A person-job-fit is sorts of algorithms in which aim to cluster talents based on their profiles and weight of competences in the target JP and allocate the best of the bests to already defined open job position.

The person-job-fit approach considers aforementioned *4p Rule* and should provide scalable algorithms in distributed computing environment. Therefore, the next challenge is to provide scalability in the person-job-fit algorithms. As a summary, Research Question (RQ) and ICs associated to skill mismatch challenge are as follows:

RQ 1 (Skill mismatch). *How efficient can a clustering algorithm allocate the best of the bests from pool of talents to specific requirements in order to solve skill mismatch problem?*

IC 1 (CKR model). *Is there any standard or reference model that classifies the CK (competences) and provides a general hierarchy of the competences in an enterprise?*

IC 2 (profiling). *How to identify competence gaps in enterprises and discover available CK in enterprises towards modeling of them?*

1.2.2 Mathematical Modeling and Regeneration of Data

As stated earlier, one of the main motivations of this research is exponential growth of HR data in regards of the volume, velocity and variety. The big data analytics is a right choice of technological solution to this problem. But, it offers a lot of choices in which vary in specifications and target use-cases. Therefore, more concentration should be paid on the selection of proper solution in the frame of big data analytics. In order to test the performance of chosen architecture, there should be real or semi-real big data (simulated) volumes. Therefore, it should be ensured that there are enough data sources in the case study of this research to test big data algorithms. Regardless of the case study of this research, does HR area deal with large volumes of the data in which providing any solution in this regard make sense of it?!

An answer to aforementioned questions and estimated volumes of the data and job openings is partially discussed in section 1.1 and will be continued for

further discussions in the next chapters. But this point itself is a big challenge. Because HR and employees data is one of the most sensible data that is not being shared even for the research goals. The fortunate of this research is that a small volume of the real 200 employees data has been collected and prepared in cooperation with industrial partners as well as from web-based resources. Based on this small volume of the data and statistical analysis of them, regeneration of the data should produce larger volumes with similar behavior as real big data. Such type of the data that is being regenerated based on statistical distribution of real data is called *semi-real big HR data* in this work.

In fact, the main challenge is how to collect, produce, simulate and regenerate enough data that are large enough to test the efficiency of designed scalable algorithms. To this aim, quality, scale and structure of the data are main artifacts that should be considered in regeneration of the test data. For this reason, the data should be simulated and scaled up with the similar behavior as real data. The quality of such *semi-real big HR data* should be good enough to judge on the accuracy of algorithms. Additional sources can also be used in order to retrieve the CK data from the web and digital sources. To this aim, there are a lot of different sources to collect CK data depending on the type and description of the jobs. For instance, the data about social competences of individuals can be streamed from social media.

As another example, suppose preparation of the professional CK data associated to the case study of this dissertation which is academic career in the computer science area. The web based data streaming as well as retrieval from digital repositories such as DBLP, AMiner provides a large volumes of data. Will it be possible to use all streamed data or should it be filtered and cleaned? Can social media be used for identification of social and professional competences. In addition, other available web based sources such as DBLP, IEEE Xplore Digital Library and ACM Digital Library or other professional sources can provide tons of the data for scientific competences. Every talent has over 3 gigabytes of profile size that should be processed for clustering. As a summary of this challenge, associated RQs and ICs in this regard are as follows:

RQ 2 (statistical distribution). *Is there any statistical method to accurately prepare and regenerate CK data in convincing volume that could test big data algorithms and their performance in scalability?*

IC 3 (big data & TA). *Do enterprises need to integrate big data in their talent analytics? If yes, how does it support and improve the talent analytics and will benefit enterprises?*

IC 4 (CK retrieval from the web). *How to retrieve CK data from digital and web based sources and assign them to the talent profiles?*

1.2.3 Scalable Matching, Recommendation and Analysis in the Career Knowledge Management

The main task is an efficient use of those initial steps in modeling of CK and accordingly retrieving relevant data sources to improve workforce development and increase HR job performance and skill-fitness. The first issue in this regard is to understand all available competences provided by talents and also collective competences of enterprises, in general. As a next issue, identification of CK gaps and matching those available resources to the required CK should be focused. Providing further recommendations to talents with specific competence goals to improve their competitiveness is the next challenge.

Decision support and recommendation algorithms in this frame should be scalable for handling large volumes and variety of data. The use and adaptation of traditional algorithms cannot solve this problem. Additionally, it should be clear how to suppose the settings of scalable decision support algorithm that fit to such data intensive CKM. From technological point of view, for instance which type of NoSQL database technology is the most suitable one for developing the data layer of the system?! or how to develop and deploy (distributed or parallel) algorithms in distributed file systems and processing environments? In addition, providing a benchmarks to test the efficiency of the system is another challenge.

As soon as the data is prepared and analytics are done for specific talents, the next challenge is to find out competence lacks of under-qualified talents and provide them further recommendations. The point here is design and application of the recommender system that is scalable for large volumes of the data. As an example, in the frame of this research a total number of 75,000 courses have been profiled in order to test the system. This delivered about 1,5 gigabytes of the data for each course concerning the course materials as well. Benchmarking indicators should cover the data management and processing on the one hand and also qualitative and analytical aspects on the other hand. The qualitative indicators mean that how accurate is the data analysis algorithm? Can one trust on the results provided through such algorithm?

As a summary, associated RQ and ICs are as follows:

RQ 3 (Scalable Clustering). *Is it possible to extend traditional DS or clustering algorithms to support large volumes of the data and how this facilitates HRM area?*

IC 5 (VET recommendations). *How to support employees to improve their CK (competence) goals by providing further recommendations?*

IC 6 (best-fit talent). *Who is the best fitting (person-job-fit) person for specific competence gap identified in an enterprise?*

1.3 How Objectives will be Achieved?

In order to achieve already defined goals of this work, innovative use of a common big data platform is demonstrated. It facilitates integration and visualization of disparate multi-modal data sources and streams. It also implements practical use of intelligent DS services in the HRM field. A solution approach of this thesis consists of four main components (steps) as of (1) *CK Assessment and Gap Identification* (chapter 3), (2) *CK Data Preparation and Regeneration* (chapter 4), (3) *Scalable Competence Analysis and Clustering* (chapter 5), (4) *Recommend Competence Development Potentials* (chapter 5). Main contribution of this dissertation are steps (2), (3) and (4). More scientific and algorithmic details of these steps are given in the following chapters.

The step (1) is based on standard assessment tools and methodologies. Defining the statistical distribution of real HR data in step (2) supports in regenerating the semi-real big HR data. In addition, the use of further data sources such as social media streaming or retrieving the data from web-based sources in this step produces disparate data volumes in which the use of big data makes sense of it. The produced data in the step (2) is used by the algorithms in the next step in order to find the best-fit talent to already opened job position in an enterprise. The recommendation process in the step (4) provides further competence improvement solutions such as Vocational Education and Training (VET) programs based on an analysis of competence goals and existing resources. A mathematical representation of CK facilitates a generic modeling of HRs and respected CK with numbers that can be used in measurements, formulas and easy integration to other sectors.

As discussed earlier, the main concern of this dissertation is to extend and utilize computerized algorithms for CKM that facilitates mapping of CK gaps with available HRs in enterprises. This process consists of measuring, assessing and representing of tangible and intangible CK in a computerized form. This is a numeric (mathematical) representation of the CK. The profiles are represented as matrices and can be easily processed for different objectives. This mathematical representation of CK provides fast and easy adaptation of the concept to a wide variety of sectors and case studies. For assessing of CK, two different methods are adopted: (1) self-assessment though questionnaire system and (2) multi-assessment through 360-degree feedback assessment method. In addition to assessments, streaming and web-based CK data retrieval delivers large volumes of the data.

In the self-assessment method, assessing the professional and tangible CK of talents is being handled by involving them in the professional and pre-configured tests. Talents receive recommendations and graphical visualization of the results upon to the completion of their tests. A theoretical study in order to prepare a pool of questionnaires for academic computer science career (a case study of this research) is the contribution in this regard. To this aim, a survey study and analysis of different recruitment actions such as interviews in academic computer science

career is used to setup questionnaires. For customization of this component to other fields, domain experts should once define and setup questionnaires specified for a target field.

For assessing intangible CK of talents such as behavioral and social CK, a 360 degree method via member's immediate work circle is more efficient and valuable. The rule is that the person who is responsible for assessing an assessee should know him for at least one year. This is because scientific studies showed that this test is accurate if an assessor knows an assessees for between 1-3 years [Eichinger and Lombardo, 2004]. In this method, a number of colleagues will provide feedback by replying to pre-configured assessment templates. The results of this phase prepare inputs of mathematical matrices. Similarly, a contribution of this part is an outcome of studies to provide configuration of standard methods to assess intangible CK, named as CKR model.

As a technological overview, the use of NoSQL database (big data) in the data layer integrates disparate data from multiple sources. Adaptation of traditional DS algorithms with new emerging data technologies is addressed by integrating the algorithm in big data technologies. Mathematical representation of CK provides a generic solution. For DM and CK analysis, a hybrid approach of Hierarchical Cumulative Voting (HCV), Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [Hwang and Yoon, 2012] and K-means algorithms supports filtering of talents for specific JP. This approach also clusters them based on assessment results and provides further CK improvement recommendations for under-qualified talents.

Additionally, real CK data of 200 talents has been anonymized and statistically analyzed to regenerate *semi-real big HR data* based on statistical distribution of the real data. This statistical regeneration of the data resulted 15 million talent data. The use of NoSQL database technology facilitates the solution approach from following points of view:

- Scalability to very large data, requires a distributed system landscape for efficient processing and an intelligent system management for cost effectiveness;
- Near real-time analysis of new data for situation aware DS;
- Integration of heterogeneous data sources from various origins, including disjoint databases, unstructured data, and outputs of other programs;

1.4 Conclusion and Dissertation Road-map

A "Compromise Model" described by Dunleavy has been used for structuring the sequence of chapters in this dissertation [Dunleavy, 2003]. According to his model, the thesis is divided into three parts as: (1) introduction and systematic literature review (chapters 1 and 2), (2) core concept and solution approach (chapters 3, 4

and 5) and (3) analysis of the results and discussion (chapters 6 and 7). The first part provides introductory and background information of this research. Chapters 3, 4 and 5 as the second part II, describe the scientific focus and mathematics of the work. This part covers practical discussion of this research as well. The last part of this thesis focuses on analyzing the results and provides research conclusion and an insight about future work.

chapter 1 covers a statement of scientific and technological problems, vision, objectives and methodology of the research. This chapter studies shortly state-of-the-art and domain specific challenges in Section 1.1 followed by general statement of contribution and indication of research approach in section 1.2. Section 1.3 provides the methodology of this research and a very short overview of the solution approach that has been researched in this work. The main research directions that have been defined by defining RQs in this section are key for structuring this thesis. This is also followed by identification of ICs for each RQ. In this regard, all of three stated scientific challenges in section 1.2 are bases of discussions in chapters 3, 4 and 5.

Guidelines provided by Kitchenham and Charters are used in writing the systematic literature review in chapter 2 [Kitchenham and Charters, 2007]. This chapter focuses on the CM and scalable data analytics (e.g. big data). Section 2.1 reviews different definitions and a history of the CM as well as theories, processes, applied research works and projects in this area. This section provides required background in order to identify scientific gaps in this area. Section 2.2 presents a survey on big data and scalable analytics. Enhanced scalability of algorithms, a more efficient usage of system resources, and an improved usability during the end-to-end data analysis processes are basic criteria focused in this section. The contribution of this work to the science beyond state-of-the-art is briefly discussed in section 2.3.

chapter 3 covers theoretical fundamentals of this research on profiling and standardized representation of the CK. After a short introduction in this chapter, a first section (Section 3.1) represents a generic CK model called CKR model that fits to a wide variety of sectors. The CKR model consists of total 64 low level (level 3) competences and covers most of possible and required CK for different sectors and job definitions. section 3.2 discusses a theory of profiling to collect HR competence data. It covers the process of tangible and intangible CK discovery and its conversion to mathematical values. Profiling theory is used not only for collecting the HR competence data, but also specification of the job specific competence data. How the CKR model can be applied in a specific sector like academic computer science career is already discussed in section 3.3.

chapter 4 is about statistical analysis, regeneration and streaming of the CK data called as *semi-real big HR data* in this research. section 4.1 covers clustering algorithms used for an analysis of real data and testing the accuracy of regenerated data. The statistical distribution of real anonymized data is given in section 4.2. The CK data streaming from sources such as social networks and web-based sources is the main issue in section 4.3.

Those collected and regenerated data should be efficiently processed, analyzed and visualized. chapter 5 targets the hybrid approach of scalable matching and recommendations. The mathematical algorithms used for analyzing CK data are covered in sections 5.2 and 5.3. Who is the best fitting exert to specific job profile can be answered with the support from mathematical algorithms stated in this section. Evaluation and analysis of the results and quality measures of the solution approach are discussed in chapter 6.

Furthermore, chapter 6 provides an evaluation and overview of practical results of this dissertation. Test and evaluation of the matching method as well as recommendation approach are two different perspectives of testing phase. Those directions are discussed in details in sections 6.1 and 6.2.

chapter 7 consists of the summary and conclusion of the research results. It concludes clearly the contribution of this work to the science, generalization of the results as well as limitations as the future work. The chapter covers a short discussion on the topics that haven't been covered in this thesis. They can be base for future research projects, theses and scientific cooperations. The proposed concept in this thesis can also be used for performance evaluation and assessment of the market products and services based on preferences (namely product configurations). This is also discussed as a future work in this chapter.

Chapter 2

Background Information and Related Work

»If you only have a hammer, you tend to see every problem as a nail. «

– Abraham Maslow

In order to distinguish innovations of this research and reveal scientific challenges, different literature and funded research projects have been reviewed. Since this work is an interdisciplinary research on CM and big data analysis, the following chapter covers both areas. A literature review of this work uses defined steps and guideline for systematic literature review in the [Kitchenham and Charters, 2007]. Proving an importance of the focus of this research and emphasizing its major challenges are the main goals of this chapter. The Innovation of this research can be clearly identified through careful analysis of state-of-the-art literature. To this aim, a clear identification of needs for conducting a review is defined as a major step in planning the review [Kitchenham and Charters, 2007].

Kitchenham and Charters defined "conducting and reporting the review" as a next step in the systematic literature review. To this aim, high quality and key publications in both of stated areas have been chosen based on their relevance to the topic and originality. Studying relevant PhD theses identifies barriers of this research with related work. Additionally, selected relevant and high valued research projects have been reviewed based on their novelty, lifetime, relevance and overlaps with this research as well as source of funding. Most of studied research projects in this thesis are funded through European Union (EU), German Research Foundation[1] or Federal Ministry of Education and Research[2]. All and all, such an intensive review identifies lacks and highlights of this research.

2.1 Competence Management

First and foremost, the definition and meaning of "Competence" and "CM" should be very well understood. The term "Competence" has been defined first by White as a performance motivation [White, 1959]. He described competence from a psychological point of view as a capacity of organism to interact with the environment in an effective manner. According to his definition, the environment is

[1] Deutsche Forschungsgemeinschaft (DFG)
[2] Bundesministerium für Bildung und Forschung (BMBF)

nowadays workforce and Competence Assessment (CA) provides an efficiency level of workers. This was an introduction to thoughts for performance evaluation of HRs. There has been a lot of research work about competences and CM afterwards, especially in recent years. Ennis provided a comprehensive literature review about competence definitions and models. Competence models are also being used for succession planning specially with respect to the mobility of workforce and retirement [Ennis, 2008].

Competence is often defined as a sort of knowledge, expertise, skills and abilities that people need to carry out job roles. Each job role has its specific competence requirements. Lundberg referred to competence from executive planning and development perspective [Lundberg, 1972]. The "knowledge", "attitude" and "ability" are stated as building blocks of executive competences from his point of view. They can be defined respectively as thinking, feeling and doing in terms of activities. Lundberg referred to a weighted combination of these building blocks for learning goals of any program. A 3-dimension conceptual scheme proposed in [Lundberg, 1972] is one of the first competence models used in the literature.

Lundberg's model suppose top managers to concentrate on developing their "conceptual competences of task, technology, actor and structure change variables". Similarly, middle managers should focus on the human competences. Low level managers should acquire technical competences [Lundberg, 1972]. Later, McClelland addressed CA and "modern competency movement" in his highly cited publication. He has also reviewed competence from psychological perspective and suggested settings for CA. This was one of initial steps in CA activities. Motivation of McClelland for proposing competences assessment rather than intelligence was inefficiency of the popular American intelligence tests [McClelland, 1973].

According to McClelland's suggestion, the best CA should consist of the criterion sampling and reflect changes. It should also include competences associated with the life outcomes and operant and respondent behavior. McClelland claimed that sampling of job skills through an assessment predicts proficiency on the job. Identification of job specific competences, protocoling those competences and testing candidates based on job specific competences has been defined as competence sampling. As an example, for recruiting software developer, he should be assessed for whatever a software developer really should carry out in his carrier. A CA shall be either theoretical or practical job sampling. The sampling should be based on the careful behavior analysis [McClelland, 1973].

Similarly, Gilbert issued the ways to measure and assess human performance. In this context, an analysis of human behavior and its consequences in the context of value provides an insight about valuable performance. This will result a worth which is a function of the value and cost and can be measured by following equation ($Worth = Value/Cost$). Accordingly, the worth becomes greater when there would be more value with less costs. Hence, people who produce valuable results without costly behavior are more competent [Gilbert, 1978]. The term "competence" has been focused not only from a psychological perspective, but also from technological and practical points of views.

Furthermore, competence has been addressed as general competences [Dessler, 2015], soft skills [Robles, 2012; Bailey, 2014], business skills [Bailey, 2014] and technical competences [Bailey, 2014]. The EC defined competence as proven ability to use knowledge, skills and personal, social and/or methodological abilities, in work or study situations and in professional and personal development [Eurpean Commission, 2008]. Competences are important in recruitment, selection, evaluation, training, development and review of HRs as well as strategic and succession planning in enterprises. However, there is still a lack of clarity in the classification of differences between competence and competency. Some literature and reports intermix the definition of both.

According to the Cambridge dictionary[3], competence is "the ability to do something well" and competency is "an important skill that is needed to do a job". Consequently, the term competence reflects the performance perspective of required skills (e.g. competency) to do a specific job. Teodorescu has referred to the definition of the competence in [Gilbert, 1978] and highlighted the worthy performance aspect as key difference between "competence" and "competency" [Teodorescu, 2006]. Due to the fact that competence reflects a quality perspective of required CK, the main focus of this thesis is on competence rather than competency. In order to cover most of terminologies that have been defined for competence in the reviewed literature, this work defines and uses the term Career Knowledge (CK) which is further discussed in chapter 3.

In 1990, Prahalad and Hamel addressed a competence model in the corporation context. Their highly cited publication indicates core competences as wellspring of new business development. They compared revenues of two large telecommunication companies (NEC and GTE) in 1988. Prahalad and Hamel concluded that NEC is one of top 5 companies in revenue because it "conceived of itself in terms of core competencies". They defined core competence as "communication, involvement, and a deep commitment to working across organizational boundaries". Authors also defined the term core products and stressed a direct link between core competencies and core products which results qualitative end products [Prahalad and Hamel, 1990].

Sandberg focused on competence at work and has seen competence as specific set of knowledge and skills required to perform specific job (work). An interpretative approach, "phenomenography" proposed in this research has been applied in Volvo Car Corporation in Sweden. According to the most central findings in this research, a human competence "is not primarily a specific set of attributes. Instead, workers' knowledge, skills, and other attributes used in accomplishing work are preceded by and based upon their conceptions of work. More specifically, the findings suggest that the basic meaning structure of workers' conceptions of their work constitutes human competence." [Sandberg, 2000].

A multi-dimensional and holistic typology of competence has been argued in [Delamare Le Deist and Winterton, 2005]. Competence has been addressed as a

[3] `http://dictionary.cambridge.org/`, accessed: July 2015

Table 2.1: Summary and history of selected scholarly competence associated definitions in the literature

Reference	Definition	Comment
[White, 1959]	Competence is a capacity of an organism to interact with the environment in an effective manner	The term "effective manner" reflects the job performance perspective and can be measured using assessment methods.
[McClelland, 1973]	Competence assessment should consists of the criterion sampling, reflect changes, include competences associated with the life outcomes and operant and respondent behavior.	He clearly defined proper settings of efficient CA and highlighted involvement of changes and behavioral and real life measures in the assessment.
[Gilbert, 1978]	Competent employees are people who produce valuable results without using costly behavior.	He also focused more on CA and addressed the cost issue beside of the performance as a part of CA.
[Baladi, 1999]	CM is about specification of competence needs, analysis of present situation and future requirements in order to identify competence gaps, competence sourcing, provisioning and procurement.	He focused more on defining the CM rather than the competence itself. The CM processes are defined in his work.
[Lindgren et al., 2004]	CM is specific information systems that help organizations to manage competences in the organizational and individual levels.	CM has been focused from two different perspectives: (1) enterprise and (2) individual.
[Delamare Le Deist and Winterton, 2005]	Competence is a fuzzy concept that has different definitions and focuses based on different practices and cultures.	The competence area still needs to reach to a common definition.
[Eurpean Commission, 2008]	"Proven ability to use knowledge, skills and personal, social and/or methodological abilities, in work or study situations and in professional and personal development."	This is a very complete definition that reflects all dimensions and perspectives of competence and associated features and processes.
[Bailey, 2014]	Non-technical Knowledge, Skills and Abilities (KSAs) are more important in the successful technical world specially in the IT sector.	The recommendations in this paper could be the base for customizing of computer science studies for preparing more competent candidates for the industry needs.

fuzzy concept in this tentative work through different practices in some countries specially US, UK, Germany and France. An extension of competence analysis depth, an investigation of greater competence details in some occupations as well as identification of the rift between rationalist and interpretative approaches are stated as main challenges in [Delamare Le Deist and Winterton, 2005]. Developing any system that integrates those wide and disparate directions is addressed to benefit an efficient competence analytics for on-the-job-training.

Additionally, Delamare Le Deist and Winterton mentioned conceptual, operational, occupational and personal competences as main dimensions required for developing a general typology of competence. Cognitive competences like knowledge and understanding are conceptual competences in occupations. Functional, psycho-motor and applied skills provide relationship between operational competences in occupation. Meta-competence as learning to learn and social competences as behaviors and attitudes are associated accordingly with conceptual and operational competences in personal competences [Delamare Le Deist and Winterton, 2005].

Similarly, Eurpean Commission defined competence as a composite definition of cognitive, functional, personal and ethical competences. Consequently, competence

is an efficient use of

- theories and informal tacit knowledge acquired experimentally (e.g. cognitive competence),
- functional abilities required in a given area of work, learning or social activity (e.g. functional competence),
- know-how to manage special situations (e.g. personal competences) and
- ownership of specific personal and professional values (e.g. ethical competence).

This study suggests self-direction as a critical factor in defining the competence level of individuals. Competence assessment of individuals based on self-direction means his/her talent for integrating these stated competences in specific challenges, goals, situations and job roles [Eurpean Commission, 2008].

Ennis covered the needs to use competency models and also practical components of competency models. In addition to current uses of competence models, they are also being used for succession planning because of mobility of workforce and retirement. The need of core competences in all jobs and professions is addressed in this paper as Knowledge, Skills and Abilities (KSAs) [Ennis, 2008]. Meanwhile, Tissot selected and defined competence as one of 100 key terms of European educations and training policy. He defined competence similar to the aforementioned ones as ability to use learning outcomes in the practice with the specific context. Functional side of the competence beside its cognitive elements is highlighted in this publication [Tissot, 2008]. The definition of the competence from different perspectives in the literature has been summarized in Table 2.1.

As a summary of this section, the definition of Lundberg approves building blocks in setting up an executive competence model as thinking (knowledge), feeling (attitude) and doing (ability) [Lundberg, 1972]. This is in lines with the fundamental assumptions in setting up the hierarchy of CKR model which is further discussed in chapter 3 on page 51. As soon as the CK is modeled, they should be assessed and represented based on the domain specifications and needs. To this aim, McClelland's research outcomes provide basics and fundamental thoughts for domain specific CA [McClelland, 1973]. In this regard, weighting and defining the importance of competences in the CKR model specifies the domain specific needs and differs the structure of this model for various sectors.

According to the Gilbert's formula, the goal of an assessment should be to find talents who could deliver higher job performance with less costs [Gilbert, 1978]. This factor is important especially in classification of talents with the same (or similar) job performance. In this case, a careful cost analysis of delivering such performance is required. Despite a domain specification support, a general and massive structure of the CKR model depicts given pillars of the general competence definition published by the EC [Eurpean Commission, 2008].

Prahalad and Hamel concluded that improving core competences results in the

qualitative end products [Prahalad and Hamel, 1990]. In this regard, enterprises should efficiently use their HRs. This needs filling the CK gaps in enterprises with mapping acquired CK to required CK which is the main goal of this research as well. In conclusion from literature analysis about the definition of competence, the CKR model given in this research covers different perspectives and dimensions defined in the literature to competence. This facilitates generalization of the model in order to support a wide variety of sectors and case studies. As a next step, the process of how to measure, analysis and manage competences and their pros and cons should be extracted from the current stats-of-the-art.

2.1.1 Theory and Processes of Competence Management

Baladi has stressed the positive reactions and effects of CM in organizations from employees, managers and organizations perspectives. Consequently, employees should focus on their competence development and extend their know-how. Managers should support employees in this regard and be flexible, fast and more accurate in their job since they clearly understand who knows what. Finally, organizations should support a systematic competence development and strategic competence supply. As a result, he addressed CM as specification of competence requirements, identification of competence gaps and competence sourcing, development and staffing. These dimensions build-up competence identification, assessment, acquisition, and usage processes in the CMS [Baladi, 1999].

Similarly, Lindgren et al. defined CM as specific information system that helps organizations to manage competences in organizational and individual levels. They focused on the definition of competence from macro level as an organizational and micro level as an individual competence analysis. Following design principles have been outlined for CMSs in order to improve the quality of organizations' competence information: (1) enhancing "formal competence descriptions with informal ones", (2) granting users "control over their competence descriptions", (3) "transparency", (4) "real-time capture", (5) "interest integration", and (6) "flexible reporting" [Lindgren et al., 2004]. In fact, the employee and manager as well as organizational perspectives of Baladi are likewise addressed as micro and macro levels by Lindgren et al..

Notably, Baladi described processes and components of knowledge and CM initiative at Ericsson. A web-based CM application has been developed in Ericsson through this initiative in which supports individual and organizational CM [Baladi, 1999]. In fact, demonstration of the micro and macro levels by Lindgren et al. complies with Baladi's perspective, accordingly [Lindgren et al., 2004]. As shown above, Lindgren et al. suppose "analysis of future requirements", "analysis of present situation", "gap analysis" and "sourcing of competences" as major CM processes [Lindgren et al., 2004]. A CMS in an organizational level requires intensive interaction with knowledge management and in fact enterprises should be aware of "who knows what?".

According to Draganidis and Mentzas's survey analysis, competence identifica-

tion, modeling, assessment, standardization and profiling are essential building blocks of any CMS. They have reviewed and examined different systems, mainly 22 CMS and 18 learning management products and services. Draganidis and Mentzas concluded that competences are important in workforce planning, recruitment management, learning management, performance management, career development and succession planning. Further contribution on standards such as Extensible Markup Language (XML), World Wide Web Consortium (W3C) and Resource Description Framework (RDF), semantic technologies, CM systems with self-service support are recommended as a research road-map in this publication [Draganidis and Mentzas, 2006].

Bailey studied the importance of non-technical Knowledge, Skill and Ability (KSA) in a successful technical world with the focus on the IT sector. Her motivation to this research is required one million IT workers in 2018 and 1.4 million IT job openings by 2022 in the US, as she described. These estimates indicate importance and novelty of research activities of this dissertation as well. She identified essential non-technical competences in IT sector in order to prepare responsive university curriculum in the next steps. In fact, she provides standardized competence model for IT sector and consequently recommend an effective VET plan. It is based on survey study of domain experts and an analysis of collected results [Bailey, 2014].

Through a survey study of collecting necessary information from different sources, Bailey identified 32 desirable non-technical, 12 business and 20 soft skills. Her conclusion is that many computer degrees have general curriculum in order to prepare candidates for a wide variety of IT jobs. At the same time, some IT companies hire candidates with less technical competences, but more competent in soft and business skills. Recommendations made in this literature could provide basic curriculum of computer science studies in universities in order to prepare competent and industry-oriented employees for the future. According to her conclusion, soft skills demands more non-technical skills than business concepts in IT sector [Bailey, 2014].

A case study of IT supported CM has been studied by Hustad et al. in Ericsson. Target system in this project consists of data collection and analysis, CM process, and global competence planning. The CM process covers competence analysis, planning and implementation. An organization's long-term and short-term strategic competence analysis identifies individual and organizational competence gaps. Consequently, competence analysis and gap identification results supports preparing an organizational and individual competence development plan. Based on this plan, further theoretical and practical action programs such as courses, project participations in different locations and education is being implemented [Hustad et al., 2004].

Through suggesting a workplace learning context model, Ley et al. covered a broad knowledge spectrum from organizational knowledge to individual's knowledge. According to [Ley et al., 2008], a learning workplace context should take care of at least (1) the work space, (2) the learning space, and (3) the knowledge

space. These spaces are unified as "3spaces" concept in their research. In knowledge worker's context, those a work space, learning space, and knowledge space reflects respectively process, competence, and domain dimensions. Therefore, all of those dimensions should be supported in work-integrated learning systems. As a result, developed system in the frame of this project provides task learning, task execution, and competence-gap based supports for organizations [Ley et al., 2008].

Rozewski and Malachowski developed Competence Object Library (COL). This library supports competence representation standards such as IEEE Reusable Competency Definitions (IEEE RCD) [IEEE, 2008] and HR-XML [Allen and Pilot, 2001]. The main goal of COL is facilitating development of tools for competence analysis in the HR and e-learning area. According to their opinion about the cost of competence gap resolution in enterprises, a proper identification of acquired and required competences contributes to a quantitative CA. Figure 2.1 shows the structure of the COL using Unified Modeling Language (UML) class diagram. The library discussed in this paper consists of competence structure and expansion modeling [Rozewski and Malachowski, 2009].

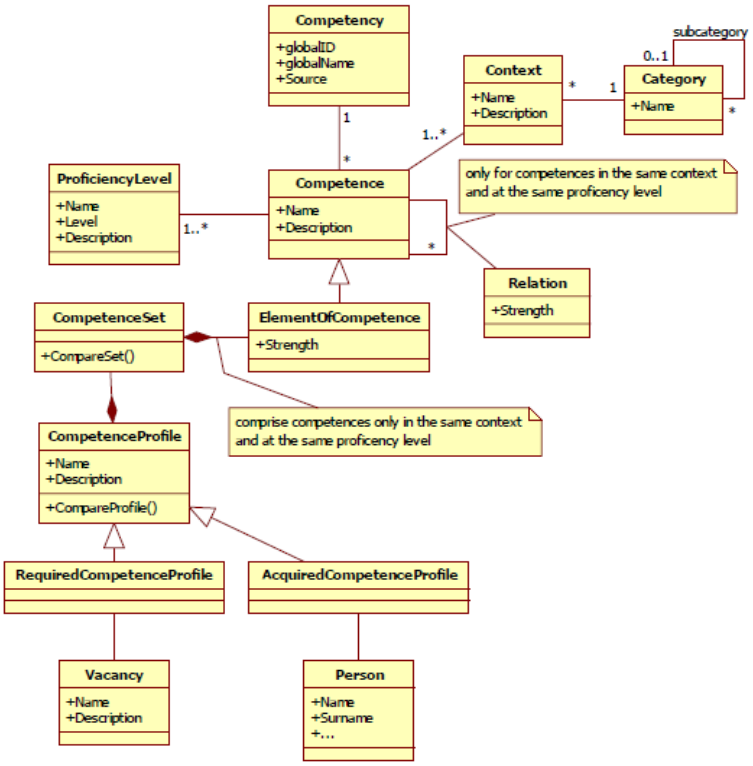


Figure 2.1: The UML Class Diagram for Competence Object Library [Rozewski and Malachowski, 2009]

In another work, Rozewski and Malachowski outlined the structure of CM processes as follows: (1) identification of market oriented competences, (2) adapting market's required competences to trainee's profiles, (3) adapting market's competences to didactic materials, and (4) supporting trainee's learning process [Rozewski and Malachowski, 2009]. For the CA, Shavelson used statistical modeling. He proposed six facets for competences. These facets identify the domain that the competence measurement should be developed. The base of Shavelson's assessment model is the task sampling [McClelland, 1973] discussed earlier [Shavelson, 2013]. In general, statistical analysis strengthens achieved results in assessments for producing empowered competence analytics.

In addition, Rozewski expanded his research by proposing following models to support CM in open distance learning: (1) knowledge processing, (2) motivation model, (3) curriculum development model, and (4) competence development model in intangible production network. The main focus of this work is on the competence acquisition and accordingly developing a curriculum based on competence gap identification. An ontology-based competence modeling is used in his research. The lack of computational methods in competence analysis is addressed as a motivation of the research [Rozewski, 2012]. The use of an ontology in this work restricts its extension and adoption to further sectors and case studies. This is in contradiction with the generalization of competence analytics.

Furthermore, Rozewski et al. compared the use of two different concepts of ontology and set theory for formalizing the CM. Their first and most important conclusion is that there is still a lack of efficient quantitative competence analysis methods. In addition, an ontology-based competence modeling is suggested to be used in defining a set of education offers based on competence requirements in the market. As a result and based on the scientific research on knowledge acquisition methods, they concluded that using the mathematics and competence set theory is more successful if a proper order is maintained. An aggregation of proficiency level, competence context and description is addressed as common understanding of the complex competence structure [Rozewski et al., 2011].

As a summary and analysis of the literature review in this section, Lindgren et al. issued transparency and granting talents to control their competence descriptions. Their issued design principles are equally important with settings of the self-assessment process in this dissertation [Lindgren et al., 2004]. This is further discussed in section 3.2.2. Enriching competence analytics through gap identification and providing further competence improvement and development recommendations, which is the main focus of this dissertation (see Industrial Challenge (IC) 5 (VET recommendations) on page 12 and Chapter 5 on page 95), is also addressed by Baladi. He stressed positive effects of competence analytics in improving organization performance without demonstration, conception or realization of practical research results [Baladi, 1999].

For the purpose of generalization and easy integration with a wide variety of systems and algorithms, a CK is represented using XML in this research. This was also recommended by Draganidis and Mentzas as open research issue in this

area. In addition, their suggested semantic technologies and Resource Description Framework (RDF) as scientific and technological challenges of the CM [Draganidis and Mentzas, 2006]. Both of these issues are supposed as the future work of this dissertation. Bailey's estimates about IT job openings in 2018 and 2022 in the US and accordingly in the world highlights the importance and novelty of this dissertation. It specially shows the growth of job seekers and accordingly their associated data and therefore the need to utilize big data in processing of such large scale case studies [Bailey, 2014].

The CKR model in this PhD covers identification of domain specific required competences in the same fashion as Bailey's model. Bailey's focus is only in the IT sector and tries to provide recommendations for IT studies and shorten the gap between IT curriculum and required competences in the industry. Her work and its achievements can be compared with achievements through applying the CKR model in the case study of this dissertation, which is an academic career in computer science. The work of Ley et al. is based on "competence performance matrix" in order to identify relationship between competence and performance. In this regard, they analyze all required competences assigned to a task in the stated matrix [Ley et al., 2008]. Their concept lack the scalability and generalization.

In addition to Ley et al.'s work, this PhD work provides mas customization, generality, scalability and applied statistical analysis for identification of correlation between competences and employees. It also complements the general structure of CM processes outlined in [Rozewski and Malachowski, 2009]. In this regard, special attention is given to the adoption of the market's required competences to trainee's profiles. This approach is addressed in the frame of RQ 1 (Skill mismatch) in this dissertation. Moreover, it affects the learning process of trainees, especially in the VET, which is discussed in chapter 5 on page 95. Rozewski and Malachowski's work lacks scalable and adaptive algorithms to support large volumes of the HR data. But in the other hand, applies efficiently mathematical modeling of competences in trainee's learning.

Different from [Shavelson, 2013], this dissertation uses statistical analysis to analyze competences. Therefore, the use of statistical analysis here is to discover correlations between competences, rather than for handling a CA. In fact results achieved through the CA are inputs of the statistical analysis and modeling. A mathematical analysis as scientific engine of the CM in this work emphasizes it from other works like [Rozewski, 2012]. In this regard, the concept could be easily adopted and generalized to other sectors without any need for further technical and professional IT-specific efforts. From another side, the current PhD work agrees and considers the same motivation of lacking computational methods in the competence analysis as suggested in the [Rozewski, 2012].

In Amiri et al.'s work, (1) the use of AHP is proper decision making based on the independent criteria. Meanwhile, the concept lacks the consideration of interdependence criteria. (2) The traits of the individuals to be considered cannot affect the decision criteria. This means, if all firms have an equal competence level at one criteria, the importance of that criteria should be reduced. In addition,

Amiri et al. focused on competence evaluation of firms rather than HRs [Amiri et al., 2009].

2.1.2 Applied CM and Funded Research Projects

In order to determine an innovation and contribution of this research, related funded research projects and their pros and cons have been clearly studied. Most of them are EU, BMBF and DFG funded projects. Table 2.2 is a summary of general information for those selected projects. The most of information about projects have been summarized from their official websites. By studying those projects and also indicating their highlights and lacks, the importance and contribution of this research can be better identified. In almost all projects, at least one Small and Medium-sized Enterprise (SME) is involved to exploit project results in a real world applications. The summary of selected projects is given in the following.

The "Confidence Competence Management as a System for Balancing Flexibility and Stability Needs (CCM2)" project (P1) is a German research project. It uses an integrated trust and CM to balance the flexibility and stability of companies for change and innovation potentials. The project first made a qualitative competence analysis of 503 employees and considers the results of this analysis as a reference competence level of specific sector. It compares the competences of assessees with this reference competence level of the sector and provides an insight about competence lacks and highlights of an assessee. A web based toolbox consisting standardized competence questionnaires in order to deposit influencing factors that match to the problems level is developed in the project [Sprafke and Wilkens, 2015].

Five competence dimensions of the (1) cooperation, (2) self-reflection, (3) combination, (4) coping with complexity and (5) self efficacy are the main competence pillars in this project (P1). The arithmetic mean measurement for each dimension is the mathematical background of the project [Sprafke and Wilkens, 2015]. The project lacks the weighting of conferences and giving an importance for any of competences. In addition, it should be stressed that having a hierarchy of competences facilitates easy decision making about specific candidate in different job positions through adjusting the weights. In this regard, this dissertation provides more flexibility and generalization of the competence models.

The "Dynamic Interdependency of Product and Service in Production Area" project (P2) focuses on the industrial product service sector. The personnel competences in heterogeneous work systems is stated as an important enabler to social actors for performing successfully. Therefore, the project is based on the specification, measurement and development of personnel competences. The project covers continuous competence development and integration of the results in everyday career life. This project supports individual (micro) and organizational (macro) competence analysis. A game-based community approach in this project prepares individuals for demands of work environment [Süße and Wilkens, 2014].

The "Business Simulation Game for HLB (Hybrid service bundles) specific skills development" project (P3) aims in developing a prototype for business game. The project focuses in the hybrid power sector and tries to optimize VET with providing an insight about later required competences. This project depends on the simulated virtual environment (game) on the model-based abstraction level and collects domain specific (HLB) business model and characteristics that could support successful management in this field. An optimization of HLB work processes, providing a domain specific knowledge through involvement of industrial partner and setting up different roles (job definitions) in the HLB sector are methods used in the project [Süsse, 2013].

Facing demographic challenges that may arise from a strategic deficit is a key point for "competence-oriented corporate coaching for sustainable CM in SMEs (4C4Learn)" project (P4). This project supports SMEs in developing occupational competence models and use of it for intra- and inter- company technical and demographic challenges. An on-line and hybrid product-service platform of the project motivates SMEs to create a CM database consisting industry-specific success factors of the CM. Transfer of lessons learned and corporate coaching is a key issue in this project. The project may work well for collaborative companies and corporate groups. This project contributes very well to identification of collective competence gaps.

The "Modeling and Measuring Competencies in Higher Education (KoKoHs)" is a funding initiative (P5) that consists of 24 research projects in different fields. These projects cover a wide competence area in the higher education. All sub-projects in the frame of this very big initiative focus on assessment and modeling of teaching competences (for teachers). Similar to earlier studied projects, this one also lacks providing a general competence model that could cover and specialized in a wide variety of sectors. As an example, research questions for one of those projects are defined as follows: "What empirical evidence can be found to support a theoretically predicted model of competencies? How do competencies in the field of scientific inquiry develop during the phase of academic science teacher education?" [Blömeke and Zlatkin-Troitschanskaia, 2013].

Furthermore, a "Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes" is a type of priority program (P6) funded through DFG. This program consists of overall 30 projects. Participating experts are with a cognitive orientation in specific disciplines. They cover a wide variety of ranges provided further research results on how to test and train competences in their respected field. A competence-based measurement method and its use in different pedagogical contexts of decision-making is a key issue in this initiative. Most of funded projects in the frame of this initiative use the pencil-paper and pedagogical concept in order to assess the competences of individuals. Similar to the "P5", this project also contributes to the development of competences in education and teaching career.

The main focus of "Pedagogical Knowledge and the Acquisition of Professional Competences in Training for Teachers" project (P7) is on improving a training

of teachers. A basic hypothesis is "Education Scientific contents and contexts represent a conceptual framework, need the teachers to interpret classroom and school events appropriate to reflect and thus to be used for managing occupational requirements." The project aims at: (1) systematizing educational scientific content in teacher training in theory and (2) identifying the importance of the educational scientific knowledge for the successful management of professional tasks. A test-based professional competence assessment method in this research can be easily extended as standard assessment method in other projects in different areas.

The "Measuring experimental Competences in the large scale assessments" project (P8) results in the computer-based test method. It allows valid and reliable measurement of experimental competence, comprehensively. Development of tests is based on the analysis of curriculum and interviews with domain experts and is tested by several discrete but evolutionary studies on its validity. The test procedure is distinguished from paper based tests, in particular through the integration of action-oriented simulated experimental environments. The method has been successfully used in a cross-sectional study with 1,165 pupils. Outcomes of this project provide very well structure in competence assessment and applied questionnaire systems.

"Technology-based Assessment of Skills and Competences in VET (ASCOT)" is also an initiative (P9) that consists of 21 projects. This initiative is formed as six main areas and needs close cooperation between research institutions, VET practitioners and facilities. Competence modeling and detection through simulated work environments is the main hypothesis in this initiative. This initiative focuses on automotive mechatronics, electronics, technician for automation technology, industrial clerk, househusband, care for the elderly, medical assistant. The target group is young people (youth) who are ending their VET. The initiative aims at the developing professional competence assessments by means of technology based methods.

The last studied project as related work is in production and logistics area. This project (P10) is entitled "Assistance System for Demographics Sensitive Company-Specific Competence Management for Production and Logistics Systems of the Future (ABEKO)". It aims at developing an assistance system for modeling of business processes and their competence requirements. It provides a catalog-based process competence structure model for technical, methodological and social competences. The project final result (developed assistance system) is based on the competence gap identification. The assistance system uses demography sensitive training and learning concepts as a basis for the design of site-specific programs for individual competence development.

As a summary of studied projects in this section, it is quite clear that there have been a lot of research and practical projects with innovative ideas in the CM recently. These projects range from scientific, theoretical and strategic to practical and industrial focus in the CM area. Some of them like P2, P3, P5, P7 and P10 concentrate on applied CM in specific field such as production, mechanical

engineering, higher education and logistics. Such projects provide a model or system which works only in an associated sector without further application in other case studies or sectors. The main scientific part of these projects depend on a domain specific competence model associated with ontology mapping. They lack mathematical representation and modeling of competences and integration with standards such as IEEE RCD, XML or HR-XML.

Some others focus on theoretical aspects of the CM and provide strategic recommendations and reports rather than practical results. These projects collect views of HRM experts by well designed surveys and provide an analysis of survey results indicating the HRM challenges and future road-maps. Most findings of such projects provide fundamentals of the practical and applied research in this area. Few of them also concentrate on the paper work in order to develop a competence model in specific areas. The outcomes and findings of such projects shall be extended in other areas, but more significant would be to extend the contextual studies in these projects and provide general outcomes in which cover different areas. Some more projects focus on the CA methods and study a paper-pencil assessment. They provide recommendations to improve paper-pencil assessments and lack use of novel technological assessment methods.

The project "P1" collects real world data for its competence model. This is comparable with the CKR model in this dissertation, but it doesn't provide further competence development recommendations. Furthermore, it doesn't focus on the competence gaps and lacks integration with well known standards in the CM area. As a results, outputs of this web-based tool can not be exported to other softwares and environments. Moreover, the project lacks weighting of competences since all dimensions are considered at the same level of importance. It covers a limited number of questions which can not be extended or customized by the user. In general, the generalization, use of modern and state-of-the-art technologies and integration with similar systems are open issues in connection to this project.

This project provides an interesting conclusion which is in line with the hypothesis of this dissertation. Through verification of the practical results, it concludes that clustering of candidates facilitates an employee selection process. It supposes clustering of assessees into less competent, mid-competent and high-competent. Both "P2" and "P3" projects focus on the specific field and improve VET in their focused areas through providing insights about required competences. As an example, the "P3" uses business gamification in order to indirectly identify competence gaps in the firms. This goal is defined as identification of required CK in this PhD work which contributes to the matching of required CK to acquired CK in order to identify CK gaps.

On the other hand, "P3" focuses on hybrid service bundling and lacks generalization to a wide variety of sectors. The project "P4" focuses more on the collaborative competences of firms rather than individuals. Therefore, this issue differentiates it from the goal of this dissertation. Most of projects in the frame of KoKoHs initiative (P5) use a survey methodology to collect required competences about specific field. They use tests to understand the current competence level of

students while teaching them. Accordingly, they lack a practical use of competence analytics and applied competence development based on achieved results. Provided outcomes of such projects may significantly provide findings specially about domain specific competences.

Similarly, the DFG funded "P6" priority program concentrates on developing and assessment of competences for teachers and education. The outcomes of projects in the frame of this research can be extended and further integrated as best practices in other areas. But, they also lack an easy and fast adoption to other sectors and also generalization aspect. Those results are carefully reviewed through designing the CKR model in this research. Due to the fact that the main focus and contribution of this dissertation is not the design of self-assessment in the CA, the outcomes of "P8" project have been used in this regard.

Table 2.2: A summary of the funded research projects in the field of CM

	Project Title	Coordinator	Funding Source	Portal	Comments/Remarks
P1	Confidence Competence Management as a System for Balancing Flexibility and Stability Needs (CCM2)	Ruhr-Universität Bochum	BMBF (2009-1013)	http://www.kompetenzmanagement.rub.de/	
P2	Dynamic Interdependency of Product and Service in Production Area	Ruhr-Universität Bochum	SFB 29: C5 DFG (2010-2014)	http://www.iaw.ruhr-uni-bochum.de/aup/forschung/projekte/dfg0710-0614.html.en	
P3	Business Simulation Game for HLB (Hybrid service bundles) specific skills development	Ruhr-Universität Bochum	SFB 29: T5 DFG (2013-2015)	http://www.lps.ruhr-uni-bochum.de/tr29/projektbereiche/transferprojekte/T5/	
P4	4C4Learn - Competence-oriented corporate coaching for sustainable CM in SMEs	Ruhr-Universität Bochum	BMBF (2013-2017)	http://www.4c4learn.de/	
P5	Modeling and Measuring Competencies in Higher Education (KoKoHs)	Johannes Gutenberg University Mainz	BMBF (2011-2015)	http://www.kompetenzen-im-hochschulsektor.de/index_ENG.php	
P6	Competence Models for Assessing Individual Learning Outcomes and Evaluating Educational Processes	German Institute for International Educational Research	SPP: DFG (2007-2013)	http://kompetenzmodelle.dipf.de/en	
P7	Pedagogical Knowledge and the Acquisition of Professional Competences in Training for Teachers	Goethe-Universität Frankfurt am Main	BMBF (since 2009- now)	http://www.bilwiss.uni-frankfurt.de	
P8	Measuring experimental Competences in the large scale assessments	Universität Duisburg-Essen	BMBF (2012-2015)	http://didaktik.physik.uni-essen.de/mek-lsa/	
P9	Technology-based Assessment of Skills and Competences in VET	German Aerospace Center (DLR)	BMBF (2011-2014)	http://www.ascot-vet.net/	
P10	ABEKO: Assistance System for Demographics Sensitive Company-Specific Competence Management for Production and Logistics Systems of the Future	TU Dortmund	BMBF (2013-2017)	http://www.abeko.lfo.tu-dortmund.de/	

2.2 Scalable Data Analytics (Big Data)

In 2011, Gartner Market Research [Fenn and LeHong, 2011] added the term "Big Data" and "Extreme Information Processing and Management" for the first time to an annually published hype cycle for emerging technologies (see Figure 2.2). In 2015, Gartner added technologies connected to the big data such as Internet of Things (IoT), advanced analytics, citizen data science to its hype cycle for emerging technologies[4]. In addition to traditional Relational Database Management System (RDBMS), so-called NoSQL and NewSQL DBs have appeared as high performance alternatives providing data storage and analytics for semi-structured and unstructured data. These DBs can also be deployed to many nodes and allow adjustable redundancy levels as required by the application.

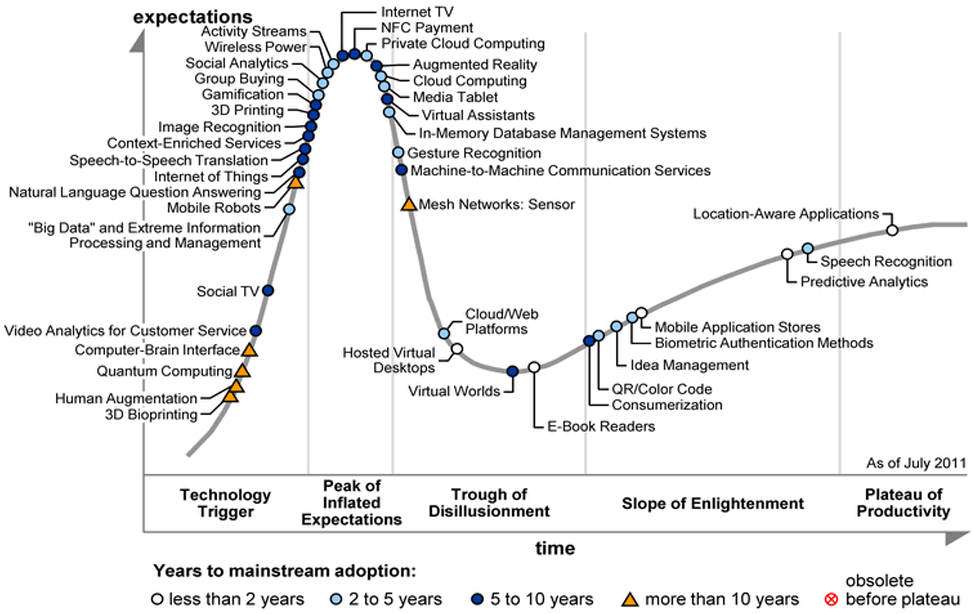


Figure 2.2: The Gartner Hype Cycle for Emerging Technologies in 2011 [Fenn and LeHong, 2011]

Furthermore, distributed file systems usually provide data redundancy by means of data replication, access transparency (clients are not aware of the way data is distributed), failure transparency (the operation of the clients is not affected by failure of the back-end nodes), concurrency (all clients see the same state of the file system, eventually make concurrent modifications). Loukides addressed

[4] Janessa Rivera, Rob van der Meulen, Gartner's 2015 Hype Cycle for Emerging Technologies Identifies the Computing Innovations That Organizations Should Monitor, Published 18 August 2015, visited on February 2016 via `http://www.gartner.com/newsroom/id/3114217`.

big data from dimension of volume and defined it as "big data is when the size of the data itself becomes a part of the problem" [Loukides, 2011]. Dumbill provided a similar definition to big data as "data that becomes large enough that it cannot be processed using conventional methods" [Dumbill, 2011]. Other authors define big data by three dimensions of volume, velocity and variety [Russom, 2011].

There are numerous domain examples for big data applications and volumes since few years ago, including web applications, recommender systems for on-line advertising, people analytics in HR, financial decision making, medical diagnostics [Bohlouli et al., 2010], or the operation of social networks or large IT infrastructures. For instance, Al was processing 20 petabytes (10^{15} bytes) per day in 2008 [Dean and Ghemawat, 2008]. In 2011, Google was able to sort one petabyte of 100-byte-strings in 33 minutes on an 8000 machine cluster[5]. Amazon reported peak sales of 398 sold items per second on July 15, 2015[6]. Likewise, eQuest reported in 2013 the importance of using big data for processing of its over 1,5 billion job board records in its Database (DB) and processing of 5 million job postings per week [eQuest Big Data for Human Resources, 2010].

2.2.1 Architectures Providing Scalability

In computing, scalability is being referred from various perspectives such as database, processing, software and hardware. Bondi defined scalability as "the ability of a system to accommodate an increasing number of elements or objects, to process growing volumes of work gracefully, and/or to be susceptible to enlargement" [Bondi, 2000]. Scalability of data storage (e.g. Hadoop Distributed File System (HDFS)) and data processing (e.g. MR) are main issues covered in this work. Scalability can be horizontal or vertical. The horizontal scaling (scale out/in) considers adding new nodes to a distributed system. The vertical scaling (scale up/down) focuses on adding resources such as CPU or memory to specific nodes in a distributed system.

In this regard, scaling up/down refers to adding more resources to the currently running ones in the cluster. This provides higher virtualization advantages, since Virtual Machine (VMs) running on the specific machine could have more resources. Such vertical scalability is less expensive than completely adding new computer to a distributed system. Because there is no need to install new software, application support as well as Virtual Machine (VM). In contrast, scaling out/in targets adding more computers to the currently running high performance computer. This has its difficulties and complexity in the management of more machines in the system. IN addition, horizontal scalability may need complex programming languages [El-Rewini and Abd-El-Barr, 2005].

When big data is discussed nowadays, the most prominent data processing

[5]Source: Google Research Blog: Sorting Petabytes with MR - The Next Episode, accessed via `http://googleresearch.blogspot.de/2011/09/sorting-petabytes-with-mapreduce-next.html` in July 2015.

[6]Source: Amazon Press Release, accessed via `www.amazon.com/pr` in August 2015.

paradigm is MR. It provides a powerful functional abstraction for the execution of parallel batch jobs. MR is first stated by Google and then its open source version called Hadoop is released. It is two phases process: (1) map and (2) reduce. Data items are being distributed between computing nodes (e.g. "mappers") and are being processed independently. These data items are paired as (key, value). Afterwards, the processing results are being collected and summarized by reducers through key-values. In fact reducers collect and integrate processes having the same key [Dean and Ghemawat, 2008]. In this way, MR provides very functional scalability for processing large volumes of the data. Figure 2.3 shows MR architecture for famous wordcount example.

After publication of the MR and Google File System [Dean and Ghemawat, 2004; Ghemawat and Gobioff, 2003], the Hadoop implementation has been initiated by Yahoo and continued as an open source Apache project [White, 2009]. Its main components are the MR implementation and the HDFS. One of its key features is fault tolerance and the ability to run on clusters of unreliable commodity hardware. The main assumption in Hadoop is horizontal scaling rather than vertical scaling. The storage part of Hadoop implementation is HDFS and the processing part is MR. The Hadoop ecosystem consists of a wide variety of tools and technologies. Hadoop can be run in any type of distributed platforms.

In order to simplify the use of Hadoop, several higher level languages and interfaces have been developed. Hive is a data warehouse based on Hadoop. It allows interacting with Hadoop using an SQL-like query language and is most suitable in a batch processing context. Hive provides storage scalability and extensibility with user defined functions and aggregations. Another extension of Hadoop is the data analysis platform Pig. It provides a domain-specific language for defining data analysis processes and a compiler to translate these programs into MR jobs. The main benefits of the Pig are the ease of programming parallel data analysis, jobs and the abstraction of the specified tasks from the actual way of execution, providing the possibility of automatic optimization.

2.2.2 Scalable Database Technologies

It should be stressed that all RDBMSs, SQL, NoSQL and NewSQL systems have their place in the complex data processing arena (see Figure 2.4). There are no magical solutions for generalizing the data input, storage, querying, and processing to fit all problems. In recent years, NoSQL DB management systems have appeared as alternatives to RDBMSs. They have come to fruition in the last 5 years, although segments of it (specific products) are still bleeding edge and new NoSQL and NewSQL DB models are being invented almost monthly. They are characterized by not using SQL as a query language - or, at least, not using "fully-functional" structured queries. Firstly, it is imperative to clearly define the strengths and weaknesses of NoSQL technology, where DBs are not relational and

[7]Source of an image: The Chair of Computer Science, RWTH Aachen, accessed in February 2016 via `http://dme.rwth-aachen.de/en/research/projects/mapreduce`.

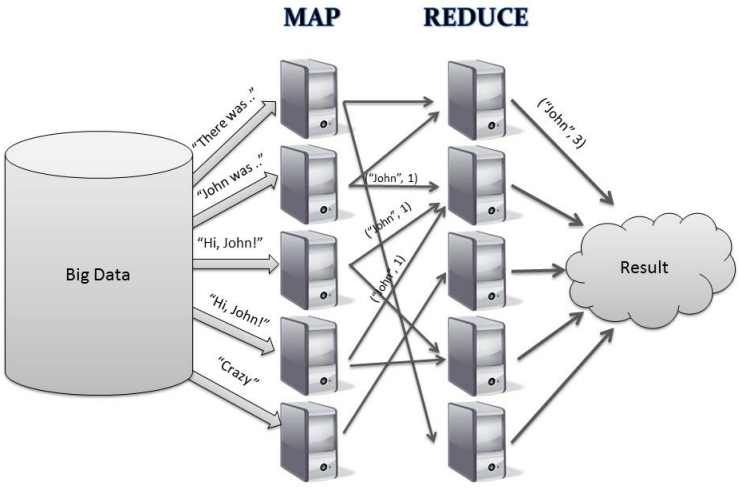


Figure 2.3: MR Architecture[7]

have no fixed data schema, complex relations or joins.

Mostly, NoSQL DBs do not offer relational operators like JOIN. They do not provide full Atomicity, Consistency, Isolation, Durability (ACID) guarantees in terms of atomic transactions, consistent DB states, transactional isolation, and durability of persistent data. The common denominator of the majority - if not all - NoSQL databases, is that they are optimized for large or massive data-store scaling. For instance, they are supposed to scale more efficiently and smoothly than RDBMSs by spreading the processing and storage load over a multitude of affordable server systems. On the other hand, RDBMSs - SQL DBs - scale up by using ever faster and memory/disk rich high-end server hardware. In contrast, NoSQL DBs offer good performance and horizontal scaling across the nodes of a cluster.

As such, they are well suited for web-scale applications and other big data domains, where the efficient storage and access of huge data volumes is more important than transactional consistency. For big data manipulation and processing, the NoSQL DBs are the best approach. Since, they add affordable horizontal scalability of storage spreading over nodes, over clusters and eventually over data-centers to vertical scalability and enable large data-throughputs - especially write-to-storage. Such a system should consciously sacrifice the RDBMS capabilities of orthogonalized data schemes consisting of tables and complex relationships (like JOINS) and a powerful query language (SQL).

[8]Updated Data Platform Landscape Map - February 2014, the 451 research group, accessed on January 2016 via `https://blogs.the451group.com/information_management/2014/03/18/updated-data-platforms-landscape-map-february-2014/`

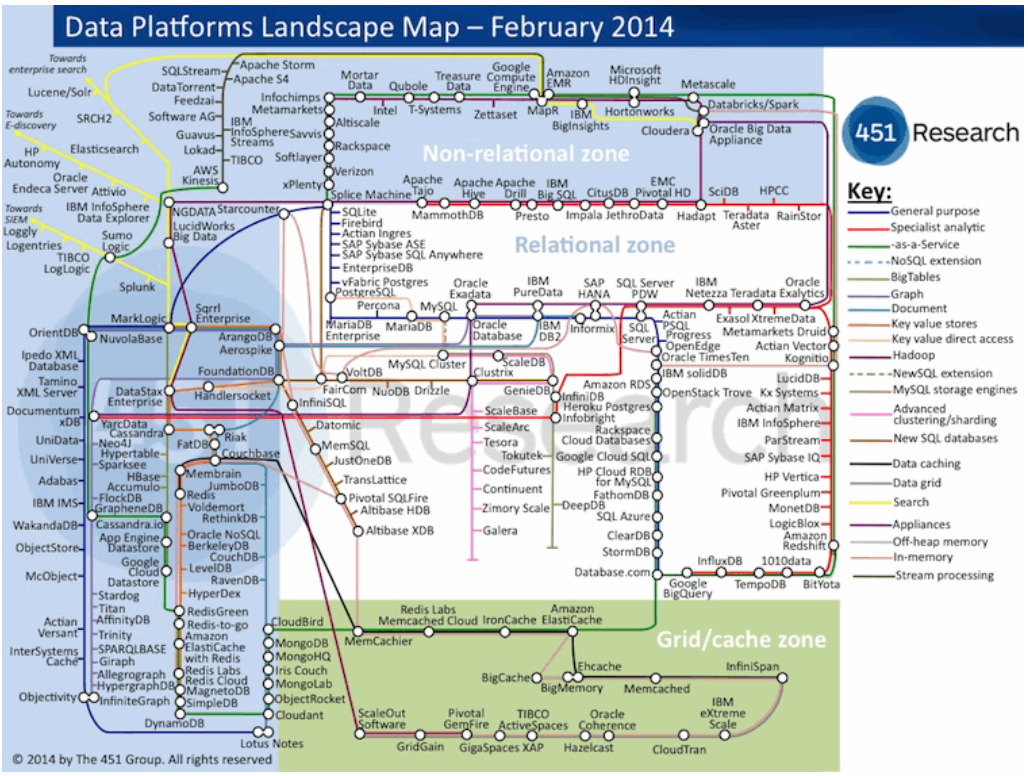


Figure 2.4: Data Platforms landscape Map[8]

At the same time, it is to be stressed that the fundamental differences between today's leading NoSQL solutions are much greater than the differences between different "strains" or products of RDBMSs. The NoSQL landscape is filled with disparate and - sometimes - diverging solutions of optimization for big data handling that can be complementary only if a unified platform with a common systems' API is implemented (see figure 2.4). NoSQL DBs scale in very different ways, having greatly differing data models and specific mechanisms for data querying. The latter are - on the main part - much more primitive than SQL although attempts are being made recently to bring more structure to querying in certain NoSQL DBs - for example by developing SQL-like interfaces, such as Pig, Hive and UnQL on top of the MR mechanism.

Furthermore, there are also significant differences in the type of scaling NoSQL products support. Some of them enable good scaling of the data-set size, some grow well in the volume of concurrent transactions. Some others excel only in the brute speed of data storage read or write, while others have a hybrid set of the before mentioned scalability capabilities, but with significant compromises stemming from this. In addition, the implementation, integration and programming of some of the NoSQL DBs is much more challenging that the incorporation of relational DB technologies in applications and middleware. This is due to the young age and documentation scarcity of some of the NoSQL products.

Consequently, the danger of using the wrong NoSQL tool for a specific large data-set processing problem is thus much more pronounced than choosing the "wrong" RDBMS for classic relational processing. Another fact is that not all NoSQL DBs are good at (horizontal) distribution over nodes and not all NoSQL DBs support effective replication (especially master-to-master) between server clusters. Usually, good scalability paired with excellent node-distribution means the underlying data model is primitive. A good case in point are graph DBs which are very single-node scalable and transaction-throughput efficient but are not optimized for efficient horizontal distribution of processing.

Due to provided reasons, it is imperative to study different types of the most common NoSQL databases, their pros and cons, and decide carefully on the selection of the most efficient and fitting technology with respect to the application specific requirements. The rest of this section provides a short summary about different NoSQL databases. With respect to this summary, it is easier to select the most proper technology for the aim of this research. There are four types of NoSQL databases: (1) key-value store (e.g. MEMBASE, Riak or Redis), (2) document store DBs (e.g. Apache CouchDB or MongoDB), (3) wide column store (e.g. HBase or Cassandra), and (4) graph databases, (for example InfoGrid, Neo4J or Infinite Graph), which implements flexible graph data models [Bohlouli et al., 2013b].

Key-Value Stores

The key-value stores collect only keys and values. Values are paired with the keys and are independent of each other. A key is an arbitrary string which is unique in the DB, the value can be any type of the data such as document, file, an image. Therefore, each data record in the DB could have different structure. Document stores and some graph DBs are also classified as key-value DBs. Key-value stores can be on the in-memory or on-disk DBs. These type of DBs could scale out by storing the data in multiple machines as well as replication. This DB is simple, flexible, portable, and without query language. Key-value stores are being used normally for user profiling such as customer, product recommendations and session management (caching). Redis, Memcached, Amazon DynamoDB, Riak KV, and Hazelcast are some types of key-value stores [Seeger, 2009].

Redis is one of the most common DB technologies, ranked by DB-Engines[9] in March 2016 as 3rd most common NoSQL DB between 299 systems in total. Redis is a type of in-memory key-value store. In fact, Redis keeps the whole dataset in the memory. Redis is extremely fast and supports over 100K *SETs* and 80K *GETs* per second. It support almost all data types such as lists, sets, and hashes. Redis operations are atomic, which provides updated values for concurrent accesses of multiple machines. It represents very fast read and write speed. Redis supports hierarchical single root replication tree which is known as master-slave replication as well. As an example, Instagram uses Redis in order to support mapping of

over 300 million photos back to the user ID [Carlson, 2013].

Document Stores

Document-oriented DBs or document stores are types of key-value DBs that extract meta-data from documents. A common examples of documents in these DBs are XML, JSON or Binary JSON (MongoDB). In document store DBs, the record can have disparate structures, which means that columns of the records can be different. Each column can have multiple values (e.g. arrays) and having nested records is also possible in these types of DBs. These types of the DBs are well suited for content oriented applications such as *Facebook*. In the document stores, a secondary index can be defined within the value content. MongoDB, CouchDB, Couchbase, Amazon DynamoDB and Marklogic are common examples of the document store DBs [Ippolito, 2009].

MongoDB is an easy to use, open source and distributed DB. According to the DB-Engines, MongoDB is the most popular NoSQL DB in March 2016. Comparing to the traditional RDBMS, MongoDB is document oriented and uses documents and collections to store the data. The elements of data are documents and can be considered similar to records in RDBMS and are JSON-like (BSON) objects. Collections are similar to tables. Records in RDBMS have same number of fields, but collections in MongoDB can consist of different number and type of fields. MongoDB has some of SQL-like properties such as queries and indexes. Queries are based on JavaScript expressions. Domains with dynamic query requirements as well as index definitions and higher performance demands for big DBs are well suited to use MongoDB [Banker, 2012].

CouchDB-BigCouch is similar to MongoDB as a Document-oriented (JSON) DB. It uses HTTP/REST as an interface to database. The big advantage of CouchDB in comparison to MongoDB is supporting Multi-Version Concurrency Control for the applications which need an access to the state of the data in different times. In addition, it differs in querying and scaling. It does not have good horizontal scaling method and users need solutions such as BigCouch for splitting and scaling issues. CouchDB uses a clever index building scheme to generate indexes and query expressions. It supports MR operations and on-line/off-line replication/sync capabilities. It is a good candidate for mobile embedded DBs on phones [Warden, 2011].

Wide Column Store

The Google's Big Table is the origin of wide column stores or column-based stores. Each storage block consists data from one column. The structure of wide column stores is similar to RDBMSs with the difference that the names

[9] The DB-Engines Ranking as a Knowledge Base of Relational and NoSQL Database Management Systems, Accessed in March 2016 via `http://db-engines.com/en/ranking`.

and format of columns can differ for each row in the table. They are, in fact, two dimensional key-value stores, since each value in the collection can consist of other key-value pairs. Wide column DBs are suitable for queries over large datasets. They provide fast search, access and data aggregation. The data in column-based stores can be partitioned across many hosts. As a result, this type of DBs provide better horizontal scaling and high availability and provide extremely faster query performance in comparison to the RDBMS. For instance, HBase, Cassandra, Amazon SimpleDB are released as wide column stores [Abadi, 2007].

Apache Cassandra was initially a Facebook internal project and is a distributed key-value DBMS. Its data model is similar to Google's BigTable and is a hybrid solution between Dynamo and BigTable. A table is a distributed multidimensional map indexed by a key. Querying in Cassandra is possible by columns and range of keys. There are tunable trade-offs for distribution and replication. Users can define the number of dedicated nodes for reads and writes before performing operations. RandomPartitioner (RP) and OrderPreservingPartitioner (OPP) are available for clustering and arranging the keys in Cassandra. RP distributes randomly key-values over the network and OPP uses the normal partitioning methods [Neeraj, 2015]. Due to replication of data through multiple nodes there is no downtime [Warden, 2011].

HBase is an open source and integrated DB with Hadoop/HDFS. It is distributed, scalable and easy to access through MR. The main focus of HBase is to support big tables with hundreds of billions of rows and columns based on Google's BigTable. It has a slow latency of individual transactions due to network traffics in distributed environments. HBase supports structured data sources. This DB type is suitable for applications such as finding small amount of data (such as top most competent employees) between over millions of the records. It provides faster read and write operations in large scale datasets. HBase provides better consistency and partition tolerance. HBase datasets can be accessed through Java API as well as REST [Warden, 2011].

As a summary and conclusion of stated technologies, MogoDB shows higher performance for applications with very high update rates, the use of memory mapped files for data storage, update-in-place (instead of multi-version concurrency control), client driver per language (not REST) and coded in C++. For many cases, MongoDB is suggested versus CouchDB due to its higher performance except the applications with versioning requirements. Writes in Cassandra are much faster than reads. Therefore it is suitable for the applications with writes more than read (logging), so one natural niche is real time data analysis. Banking and financial applications are example scenarios for using Cassandra.

In general, HBase provides higher performance when it is being accessed through many distributed clients. In addition, it is well suited for real-time data analytics and is good for applications creating recommendations, applications with ad hoc queries with aggregation through large similar datasets. Accordingly, HBase is also suitable for the aim of this PhD work, since it provides better performance when a small piece of data is being searched and processed across

large volumes of the data. An explicit example of this case could be for instance searching for the most competence employee or the best fitting employee to specific job position between hundred thousands of candidates.

2.2.3 Scalability and Decision Support Systems

Decision Support Systems (DSSs) are nowadays ubiquitous in industrial and research applications, and a large variety of commercial and open source tools and libraries exist. Furthermore, there is a rich theoretical background from various disciplines such as statistics and operations research that lays a solid foundation for decision making systems. The use of statistical and data mining methods has been limited to specific data from specific sources, depending on the application domain. There are notable open source tools like R, Weka, Mahout to develop a Decision Support System (DSS) and clustering algorithms. Similarly, commercial products like Hugin offer a large variety of methods ready to be used in various application domains such as HRM, TA and employee selection.

The proper application of tools and algorithms for decision support as well as clustering increases the productivity, efficiency, effectiveness, competitiveness and as consequence, making the planning, organization and investment more secure. DSSs are interactive computer-based information systems, which help decision makers utilize data, models, solvers, visualization and the user interface to solve structured, semi-structured or unstructured problems. They are built by different levels of system developers to support different levels of decision maker users [Dong and Srinivasan, 2013]. The decision process can be decomposed into three stages [Holtzman, 1989]:

- Formulation of the decision model that reflects the decision problem, i.e. generating alternatives and identifying evaluation criteria.
- Evaluation of the decision model, i.e. computing the implications of the decision model, evaluating it using a formal decision method and producing a recommendation.
- Appraisal of the recommendation, i.e. analyzing the recommendation and presenting the interpretation in a natural language form.

Amiri et al. defined two types of Multi-Criteria Decision Making problems: (1) Classical Multi-Criteria Decision Making (MCDM) problems which consist of crisp numbers for ratings and weights and (2) fuzzy MCDM that used impression, subjective and vagueness through linguistic terms such as "Not very clear", "Probably so", "very likely" for ratings and weights [Amiri et al., 2009]. Notable open source tools which are summarized in the following, include: (1) the R project for statistical analysis, (2) the WEKA project for data mining, (3) the KNIME platform for data analytics, and (5) the Apache Mahout for machine learning and decision support on top of the MR framework Hadoop.

The **R Project**[10] is an open source statistical language and in fact a comprehensive suite of tools providing to the users a vast variety of statistical and graphical techniques for data analysis. Furthermore, the R can be linked with other languages such as C and C++ and can be used for advanced massive data analysis. The R can be integrated and used on top of Hadoop for parallel and distributed statistical analysis. It supports for instance, time-series analysis, classification and clustering algorithms as well as linear and nonlinear modeling. There are numerous tools and packages in order to add a wide variety of functionalities such as data visualization (e.g. ggplot2), web application frameworks (e.g. Shiny) to the R language [R Development Core Team, 2011].

The **Predictive Model Markup Language (PMML)** is an XML-based language developed by the Data Mining Group. It provides ways to represent models related to predictive analytics and data mining. The PMML enables sharing of models between different applications which are otherwise incompatible. In this way, it could be implemented and integrated into Hadoop and MR as well. The primary advantage of PMML is that the knowledge discovered can be separated from the tool that was used to discover this knowledge. Therefore, it provides independence of the knowledge extraction from application, implementation platform and operating system. PMML consists of two clustering models as of center-based and distribution-based cluster models[11].

The **Weka workbench** is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. It is very flexible for users who can easily apply a large variety of machine learning methods on large datasets. It can support the whole process of data mining, starting from the preparation of data to the statistical evaluation of the models. The workbench includes a wide variety of methods such as regression, classification, clustering, association rule mining, and attribute selection. Furthermore, it supports streamed data processing. The system is open-source software, written in Java and freely available [Hall et al., 2009]. Its recent version provides base *map* and *reduce* tasks, Hadoop-specific wrappers and Spark-specific wrappers.

According to Berthold et al., the **Konstanz Information Miner (KNIME)** is an an open source modular environment which enables easy visual assembly and interactive execution of data pipelines. It is designed as a teaching, research and collaboration platform and provides integration of new algorithms and tools and data manipulation or visualization methods. Its great advantage is the powerful user interface, offering easy integration of new modules and allowing interactive exploration of analysis results or models. Combined with the other powerful libraries such as the WEKA and the R language, it provides a platform for complex and massive data analysis tasks. The Konstanz Information Miner is continuously maintained and improved through the efforts of a group of scientists and is offered freely for non-profit and academic use [Berthold et al., 2008].

[10] R: The R Project for Statistical Computing, accessed in May 2016 via `https://www.r-project.org/`

[11] Data Mining Group, Accessed on March 2016 via `http://www.dmg.org/`

Apache Mahout is an open source software project hosted by the Apache foundation. It provides scalable machine learning library on top of Hadoop, with the goal to provide machine learning algorithms that are scalable for large amounts of data. The development has been initiated with the paper from [Chu et al., 2006]. Up to now, several dozens of algorithms have been implemented for data clustering, data classification, pattern mining, dimension reduction. All algorithms are written in Java and make use of the Hadoop platform [Owen et al., 2011]. Apache Mahout supports in particular collaborative filtering (user-based and item-based collaborative filtering, (weighted) matrix factorization), classification (logistic regression, Hidden Markov Models, Multilayer Perceptron and Naive Bayes) and clustering (Canopy, (Fuzzy and Streaming)K-means and spectral clustering).

As a summary of reviewed DSS tools and technologies, Apache Mahout is the only tool that is released with the origin of Hadoop and MR based support. Furthermore, it provides very well performance for disparate data sources specially for distributed sources across multi geographical servers. But it lacks some machine learning algorithms in which most algorithms are not yet supported and in-progress or being released from time to time. The R statistical language and the PMML offer the opportunity to combine a wide range of statistical methodologies and models. They are able to cooperate for processing massive data from diverse sources and producing output for feeding the DSSs and clustering algorithms. They lack integration with Hadoop based data processing.

In addition, there are not efficient and easy to use open source Hadoop-based solutions for both of R language and PMML. Moreover, IBM and Oracle offer commercial use of Hadoop based R programming. At the same time, PMML lacks easy and fast integration with Hadoop and MR ecosystem and provides less performance in this regard. Also, Weka provides basic maps and reduce tasks and lacks complex and hybrid machine learning approach deployment in the Hadoop and MR. These stated tools show low performance especially when there should be hybrid approach of machine learning algorithms in order to provide recommendations and clustering algorithms. As a result, such approaches can be developed using Java APIs designed for MR programming. This is further discussed in chapter 5.

2.3 Contribution to Science beyond state-of-the-art

The contribution of this work to the science is addressed from two perspectives (disciplines): (1) CKM as well as (2) data science and big data analytics. It consists of theoretical and practical contributions. Studied research projects in this chapter show that sector experts define domain specific required CK based on their own experience. Such domain specific CK does not match to other enterprises or sectors. In fact, all studied projects are applied in specific sector without an insight to be adapted in other sectors. As stated earlier in the RQ 1

(CKR model), a lack of common understanding of CK in heterogeneous enterprises and sectors results in difficulties like IC 1 (CKR model). In this regard, providing a general CKR model which integrates a wide range of competence definitions is supposed to be a first contribution of this work.

This work suggests a generalized CKR model in order to cover most of job requirements in different enterprises and sectors. The general CKR model proposed in chapter 3, provides easy adoption to any other sector and job description. This is because of its mathematical background and representing domain specific CK with matrices and indicating the weights. The weights can be different for various sectors or enterprises. One can also add or remove further competences in the CKR model. As an example, current studied competence models for software engineering sector cannot be easily adopted and used in the teaching sector. It is because most of them use ontologies and depend mainly on the context of target sectors. This contribution is not a type of software and/or algorithm, but a general and theoretical CKR model with mathematical background mainly using matrices. It is further discussed in chapter 3.

As a second contribution, proposed CKR model contributes to a global job performance measurement and assessment. As an example, current assessment models of employees' CK doesn't support their EDR and assistance in making key decisions in enterprises. Most of CKM studies depend on the context and specifications of target domains. In this way, it is not required to study which CK is required for which specific sector. This harmonizes job description model as well as workforce development and planning. As a result, further and future research should adopt the CKR model and focus on the contextual research in assessment rather than developing a competence model from scratch.

The next contribution is an efficient modeling and representation of intangible CK through profiling method which is discussed in chapter 3. In this regard and in the frame of the IC 2 (profiling) (see Section 1.2 on page 8), enterprises will succeed in computerization of their CK and will be aware of: (1) CK lacks in specific sectors (competence gaps), and (2) the CK strength and highlights of employees in order to better match them to identified CK gaps. (3) the EDR processes and better planning and development of the competence development (4) prediction and mining of the CK and get prepared for them (this is discussed in the future work of this dissertation) (5) updating the curriculum and content of VET and formal training in universities based on the labor market needs.

In addition to CKM area as a case study of this research, the current work contributes to the data science and big data analytics as well. Accordingly, traditional data analysis and management technologies such as RDBMS are unable to fully exploit the potential of big or complex data. This is due to three main reasons:

1. the sheer amount of data nowadays that cannot be handled by traditional approaches (see section 1.2 on page 8),
2. unstructured and heterogeneous nature of today's data sources, and

3. the missing DM capabilities that can process huge data and bridge the gap between raw data and the sector-specific questions of data analysts.

Consequently, this work contributes to the utilization of data-stream processing that is (1) redundantly scalable for huge volumes of data, and (2) capable of working with heterogeneous multi-modal data sources and streams. In this regard, applied big data analytics is one side of the contribution, and integration and modeling of disparate data is the other side of it. An outcome here is a practical algorithm that can be applied in industry and/or science of analyzing disparate volumes of the HR data and can be adopted in other sectors and case studies. The target solution reduces data movement and replication overheads, scaling analytics horizontally to be more economical and efficient. The question of how to analyze the data and which goals to achieve through big HR data analysis is being answered in chapter 5 and addressed by RQ 3 (Scalable Clustering).

It is clearly discussed in chapter 5 that processing of about 15 million employee workforce profiles takes about 12 days through traditional computers. The delay is partially because of distributed and disparate data sources as well as large data sets that traditional algorithms are unable to process them. Very simple data set of those employees without involving any special streamed data from social networks is about 1 gigabytes and cannot simply be processed by personal computers. Processing consists of preparing, cleansing, clustering, analysis, and visualization of the job qualification data. The processing can be reduced down to 9 minutes through suggested solution in this research (see chapter 5). It can even be improved to reduce this processing time by scaling up the system or developing further algorithms as discussed in section 7.2.

A big challenge in this work as stated earlier is to prepare enough volumes of real big HR data in order to test and evaluate proposed big data algorithms. In this regard, the only success was to summarize CK data of 200 talents in cooperation with industries. One important contribution of this work is to discover the statistical distribution of this small dataset and regenerate and enlarge it for up to 15 million talents' data. In this way, the algorithms can be tested and evaluated with real big data volumes. Moreover, the data retrieval from digital sources such as web and social media streaming is the main contribution in this regard. This is discussed in details in chapter 4 and addressed by RQs 2 (statistical distribution), 3 (Scalable Clustering) and ICs 4 (CK retrieval from the web) and 5 (VET recommendations).

Through data retrieval from digital sources in the case study of this research, the bibliographic data of over 49 million research publications and 29 million authors have been collected from freely available source in the web. This is a part of integrated CK data for target case study of this research which is an academic research career in computer science area. Prepared datasets of this research can be used as an standard CK dataset for further research in the big HR data in future. This is very important contribution, since there is always a lack of accessible datasets in the HR area. The data sets can be shared with future researchers

upon to request. This final dataset is about 1 terabytes in size and can be used for any type of competence analytics and prediction algorithm which deals with HR qualification data.

Considering the European refugee crisis and the main problem of skill-mismatch, this work efficiently facilitates e-recruitment in a form of skill discovery and matching solution. As a final issue in the contribution of this dissertation, suppose that there are some talents who are aware of their competence gaps and would like to improve their qualifications and competence development plan. They will need a DSS in this regard. One novel part of the dataset preparation and regeneration in this work is the process of setting up 75,000 course profiles in the dataset. Using this dataset, a hybrid DM and recommendation approach provides further competence training and improvement recommendations. This hybrid DM approach consists of TOPSIS, AHP and K-means clustering algorithms. It is discussed further in details in chapter 5.

As a summary of this section, the solution approach given in this research contributes clearly (1) to the competence gap identification in enterprises, (2) better skill matching algorithm (skill miss-match problem), and (3) scalable analysis of large volumes of the data. These are stated as RQs 3 (Scalable Clustering), 3 (Scalable Clustering) and ICs 5 (VET recommendations) and 6 (best-fit talent) in section 1.2. In general, the contribution of this PhD is abstracted in Table 2.3.

Table 2.3: Summarizing the contribution of this work to the science beyond state-of-the-art

Description	Problem statement as	Solution Approach	Contribution type
Standardized and general CKR model with inclusion of a wide competence descriptions and supporting different sectors and enterprise strategies	IC 1 (CKR model) IC 2 (profiling)	chapter 3	Theoretical
Utilization of Redundantly scalable data-stream clustering for huge volumes of data for processing of heterogeneous multi-modal data sources and streams	RQ 3 (Scalable Clustering) IC 3 (big data & TA)	chapter 5	Theoretical
Identification of potential HR data sources, regeneration, simulation, interpretation of data and argumentation of big data pros and cons to target case study	RQ 2 (statistical distribution) IC 4 (CK retrieval from the web)	chapter 4	Practical
Providing a hybrid approach for matching of job seekers to already opened job positions	RQ 1 (Skill mismatch) RQ 3 (Scalable Clustering) IC 6 (best-fit talent)	chapter 5	Theoretical & Practical
Support skill workers in making the right decision to improve their competence gaps and competence development plan through recommending further courses	RQ 3 (Scalable Clustering) IC 5 (VET recommendations)	chapter 5	Technological

2.4 Conclusion of the Chapter

All in all, an intensive literature review in this chapter shows that the CM and CKM research and analysis demands further contribution in order to improve job performance in enterprises. It even became recently more important in computer science, since earlier efforts are mainly from psychological and HRM points of view. Particularly, reviewed definitions of key publications about CM results in providing a summary about general definition of competence and CM in section 2.1. In addition, it covers required background for identifications of different CK dimensions and its hierarchy in the CKR model. Table 2.1 on page 20 summarizes all studied definitions. The most common definition which is also referred and used in this dissertation is from EC.

As a conclusion of section 2.1.1, any CMS should consist of at least: (1) identification of required competences (competence discovery), (2) assessing acquired competences (competence assessment), (3) matching acquired and required competences (competence analysis), and (4) providing further competence development plan as well as recommendations for improving the competence gaps (competence development). Studied literature in section 2.1.2 show that most research efforts focused on the utilization of ontologies and traditional paper (or interview) based competence assessment methods. These methods cannot be applied to new sectors in a short time. In general, 86% of studied relevant research projects in section 2.1.2 focus just on one specific sector. Therefore, their results and algorithms cannot be easily applied in an analysis of competences in other areas. Most of them focus on assessment and modeling of teaching competences.

The main challenge and limitation that similarly addressed in most of the literature is a lack of efficient and generalized competence matching method. The common functions of such competence management and analysis systems consist of modeling of required and acquired competences, gap identification and curriculum development. Another challenge identified through literature review is the growing volume of the HR data and lack of scalable algorithms in this area. Utilization of the big data in HRM is always stressed as an important challenge in this area, but real practical work in this regard is missing. For instance, all reviewed research projects lack utilization of big data and cloud computing technologies in competence assessment and management [Keshavarzi et al., 2013]. A summary of the literature and goals associated in using them are given in Table A.1 as an Appendix A. This table clearly shows for instance which works are background of this dissertation, which ones are fundamental literature, which ones are disagreed works and vice versa.

Moreover, in the vast ocean of big data tools and technologies, it should very well understood which big data technology provides better results for domain specific preferences. In this regard, short review of available big data technologies and their pros and cons have been given in section 2.2. In addition, platforms providing scalable multi-criteria decision making algorithms are studied in section 2.2.3. Tools such as R language, PMML, Weka, KNIME and Apache Mahout

have been reviewed in this regard. These tools are traditional and commonly used tools in the DS and knowledge management areas. They are also adopted and improved recently in order to support big data platforms as well. The conclusion is that those studied tools do not match individually to the requirements of this research, so a hybrid approach has to be developed in order to support different perspectives. This is further discussed in chapter 5.

As final conclusion of this chapter, contribution of this work has been defined in section 2.3 beyond state-of-the-art literature review. This contribution covers defined RQs and ICs in section 1.2 in one hand and identified gaps through the literature review on the other hand. The main contribution of this work which is from theoretical and practical perspectives consist of: (1) CKR model, (2) job performance analysis and measurement, (3) scalable data stream processing, (4) HR data retrieval and regeneration, and (5) scalable DS and recommendation for competence planning and development. These contributions and further details have been summarized in Table 2.3. The reviewed literature in this chapter and their relevance and connection to the goals of this research have been summarized in Table A.1.

	HBase	MongoDB	CouchDB	Cassandra	MySQL	Redis	Riak	Membase
Data Model	Colum-oriented	Document-oriented (BSON)	Document-oriented (JSON)	Key-value	Relational	Key-value	Key-value	Key-value
Licence	Apache	AGPL (Drivers: Apache)	Apache	Apache	Commercial or GPL or FOSS	BSD	Apache	Apache 2.0
Data Type *(review)	row	document	document	value	record	value	value	value
Horizontal Scaling	Partitioning	Sharding built-in	Sharding only with BigCouch or Couchbase	Multi-node capabilities	Partitioning	Redis Cluster, (in development)	RIAK cluster, multi-node	cluster; multi-node (sharding)
Replication	HBase cluster; master-push	One server in set/shard is active for writes; Replica Sets and Master-Slave	Excellent; Master-master replication; Master-slave	Good; Master-slave	Master-slave, multi-master, and circular replication	Master-slave	Masterless multi-site replication	Master-master replication; Master-slave
Query Method	Map/reduce with Hadoop	Dynamic; some SQL. (Query, index) properties	Pre-defined JS queries (views);	Map/reduce possible with Apache Hadoop	Dynamic; SQL	Simple values or hash tables by keys	FT search, indexing, querying with Riak Search server (Beta)	Memcached-protocol
Interface	HTTP/ REST (also Thrift)	Custom, binary (BSON)	HTTP/ REST	Custom, binary (Thrift)	Native drivers	Telnet-like	HTTP/ REST or custom binary	Memcached & extensions
Written in	Java	C++	Erlang	Java	C & C++	C++	Erlang, C & JS	Erlang & C
Atomicity	row-level	on single documents	on single documents	within a single column family.	row-level; ACID-compliant if transactional storage engine used - like InnoDB	on individual data structures, like counters, lists and sets	no	on a single item at the server level
Some of application domains	Search engine look-alike solutions	Dynamic queries; alternative to SQL	CRM, CMS systems	Banking, financial industry	RDBMS typical applications - most of which is not Big Data	Analytics. Real-time data collection. and communication	Point-of-sales data collection. Factory control systems	Highly-concurrent web applications;
Main Point	Hadoop/HDFS; realtime BigTable-like data	Big DBs with SQL-like querying	Accumulating data with little change; good versioning	Write more than read (logging); completely in Java	Good all-around RDBMS	For rapidly changing data with database in RAM	Good single-site scalability, availability and fault-tolerance	Low-latency data access, high concurrency support and high availability

Figure 2.5: Comparison Grid of the Big Data Technologies

Chapter 3

Career Knowledge (CK) Profiling and Representation

»All our knowledge has its origins in our perceptions.«

– Leonardo da Vinci

Profiling is the first and most important step in computerization and mathematical representation of the CK in the HR information systems. The CK profiling consists of methods for identification of knowledge sources, discovering associated CK and representing them for proper analytics. It is a key method used in this work in order to model and convert those tangible and intangible CK into numerical values. Tangible CK refers to qualifications like degrees, certificates and all others that are issued by certificates. Assessing such types of knowledge is easier than non-certified intangible CK. In contrast, intangible CK refers to personal expertise, skills, behavioral issues and competences that are not (and cannot be) issued, assessed or certified. In fact, this is about personal qualifications of talents and depends heavily on their character and personal behavior.

The CK addressed in this work is not only about the competences of talents, but also skills and competences that jobs need to be handled successfully. In fact, it is about HR knowledge as well as a job specific knowledge. The first one (i.e. acquired CK) should be measured and identified practically through assessment methods. A job specific knowledge (i.e. required CK) should be defined by domain experts or top managers. It is imperative to use the same representation model for acquired and required CK in order to efficiently provide competence gap analysis and person-job-fit solutions. To this aim, a CKR model is developed in this research as standard CK definition and representation model. This model is discussed more in details in section 3.1 on page 54. The CK is being defined as following in this dissertation:

Definition 1. *The Career Knowledge (CK) is any type of skills, knowledge, expertise, tasks, qualifications, degrees and professionalism that is connected to a job definition or human resources. A CK may address partial or all requirements of talent's or job profile. A talent's profile specifications are acquired CK and job profile specifications are required CK. The CK is being referred as job knowledge in some literature.*

The *CK profiling* refers to extracting, collecting, assessing, measuring and

discovering job specific knowledge of talents. It is based on the outcomes of (1) self-assessment through involving talents (2) multi-source assessment through involving talents' work circle (e.g. colleagues), and (3) retrieving the talent associated data from web and digital sources. A TP consists of general demographic information of a person as well as his assessment results. As soon as a TP is completed with real values (i.e. acquired CK), it can be used for person-job-fit decisions. A person-job-fit decision resolves the skill mismatch challenge addressed by the IC 1 (CKR model) on page 10 through assigning (fitting) a talent from the pool of candidates to the desired job position. It is based on a classification of talents with respect to the required CK in the job description and acquired CK of talents.

Definition 2. *The person-job-fit process is a method of prioritizing and selecting the best-fit candidate from the number of candidates (i.e. alternatives) based on already defined competence levels (i.e. criteria) in specific job position. The policy of selecting the best-fit candidate (i.e. algorithm) may vary depending on an enterprise and its job quality measures (assessment methods).*

The CK profiling covers two different dimensions: an employee (i.e. talent) and an employer (i.e. talent seeker). The CK profiling from employer's side is goal oriented and identifies all required CK for specific job definition. Outcomes of this process are referred as *JPs* or *required CK* in this research. The JP consists of required competences, roles, expertise, working conditions and importance level of competences for specific job. A major output of this process is a weighting matrix of the target job referred as RCK matrix. Based on defined weights in the RCK matrix, talents are classified in clusters like over-qualified, best-fit, and under-qualified. Different JPs result different talent clustering results. For instance, two talents may belong to the same cluster like over-qualified for specific job, but perhaps grouped differently for another job definition.

Employees' CK profiling is referred as acquired CK in this work and uses CA to retrieve the level and quality of talents' competences, certificates, skills and all other activities in order to store as Acquired Career Knowledge (ACK) matrix in talents' profiles. Outcomes of this CA provide required data to decide if a talent is competent enough to handle specific job or not. The CA process is independent of desired JPs and doesn't consider required CK while assessment and data collection phase. In fact, this method answers to the question of "What and how much does a talent know?". An assessment is on the basis of standard CKR model discussed in the following sections. This type of assessment should cover different perspectives and types of collecting the talents' job qualitative data.

Consequently, the CA consists of two methods: (1) multi-source assessment, and (2) self-assessment. Every assessment method has a weight in the JP, meaning that the importance of assessments differs depending on the job and TPs. The multi-source assessment method is on the basis of 360-degree feedback and involves an immediate work circle of a talent to provide their opinion about talent's qualifications and professionalism. Since there are multiple respondents to this

assessment, their replies are summarized using weighted arithmetic mean. In the case of not being an employee of an enterprise which hardens justification and assignment of people from his direct work circle, a talent has to nominate referees to employer for 360-degree feedback process. The 360-degree feedback method in this case receives a lower importance (weight) in matching algorithms, since respondents are not familiar to employer and the quality and trueness of their feedback is not known to an enterprise.

In the self assessment method, a talent participates in an assessment and evaluates the level of his CK. The base of self assessment is questionnaires that consist of associated statements (questions) to the competences from the repository of competence definitions. The self assessment plays a key role in identification of intangible CK. In the self-assessment method, a talent should respond to domain specific and contextual questions in a form of an exam. The questions in such an exam are for instance professional questions about programming for software engineering jobs. The challenge in this assessment method is to manage an on-line exam which has some organizational difficulties. Those difficulties are not focused in this dissertation due to the fact that there are some known methods for on-line tests that enterprises manage them based on their own strategy.

In addition to assessment results, web based sources like digital sources of enterprise as well as web-based DBs such as DBLP are used to discover talents' competence data. These sources cover most of competences like social or even professional CK. Data retrieval and streaming methods collect the data from web. The outcomes of this method should also be processed and summarized in talents' profiles. It delivers a large volumes of the data specially through streaming social networks and crawling of job descriptions from the web [Bohlouli et al., 2015b]. This method is discussed in section 4.3 on page 90. As stated earlier, the results of different profiling methods are called *TPs* or *acquired CK* in this research.

In the following, the CKR model developed in this work is represented and discussed in section 3.1. The CKR model is the basic model for profiling and collection of HR and job data. It can be adapted to any field of jobs and sectors like IT, chemistry, mechanical engineering or administrative jobs. An academic computer science career is the case study of this research which is described in section 3.3. The argumentation about why this area is selected as the case study of current dissertation is discussed in this section as well. The study in this section supports further development and adaptation of the CKR model in other scenarios/sectors/case-studies.

In addition, the theory of CK profiling is discussed in section 3.2. This section covers the goal of profiling and also mathematical definitions and background that will be used for profiling in the target case-study of this research. Different profiling methods used for two different perspectives that stated earlier (employee and employer) are addressed in this section. Additionally, it covers the mathematical outcomes (job and talent matrices, ACK and RCK) of the profiling algorithms. The latest section (3.4) provides a conclusion and how results of this chapter facilitates issues in the later chapters.

3.1 Career Knowledge Reference (CKR) Model

Specifications of the CK is not same in various sectors. It may even differ for various enterprises in the same sector. According to such heterogeneity of CK models in different sectors and enterprises, talent analytics solutions may not work efficiently in all of them. In order to solve this problem, developing a generic CK model facilitates easy adaptation of talent analytic solutions to different sectors and enterprises. It has been developed in the frame of this work as a general Career Knowledge Reference (CKR) model. A wide variety of sectors and their required CK have been studied while specifying this model. In other words, the most (probably all) required CK in different sectors are foreseen in the CKR model. This model has been inspired from the *Professional, Innovative, Social (PIS) competence pyramid* in the frame of ComProFITS project[1] [Bohlouli et al., 2013a] and [Mittas et al., 2015] and the *PIS competence matrix* at the CoMaVet project[2].

In addition to the literature study and practical experiences in this research, a survey analysis of 186 participants from industry and academia played a key role in developing the CKR model. About 44% of participants in this survey are academic experts and the rest (i.e. 56%) is from industry. Likewise, 73% of survey participants have a background in computer science and the rest (i.e. 27%) is from different disciplines. Two objectives are followed in conducting this survey: (1) To find out whether CKR model is comprehensive enough to cover all required CK of different sectors, disciplines, cultures and enterprises. (2) To customize CKR model in academic computer science career and get expert feedback in specification of required CK in this sector. Outcomes of the first objective results in fine tuning of the CKR model which is discussed in section 3.4. In addition, the second objective specifies weights of CK categories in the CKR model.

According to the responses in this survey, 85% of respondents believe that the CKR model covers all competences that they expect from such a model. In addition, 78% agrees that having such a model will improve documentation and standardization of their job description and e-recruitment processes. Interestingly, 83% confirmed that they are interested to use an e-recruiting software which is based on scientific algorithms for recommending qualified candidates. Likewise, 66% of respondents doesn't use any talent assignment and recruiting software as well as any CM tools. Results collected through this survey are discussed in the following paragraphs and also in Section 3.3. Figure 3.1 on page 55 shows the distribution of survey participants with further details.

The CKR model is a generic categorization of CK with multi-layered architecture. A tree-like structure of the CKR model consists of three layers. All level-1,

[1]Competence Profiling Framework for IT sector in Spain (ComProFITS) is funded by EU. A project information accessed in November 2015 through `www.comprofits.eu`

[2]Competence Management in the European VET-Sector (CoMaVet) project is funded by EU. A project information accessed in November 2015 through `http://www.adam-europe.eu/adam/project/view.htm?prj=3962`

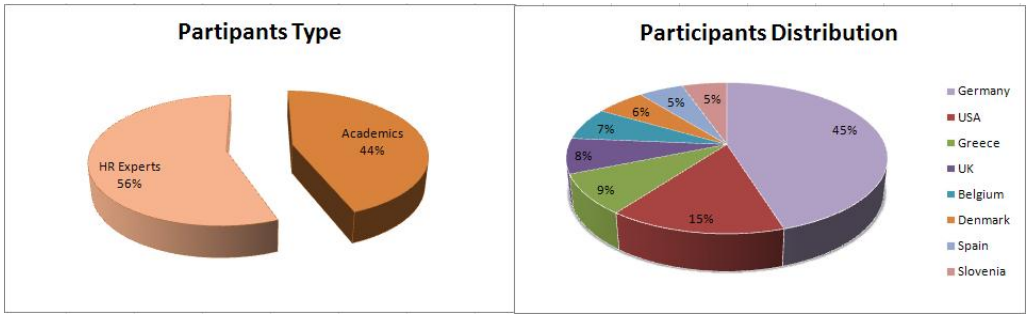


Figure 3.1: Distribution of participants in the survey analysis of this research for conducting CKR model

level-2 and level-3 categories in the CKR model are referred respectively as CK category, subcategory and sub-subcategory. The lowest layer of this model covers 64 CK sub-subcategories (e.g. competences). Any CK category in all of three levels in this model contains a *heading* and *context*. The *heading* covers the subject, title and id of any CK category. A *context* consists of sector specific *attributes*. *Attributes* are, for instance weights of required CK, a parent CK category (higher level) in the tree, importance factor for assessment methods, and CK description. A hierarchical architecture of the CKR model is shown in Figure 3.2 on page 56.

A title, total number of competence categories in higher levels (i.e. level-1 and level-2) as well as the depths of levels (i.e. 3 levels) in the CKR model is fixed for different sectors and enterprises, meaning that the model uses the same hierarchy for all different case studies. In fact, the heading of the CKR model is independent of target sector or case study and is same in different case studies. But, the context is being defined differently for various sectors and case studies. The context covers at least (1) statements such as "What is this competence category and which expertise are required for it?", (2) importance, (3) weight such as "How important is this competence?", and (4) the level of required CK such as "strong, basic or medium knowledge in competence X".

Any CK category (level-1) consists of 4 sub-categories (level-2). Mutually, a CK sub-category is composed of various numbers of sub-subcategories (level-3). Assuming a not fixed number of sub-subcategories provides more generalization in the CKR model, whereas enterprises can fully customize it depending on their needs and strategies. It is also possible to define a fixed number of sub-subcategories and give a 0 weight to some of them that are not required. This will provide more homogeneity to the model. In the frame of this research, the fixed number of 4 sub-subcategories have been considered for each sub-category. As a result, there are total number of 64 sub-subcategories in the model. Listing 3.1 on page 60 shows a simple XML representation of a JP with CKR model.

Normally, domain experts, head of departments or HRM employees use CKR model to configure job specific needs of a target job. In setting up a JP using CKR

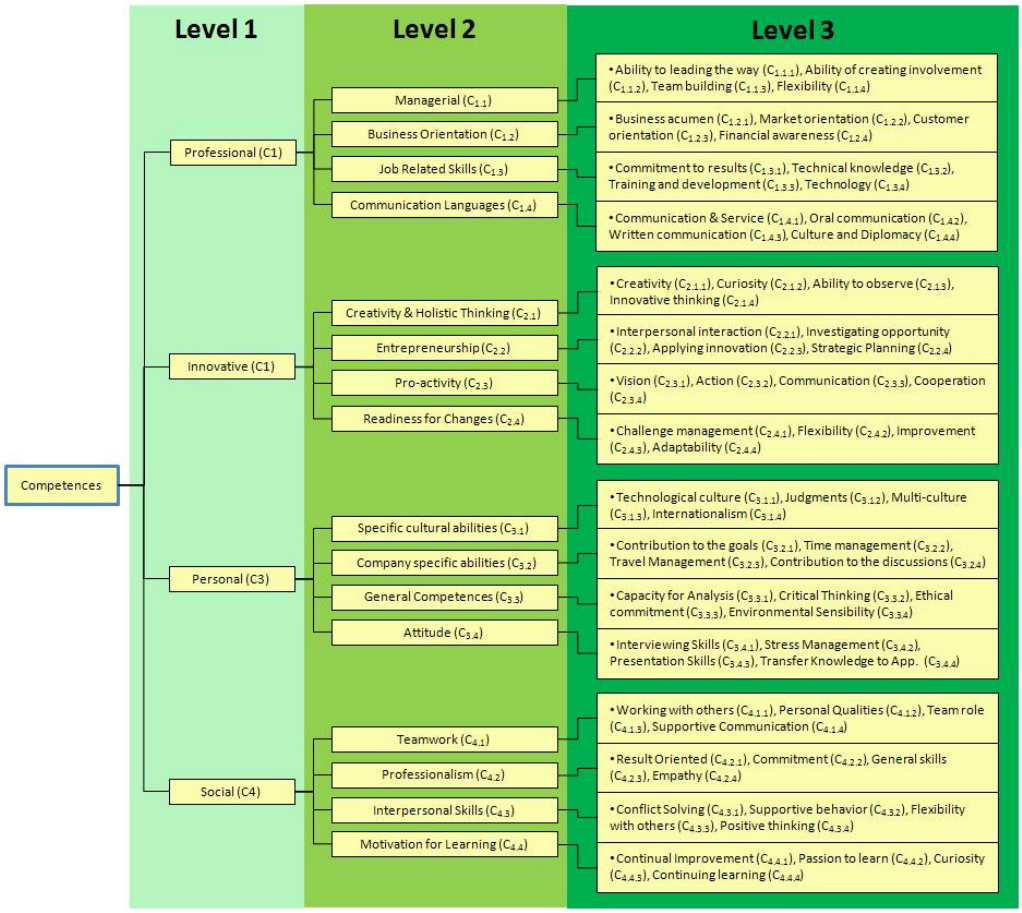


Figure 3.2: Career Knowledge Reference Model

model, a level of expertise required for a competence should be defined by giving the weights and identifying their level of importance [Bohlouli et al., 2015a]. This means defining required CK for initiating a JP. Categorization of the CK is key to person-job-fit approach and skill miss-match challenge addressed as IC 1 (Skill mismatch) in section 1.2.1. Definition of *attributes* for each category initiates, for instance, required CK matrix called RCK which is described in Section 3.2. This matrix consists of defined weights in the *context* of the CKR model. Additional *attributes* to the currently available ones can be defined by enterprises in the *context* of categories.

The level-1 categories in the CKR model are professional, innovative, social, and personal competences. This level is the highest level consisting abstraction of lower layers' measurements. For example, assessment results of 16 sub-subcategories from $C_{1.1.1}$ to $C_{1.4.4}$ are summarized (using weighted mean calculation) as a C_1 in the level 1. This abstraction follows a hierarchical architecture of the model. Interpretation of the level-1 values doesn't support an accurate person-job-fit decisions. Therefore, this level is used mainly to visualize *collective competences*

of enterprises. As a result, an insight about general strengths and weaknesses as well as collective competence gaps of enterprises can be achieved in this level [Bohlouli et al., 2012b]. As an example, facts such as "An enterprise X lacks, in general, social competences." can be achieved by analyzing the values in this level. Figure 3.3 shows achievement of such results through visualizing the level-1 CK categories.

Boreham defined *collective competences* as using three normative principles: "making collective sense of events in the workplace", "developing and using a collective knowledge base", and "developing a sense of interdependency" [Boreham, 2004]. In fact, collective competences define traditional individual competences in the context of teamwork and group competences. An accurate analysis and visualization of the level-1 values support collective competence gap prediction which is key in an orientation and definition of new jobs. Figure 3.3 shows visualized level-1 CK in the case study of this research. A spider plot of visualized level-1 values in this figure shows that this enterprise lacks in general social competences. Therefore, its future job announcements will emphasizes social competences more in order to increase social competitiveness of the enterprise.

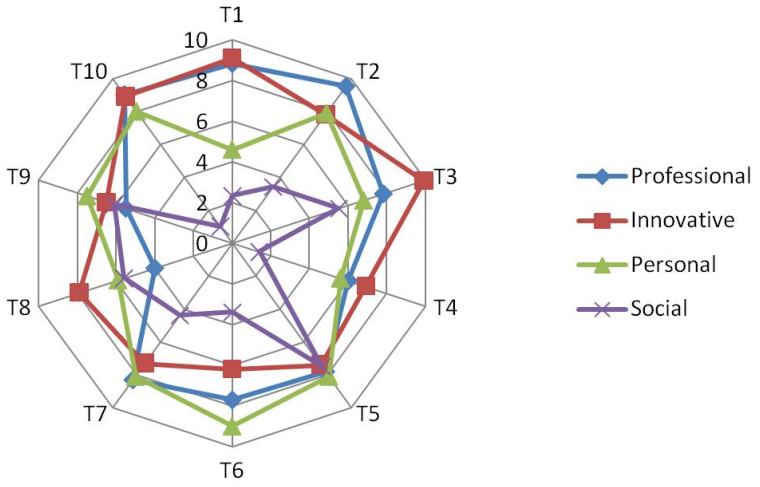


Figure 3.3: Visualized Collective Competences of an enterprise based on Level-1 CK from CKR model for an enterprise with 10 employees

However, a visualization of level-2 CK grasps inner competence needs of an enterprise. This is in mutual connection to enterprises' policies and requirements. Identification of social competences as a major competence gap in aforementioned example (Figure 3.3) does not efficiently provide detailed information about enterprises' competence gaps. Emphasizing social competences as a major competence gap should evolve more detailed competence configuration as well as enterprise specific strategic and cultural needs. As a result, setting up a JP based on the

level-2 sub-categories is the optimum solution that fulfills most of competence gaps. In fact, this level is a milestone of defining competence gaps in details. Visualized level-2 CK in the case study of this research is given in Figure 3.4.

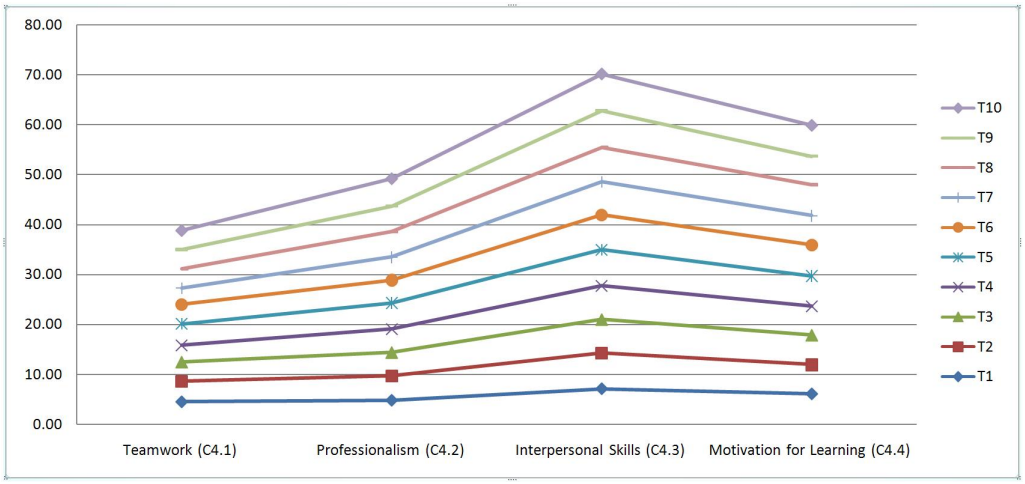


Figure 3.4: Competence Gap Identification and Analysis through Visualization of level-2 Competences in the CKR model

The most meaningful and valued talent data is assessment results of level-3 CK. In addition to the person-job-fit approach, values in this level can also be used for providing on-the-job-training for currently employed people. Visualization of collected data in this level is more effective in providing insights about potential competence lacks or improvement potentials of individuals. Therefore, this level focuses more on the individual level rather than an enterprise (collective) level. The assessment results in the level-3 provide enough information to decide if a talent is competent enough for specific job position or not. Values in higher levels of the model (level-1 and 2) are being computed based on the values of this level.

Each sub-subcategory ($C_{1.1.1}$...$C_{3.4.4}$) may consist of different CK *context*. As an example, if an enterprise needs to recruit a person with COBOL programming skills, this should be specified in the *context* of "Technical Knowledge" in the level 3 sub-subcategory ($C_{1.3.2}$). The $C_{1.3.2}$ focuses on required sector specific technical knowledge of any job. Collected results through assessments in this level are integer numbers in the range of [1..10]. There is a confidence ratio in this level for values collected through web-based sources and social networks. For values with confidence ratio less than 0.5, the collected data from the web is not confident enough to be considered in mapping algorithms. In fact, they do not provide adequate information about person's competences. More about this ratio is discussed in chapter 4.3.

Professional CK category (C_1) is more sector specific and covers required

technical, communication and managerial skills of the jobs. This group consists of "Managerial" ($C_{1.1}$), "Business Orientation" ($C_{1.2}$), "Job Related Skills" ($C_{1.3}$) and "Communication Language" ($C_{1.4}$). This category has higher weight in JPs like top management or very specific professional jobs such as database engineer. Requirements, weights and definitions in this category look like the equivalent for similar jobs in different enterprises and/or sectors. The self-assessment method has a higher importance than other assessments for this category. Retrieving web based professional CK data for computer scientists career as the case study of this research are collected through sources like dblp[3] and AMiner [Tang et al., 2008]. This is a bibliographic data of computer scientists.

A second CK category in this model is *innovative CK (C_2)* which focuses on creativity and readiness to changes. This category depends intensively on the culture and policy of an enterprise. The 360-degree feedback assessment is the main profiling method in this category. "Creativity and Holistic Thinking" ($C_{2.1}$), "Entrepreneurship" ($C_{2.2}$), "Pro-activity" ($C_{2.3}$) and "Readiness for Changes" ($C_{2.4}$) are sub-categories of this group. Normally, top managers and policy makers define the weights of sub- and sub-subcategories of this category. It is nearly impossible to find corresponding sources for this category through the web. An information about patents and inventions registered with the name of a talent in the web are convincing data sources for competences in the frame of this category.

A *personal CK category (C_3)* covers enterprise or talent specific needs, conditions, and cultural issues. This category consists of "Specific Cultural Abilities" ($C_{3.1}$), "Company Specific Abilities" ($C_{3.2}$), "General Competences" ($C_{3.3}$) and "Attitude" ($C_{3.4}$). Configuration of required *personal CK category* looks significantly dissimilar between various enterprises and case studies even for the same JPs. As an example, the responsibilities of a job like business intelligence analyst and data warehousing engineer look like similar for different enterprises/case studies. But, the weights given by different enterprises for required personal CK subcategories in this job are different. Because data warehousing jobs depend heavily on the ethics of data supposed to be analyzed. Accordingly, sub-subcategories like "Ethical Commitment" should have different weights.

The 360 degree assessment method has higher impact in collecting data associated with *personal CK category (C_3)*. As a result, this assessment method will have higher weights in the measurements. Discovering and identification of sources for this category from the web is not an easy task. Because it depends to the personal data which is subject to the ethics. In addition, it depends heavily to the personal character of talents as well the cultural and strategic policies in enterprises. It means that one specific talent may have different values for competences in this categories in different enterprises. The workforce culture has an important impact on this category. Therefore, exporting assessment results of this category from one enterprise to another one doesn't make sense.

[3]DBLP: The Computer Science Bibliography, Retrieved on 2015-01-28 from data maintained by the dblp team at `http://dblp.uni-trier.de/`.

The *social CK category (C_4)* in the CKR model focuses on teamwork and social inclusion of employees in the workforce development. This category consists of "Team Work" ($C_{4.1}$), "Professionalism" ($C_{4.2}$), "Interpersonal Skills" ($C_{4.3}$) and "Motivation for learning" ($C_{4.4}$). The *social CK category (C_4)* is enterprise specific CK and significantly differs for different enterprises. The momentous profiling methods of this category are 360-degree feedback assessment and data retrieval from the web-based sources and social networks. Social networks such as LinkedIn, ResearchGate and Bibsonomy are potentials web-based data sources of this category in the case study of this research.

```
<?xml version="1.0" encoding="UTF-8"?>
    <!-- Example consists of one CK category and one of its sub-categories. This model
    should be completed for all categories, sub-categories and sub-sub-categories of CKR
    model in practice. -->
 <header>
  <jobID>123456789</jobID>
  <jobTitle>Research Assistant in Software Engineering sector</jobTitle>
 </header>

 <body>
  <ckHeading>
   <id>1</id>
   <title>Professional</title>
   <level>1</level>
  </ckHeading>
  <ckContext>
   <!-- attributes list in the context-->
   <parentID>0</parentID>
   <weight>7</weight>
   <selfAssessmentWeight>9</selfAssessmentWeight>
   <multiAssessmentWeight>3</multiAssessmentWeight>
   <webAssessmentWeight>7</webAssessmentWeight>
   <description>professional requirements and expertise!!</description>
  </ckContext>

  <ckHeading>
   <id>7</id>
   <title>Job Related Skills</title>
   <level>2</level>
  </ckHeading>
  <ckContext>
   <!-- attributes list in the context-->
   <parentID>1</parentID>
   <weight>9</weight>
   <selfAssessmentWeight>7</selfAssessmentWeight>
   <multiAssessmentWeight>0</multiAssessmentWeight>
   <webAssessmentWeight>4</webAssessmentWeight>
   <description>Job specific professional knowledge requirements!!</description>
  </ckContext>
 </body>
```

Listing 3.1: Representation of a JP using CKR model with XML

3.2 The Theory of Profiling Career Knowledge

A proper profiling method improves quality of talent analytics in discovering tangible and intangible CK. In the frame of this work, profiling refers to modeling and extraction of domain specific CK for semantical and visual analysis of workforce development. Generally, it is aimed as data collection process. Given these points, inappropriate and non-relevant data collection method imposes further difficulties to the skill miss-match challenge. Outcomes of the profiling method are talent and JPs. Therefore, algorithms used for profiling differ depending on a target group (talent or job). Profiling of a talent's CK consists of three assessment types in this work: (1) The 360-degree feedback assessment, (2) Self-assessment, (3) and data retrieval from web-based and digital sources.

The assessment results of talents are stored as acquired CK matrices ($ACK_{m\times n}$), where m is a total number of assessment types, and n is a total number of competences. In this work, a total number of 3 assessment types ($m = 3$) and 84 competences ($n = 84$) have been defined. They can be any number in order to support full adaptability to other case studies and general approach. Assessment results are integer values in the range of [1, 10], where 10 is the highest (best) competence level. Any row in this matrix shows results achieved through one specific assessment method for different competences. Likewise, a column in this matrix indicates values assessed for one specific competence through different assessment methods. In fact, the ACK matrix is a mathematical representation of talent's profile.

This matrix plays a key role in the competence analytics algorithms. Values of the first two levels from the CKR model in the ACK matrix are computed based on assessment results in level l_3. As an example, the $ACK[1][2]$ indicates the assessment result of the C_2 in the CKR model achieved through assessment method 1 (i.e. 360-degree feedback). Similarly, $ACK[3][25]$ refers to the $C_{1.2.1}$ in CKR model which is achieved through web based CK discovery. In fact, data retrieval from the web results in the production of real competence values such as SCF and AIS as discussed in sections 5.2.2 and 5.2.3 for each talent.

The higher level competences (l_2 and l_1) are computed using level l_3 assessment results and their defined weights without using CA in this regard. An equal weight is given to the l_3 sub-subcategories for computing l_2 subcategories in this work. It is theoretically possible to define various weights to l_3 sub-subcategories as well. The weights in higher level competences are defined using HCV method. The HCV method is a type of cumulative voting. In a cumulative voting, all categories of the same level are being prioritized at the same time. In comparison, the HCV method represents higher correlation between categories that are being prioritized in the groups, and lower cohesion with other groups.

In the HCV method, the assignment of the weights is based on the allocation of imaginary units supposing that a sum of all amounts is equal to a fixed number of f (like 100) [Berander and Andrews, 2005]. In this method, items to be assigned

are structured into groups and the elements of the higher level are assigned first amounts summing to f by cumulative voting. Then the items of the lower level are prioritized separately within each group. In this manner, the prioritization of competences takes into account both hierarchical organization of competences (Figure 3.2) and the quantitative prioritization. It clearly demonstrates how much a specific competence is more important or relevant to a specific job than any other.

There are 4 categories in the level l_1 of CKR model (Professional, Innovative, Personal and Social) to be prioritized. Let C_i where $i = 1, 2, 3, 4$ denotes the i-th category and that to each one of the groups the following amounts are assigned:

$$w_i, 0 \leq w_i \leq f, i = 1, 2, 3, 4 \text{ such that} \sum_{i=1}^{4} w_i = f \tag{3.1}$$

The sum of allocated values (f) should be predefined. It is usually being set to $f = 100$ without loss of generality, since for any value of f, the weights can be easily transformed (through division by f) to have sum equal to 1. Since C_i is a l_1 competence group containing 4 items of level l_2, these have also to be prioritized in a similar way using HCV. For any category C_i in the level l_1, $i = 1, 2, 3, 4$, the prioritization of its 4 sub-categories are denoted as:

$$w_{ij}, 0 \leq w_{ij} \leq f, j = 1, 2, 3, 4 \text{ with} \sum_{j=1}^{4} w_{ij} = f \tag{3.2}$$

Similarly, the weights of level l_3 sub-subcategories can also be defined using HCV method using equation 3.3. For the ease of use, an equal weight for each sub-subcategory is defined for level l_3 competences in this research. Therefore, w_{ijk} is equal to 1/4.

$$w_{ijk}, 0 \leq w_{ijk} \leq 100, k = 1, 2, 3, 4 \text{ with} \sum_{k=1}^{4} w_{ijk} = f \tag{3.3}$$

Schematically, the weights are assigned according to the hierarchical architecture of the CKR model showed in Figure 3.2 on page 56. As it is clear from this figure, the HCV is a top-down approach that is initiated at level l_1 and terminated at level l_3 of the hierarchy. As an example, suppose that a distribution of f=100 units for the first level categories C_1 (Professional CK), C_2 (Innovative CK), C_3 (Personal CK) and C_4 (Social CK) is as follows: $w_1 = 60$, $w_2 = 20$, $w_3 = 10$ and $w_4 = 10$. It is clear from this distribution that the C_1 is the most important CK of level l_1 which is three times more important than C_2 and 6 times more important than C_3 and C_4. In addition, C_3 and C_4 are equally important.

It is important to note that if ACK and prioritization of competences (known as required CK, RCK) are combined in the 2nd level, the values assigned to 2nd level should take into account the values assigned to 1st level. This can be achieved by a simple multiplication and a normalization which results in values summing

to 1. The normalization and adjustment results in the absolute assignment of importance to each subcategories such that their sum regardless of a category is 1.0. The result of HCV method is RCK matrix which consists of CK weights in the l_1, l_2 and l_3 levels. Overall this thesis, i refers to the index of a level l_1 competence category, j is used as a level l_2 subcategory index and similarly k is for indicating level l_3 sub-subcategory index.

Rows in the RCK indicates the weights of one specific assessment method which is measured using Equation (3.4). It is clear that one specific CA method doesn't deliver accurate results for any types of competence, so competences shall be assessed using different CA method. As a result, different assessment methods should have various weights for various competences. Defining such variable weights for different CA methods provides even further generalization and mas customization to this model. Each row in the RCK matrix are calculated using Equation 3.4.

$$\omega_{a,x} = \begin{cases} w_i/10 & ,1 \leq x \leq 4, i = x \\ w_{ij}/10 & ,5 \leq x \leq 20, i = (x-1)/4, j = (x-1) \mod 4 + 1 \\ w_{ijk}/10 & ,21 \leq x \leq 84, i = ((x-5) \operatorname{div} 16) + 1, \\ & j = (((x-5) \mod 16) \mod 4) + 1, k = ((x-1) \mod 4) + 1 \end{cases} \tag{3.4}$$

where $\omega_{a,x}$ is the weight of competence category x in the assessment type a and w_i, w_{ij} and w_{ijk} are calculated using Equations 3.1, 3.2 and 3.3, and indicate weights of levels l_1, l_2, and l_3 competences, respectively. For example, w_5 shows assigned weight to the $c_{1,1}$ competence subcategory. Similarly, w_6 for $c_{1,2}$, w_8 for $c_{1,4}$, w_9 for $c_{2,1}$ and w_{20} shows the weight of competence subcategory $c_{4,4}$.

Initiation of the RCK matrix in the profiling is in fact the process of setting up a JP. This is referred as job profiling process. Computation of each row in the RCK matrix is discussed earlier and also in Equation 3.4. There are 84 columns in the RCK matrix based on CKR model and 3 rows (assessment types). The structure of showed as Equation (Eq. 3.5).

$$RCK^{\alpha}_{m \times n} = \begin{bmatrix} \omega_{1,1} & \dots & \omega_{1,x} & \dots & \omega_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \omega_{a,1} & \dots & \omega_{a,x} & \dots & \omega_{a,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \omega_{m,1} & \dots & \omega_{m,x} & \dots & \omega_{m,n} \end{bmatrix}_{m \times n} \tag{3.5}$$

where α is an already opened job position α (JP_α).

One important advantage of this matrix (RCK) is the separation of assessment methods with differing their weights and effects in the measurements. Separation of these assessment methods through different weights in the JP has an advantage of giving different importance to various assessment types. In this way, different strategies, policies and priorities can be given to specific assessment methods while setting up the JP. For instance, if self-assessment is more important due

to specific policies in an enterprise, it gets higher weights in the required CK. Definition of the weights in ACK is based on already described HCV method in the RCK matrix.

In order to calculate values of higher levels (l_1 and l_2) in the ACK matrix, values achieved through assessments in the level l_3, plus weights computed through HCV are required. Results of these measurements are collected in the ACK matrix as acquired CK. The structure of ACK matrix is showed in Equation 3.6.

$$ACK^{\tau}_{m\times n} = \begin{bmatrix} \sigma_{1,1} & \dots & \sigma_{1,x} & \dots & \sigma_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sigma_{a,1} & \dots & \sigma_{a,x} & \dots & \sigma_{a,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \sigma_{m,1} & \dots & \sigma_{m,x} & \dots & \sigma_{m,n} \end{bmatrix}_{m\times n} \tag{3.6}$$

where $ACK^{\tau}_{m\times n}$ is an acquired CK matrix of a talent τ and $\sigma_{a,x}$ denotes assessment result for CK category of x through assessment method a. The $\sigma_{a,x}$ is recursively computed from the (Eq. 3.7).

$$\sigma_{a,x} = \begin{cases} \frac{\sum_{i=4x+1}^{4(x+1)} \omega_{a,i}\times c_i}{\sum_{i=4x+1}^{4(x+1)} \omega_i} & , 1 \leq x \leq 20 \\ c_x & , \text{else } 21 \leq x \leq 84 \end{cases} \tag{3.7}$$

where c_i is assessed CK level of the level $l3$ sub-subcategory i.

Depending on the assessment type, values of the level l_3 (C_{21} up to C_{84}) are measured through specific assessment algorithm which are discussed in the corresponding subsections in the following. In fact values of level l_3 are based on the practical assessment processes of individuals. Values of the levels l_2 and l_1 are computed using (Eq. 3.7) based on assessment results at level l_3.

3.2.1 360-degree Feedback Method

A talent, three of his immediate work circle and his immediate manager form an assessment team in the 360-degree feedback[4]. The results are summarized through a weighted arithmetic mean of all participants in the 360-degree feedback. This method is entitled as a multi-source assessment competences in talents' profile. It is helpful in collecting the viewpoint of colleagues who work closely with a talent at the same level or department. The values collected through 360-degree feedback are facts and figures like rating of talents' qualification and professionalism level for associated competences. In fact, it collects intangible acquired CK. Normally, a head of department initiates an assessment team and manages the assessment procedure.

[4]The development and discussions about 360-degree feedback method in this section and overall this thesis are based on the research, development, experience and documentation achieved through coordination of the Competence Profiling Framework for IT sector in Spain (ComProFITS) project.

Each of competence sub-subcategories in the level l_3 is assessed through at least four statements. Accordingly, four competences at level l_3, each has an assessed competence value achieved through the mean of the statement evaluations. All data are stored in the lowest level, which means that obtained integer value is related to the following information:

- Selected competence in level l_3,
- date and time of the assessment,
- specific statements' text,
- specific value related to the statement, and
- the person who has given the assessment value

Based on the assessment results in the level l_3, it is possible to create competence analysis on several higher levels by specific calculations and statistical analysis. The HRM department in cooperation with the specific Head of Department maintains the weights, when the job is being defined. This process is suitable when an expression of the competence level of a talent or a group of talents (a department) are wanted (human capital). The specific job related competences are weighted by means of the matching weight factor. The 360-degree feedback is suitable for employees in an enterprise in order to identify their competence gaps and match them to the goals defined in their EDRs. It is not recommended for job seekers who are not currently employed in an enterprise. To this aim, a self-assessment method is preferable.

Outcomes of the 360-degree feedback assessment are integrated in the ACK matrix. In addition to the achieved values through an assessment, the weights of this assessment should be defined in the RCK matrix. Both weights and achieved values are required in the matching algorithms. These values get lower weights for people who are not employed in an enterprise, since there is not convincing confidence for this type of assessment for not-employed users.

3.2.2 Self-Assessment Method

There is a pool of statements for any of competence sub-sub-categories in the level l_3 of the CKR model. These statements are like exercises or questions in exams or tests. The focus of this work is not to define or provide a description to those exercises or to study their influence in the results of an assessment. The statements target specific competences and are defined by the head of the department or person who initiates a JP. Results of this process are exams like Test of English as a Foreign Language (TOEFL) or Graduate Record Examinations (GRE) in order to evaluates tangible and intangible CK of talents. For each JP, a set of related competences is selected from the CKR model which results initiation of a proper test for selected competences.

As an example, imagine that the goal is to employ a software developer with background in the COBOL language. To this aim, setting up a self-assessment test evaluates his technical knowledge in COBOL. Evaluating this type of knowledge through 360-degree feedback is not accurate and representative. Because the specific competence can be more or less important for the job function, a similar weighting mechanism to 360-degree feedback is adopted to specify a weight to each competence that belongs to the specific JP. This method is useful specially for evaluating tangible CK. It consists of multi-choice questions with one or more correct answers. The process of computing final result through self-assessment is showed in Figure 3.5.

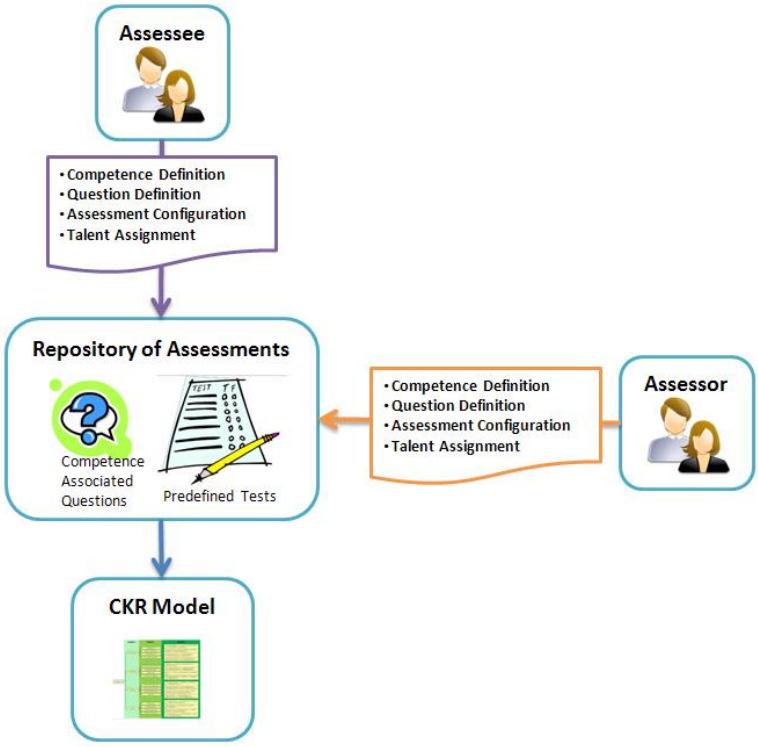


Figure 3.5: Architectural overview of the self-assessment method

For calculating the total score in multi-choice questionnaire, standard formulas described in [Bandaranayake, 2008] are used. Further details are described in [Yang, 2015].

Symbol	Definition
m	Total number of assessment types
n	Total number of competences
$ACK^{\tau}_{m\times n}$	Acquired Career Knowledge (CK) Matrix consisting values of m assessment types for n competences
$RCK^{\alpha}_{m\times n}$	Required Career Knowledge (CK) Matrix consisting weights of n competences through m number of different assessments
w_i	The weight given through HCV to the level l_1 category i in the RCK
w_{ij}	The weight given through HCV to the level l_2 category ij in the RCK
w_{ijk}	The weight given through HCV to the level l_3 category ijk in the RCK
$\omega_{a,x}$	Computed weight for assessment type a and competence category x based on weights defined in w_i, w_{ij} and w_{ijk}
τ	A talent profile τ
α	A job profile α
$\sigma_{a,x}$	Final computed elements of the ACK matrix for assessment type a and competence category x

Table 3.1: The summary of defined mathematical symbols and equations in the chapter

3.3 CKR Model in Academic Computer Science Career

The CKR model has been adopted to academic computer science career as a case study of this research. In this regard, a survey study with participation of 186 volunteer experts has been conducted. The questionnaire was distributed between around 2,450 experts. It identifies the importance and weights of competence categories, subcategories and sub-subcategories from the CKR model in the target case study. The outcome of this step is first, to understand whether the CKR model supports general modeling of required competences in different sectors and enterprises. This issues a general perspective of the CKR model. Second, to identify a proper configuration of RCK matrix in order to define the weights of required CK in the target case study. The selection of academic computer science career as a case study is based on the following four important reasons:

1. This field requires a wide variety of novel and very fast ever-changing expertise specially for new technologies.
2. This field proposes a high demand for new skills and indicates consists of large competence gaps.

3. The total number of active enterprises in the computer science area is promising and therefore, the research results and findings can be directly applied into the market.
4. There is promising large volume of scientific competence datasets (bibliographic data) freely available in this area which facilitates test and evaluation of scalable analytics.

Participants of the survey were academic and industrial experts mainly from Germany, Belgium, Denmark, US, Greece, Slovenia, Spain and UK. This geographical distribution of the participants considers regional factors, specially European and US differences as well. The main goal of the survey, as stated earlier, was to identify the weights and importance of required competences in the CKR model for the target case study. Prioritization of the competences in this survey was based on the HCV method as described earlier. The questionnaire has been distributed in both German and English languages using the SoSci Survey[5] platform which provides free license for researchers and scientific activities. The survey results have been exported as *CSV* format for further studies and analyses.

With respect to the total number of distributed invitations, a success rate of participation was about 7.6%. In addition to prioritization of competences from CKR model to the academic computer science career, further questions have been asked from participants. Those questions and their results have been discussed in the following. Participants are grouped into HR experts from enterprises (named as experts group) and academics (named as academics group) in the computer science area. In total, experts group consists of 56% of participants and the rest is academics group (i.e. 44%). Further details of participants are showed in Figure 3.1 on the page 55. The final result of this survey is summarized as a RCK matrix of the target case study and is used in the future chapters.

The prioritization of l_1 categories, l_2 subcategories and l_3 sub-subcategories is based on HCV Method. This method is discussed once before in section 3.2 on the page 61 and also uses the proposed hierarchy in Figure 3.2 on page 56. The weights are integer values between 1..10 and the sum of weight in each step is equal to 10 ($\sum_{i=1}^{4} w_i = 10$ in the equation 3.1). Professional competences category (C_1) received the highest weight ($w_1 = 3.3$) in the level l_1. Accordingly, innovative competences is the next important required competences in the level l_1 ($w_2 = 2.4$). Personal (c_3) and social (c_4) competences are equally important in this level ($w_3 = 2.1$ and $w_4 = 2.2$). These are defined as required competence weights, but should be separated for aforementioned three CA types.

It should be stressed that all achieved results through this survey study are for self-assessment method. The weights of two other methods for all competences in the CKR model have been defined through literature review and experts feedback. The final result consisting all assessment types are summarized in table 3.2. In the level l_2 competence subcategories, Job-related Skills ($w_{1.3} = 3.5$),

[5]The SoSci Survey Platform, visited in January 2016 via `www.soscisurvey.de`.

Creativity and Holistic Thinking ($w_{2.1} = 3.3$), Attitude ($w_{3.4} = 3.3$) and Teamwork ($w_{4.1} = 2.8$) are listed as the most important required competences. As it is clear from this prioritization in the level l_2, all level l_1 categories are involved in the top 4 competence subcategories. In addition, job related competences such as programming language knowledge or data analysis expertise and familiarities with required technologies in this regard are the most important required competences.

	w_1	w_2	w_3	w_4	$w_{1.1}$	$w_{1.2}$	$w_{1.3}$	$w_{1.4}$	$w_{2.1}$	$w_{2.1}$	$w_{2.3}$	$w_{2.4}$	$w_{3.1}$	$w_{3.2}$	$w_{3.3}$	$w_{3.4}$	$w_{4.1}$	$w_{4.2}$	$w_{4.3}$	$w_{4.4}$	...
360-degree	5.4	1.7	1.5	1.4	3	2	1.1	3.9	1.9	2.7	1.9	3.5	5.2	1.4	1.4	2	3	2.7	3	1.3	...
self-assessment	3.3	2.4	2.1	2.2	1.8	2.5	3.5	2.2	3.3	1.7	2.6	2.4	1.3	2.7	2.7	3.3	2.8	2.7	1.9	2.6	...
Web-data	1.8	1.1	4.3	2.8	2.1	2.8	2.2	2.9	2.8	1.9	2.9	2.4	4.9	1.3	1.3	2.6	2.1	3.1	2.1	2.7	...

Table 3.2: Weighting of required CK for computer science career according to the CKR model as well as identifying an importance of assessment types in this domain, achieved results through survey study of this research (RCK matrix)

3.4 Conclusion of the Chapter

An overview of the assessment and profiling methods as well as general CK reference model called CKR model is discussed in this chapter. The terms CK and person-job-fit have been clearly defined in this chapter. The CK is being referred as job knowledge in some literature with exactly the same meaning and goal. A person-job-fit process contributes to the RQ 1 (Skill mismatch) which is well-known in the HRM area. This process depends on the assessment results and profiling of the talents' competences as well as profiling of required competences identified as competence gaps. As soon as competence gaps are identified and acquired competences are profiled, mapping of them can be efficiently handled by means of machine learning and clustering algorithms. The goal of person-job-fit algorithm is mainly to classify people in at least three groups of overqualified, best-fit and under-qualified ones. As stated earlier, the general structure and usefulness of the CKR model has been evaluated through survey study.

Three different assessment and job knowledge discovery methods have been defined in this chapter: (1) Multi-source assessment, (2) Self-assessment, and (3) Data retrieval from the web and digital sources. In the multi-source assessment, colleagues and immediate work-circle of a talent provide their opinion about questioned competences of a talent. Their replies to these questions are Likert-scales [Likert, 1932] like "Strongly disagree", "Disagree", "Neither agree nor disagree", "Agree" and "Strongly agree". The assessment is handled through 360-degree feedback method. In addition to collecting the competence data from other colleagues, individuals should assess their job knowledge themselves through self-assessment. The self-assessment method is like participating in on-line tests or exams. The job knowledge data retrieval from digital sources is discussed in chapter 4.

Identification of innovative and personal competence data is not being retrieved from the web and digital sources in this work. In addition, organizational issues of holding an on-line test in the self-assessment are not addressed. These issues are defined as future work in section 7.2. All used and defined mathematical formulas and symbols in the theory of CK profiling in section 3.2 are summarized in Table 3.1. The importance of each assessment method as well as weights of all categories in the CKR model in the computer science academic career have been defined through survey analysis and are summarized in section 3.3.

Chapter 4

Mathematical Modeling, Interpretation and Regeneration of CK Data

»Try not to become a man of success, but rather try to become a man of value.«

– Albert Einstein

Big data facilitates an efficient and productive Talent Analytics (TAs) and improves scalable value creation and Career Knowledge (CK) discovery from large scale disparate HR data. Associated difficulties and challenge in this regard have been discussed in section 1.2 as IC 3 (big data & TA). Consequently, any scientific solution and algorithm should be evaluated and analyzed through integration with real big datasets. From ethical perspective as well as being able to prepare large volumes of the data, it is difficult to retrieve enough and qualitative HR data from free and open sources such as web based databases or any enterprise. Therefore, an artificial data has to be regenerated based on available datasets with 200 employee data (see RQ 2 (statistical distribution) and IC 4 (CK retrieval from the web) in section 1.2) in order to test and verify developed big data algorithms.

Due to ethical issues associated with employee's data, it was not possible to acquire real large scale employee data. The small collection of anonymized 200 employee data has been collected in cooperation with industrial partners. statistical analysis of this dataset results in the regeneration of equivalent big HR data for the evaluation phase. Statistically regenerated big HR data from original datasets consists of 15 million talent data which is sufficient for evaluation of the proposed methods. As discussed before, the real small dataset has been retrieved from industrial partners in a form of three job applicant groups: (1) under-qualified applicants, (2) best fit candidates and (3) overqualified applicants. Such a scenario of considering primary data into three groups was recommended from domain experts, due to the fact that it fits to the real world recruitment practices as well.

It is mandatory to test proposed Hadoop and MR based methods with really big data. Otherwise in cannot be ensured whether proposed approach is correct and successful and working as expected in practice. But the problem was that real test data was not existing and therefore the Hadoop and MR based implementations had to be tested with artificial data. However, any arbitrary artificial data would

not work since proposed method is going to practically solve HR competence data problem in real life. Therefore, it was important to have a set of test data which represents real world situation and data. In this regard, an analysis, formulation and regeneration of the test data based on the real data is a key requirement and innovation of this work.

How the data is being interpreted and clustered is discussed in Section 4.1). According to the CKR model, the CK data in the third level provides 64 dimensions. Therefore, clustering and interpretation of such data with huge dimensions is difficult. The statistical analysis and finding the distribution of original anonymized HR data has been covered in Section 4.2. Based on the findings in this section, regenerated datasets is based on the uniform distribution of original data. In addition to regenerated data, data streaming from social networks such as Twitter is discussed in Section 4.3. Due to the fact that a case study of this work is computer science career, the bibliographic data from DBLP and ArnetMiner have been retrieved and integrated from the web.

4.1 Clustering of CK Data

Data clustering is an unsupervised method for classification of objects with similar properties. The objects are categorized into different groups (clusters) in which the members of one group are quite similar and simultaneously are dissimilar with the members of other groups [Gan et al., 2007; Xu and Wunsch, 2008]. In an unsupervised method, the final clustering results are not being affected by initial distribution of clusters. In supervised clustering, all clusters and their spectral properties should be known before clustering. Unsupervised classification methods are useful, when there is not any preexisting training phase. The majority of the clustering methods is typically categorized to two kinds of partitioning and hierarchical methods.

In partitioning methods, the aim is to partition n elements into k groups ($k \leq n$) provided that each group is non-empty and each element solely belongs to just one group provided that the members of one cluster are more similar to each others comparing to members of other clusters. To this aim, a clustering criterion such as square error is employed. The value of k is often provided by the user, although not every value of k results in natural or conceptually correct clusters [Kaufman and Rousseeuw, 2009]. Therefore, typically, different values of k are tried out and the most meaningful value for k which shows best characteristics or interpretation is then selected. Partitioning methods aim to find a good partition in which similar or closely related elements are grouped as one cluster and members of other clusters have less similarity with them.

In hierarchical methods, the aim is to build a binary tree of hierarchies in which the sub-trees of each nodes are joined, i.e. considered to be a cluster at that level, based on the similarities of their sub-trees. Specifying the number of clusters is not required in hierarchical methods, since the number of nodes in each level

implicitly convey it. Two kinds of hierarchical clustering algorithms are available: (1) top-down and (2) bottom-up. In the first approach, all data are considered to be one cluster in the beginning and are split recursively until individual elements are reached. Contrary, in the bottom-up approach each element is regarded as a singleton cluster and then each pairs of them are recursively agglomerated until one cluster, which consists of all elements, is reached [Manning et al., 2008].

Considering n as the number of elements to be clustered, hierarchical clustering methods form all possible clusters of size k ($1 \leq k \leq n$) in one run. Therefore, it might be argued that one does not need the partitioning algorithms since all k-clusters are found. Actually this is not the case since forming all clusters in one run does not necessarily yield the best possible clusters as once a decision for split or agglomeration is made in hierarchical methods. It can never be either changed or improved [Kaufman and Rousseeuw, 2009]. In contrast, the aim in partitioning methods was to find the best partitioning in order to categorize n elements into k groups.

Consider n objects which are characterized by p features. They can be represented by an $n \times p$ matrix (Eq. 4.1) in which rows are objects and columns are variables and $x_{i,f}$ represents the f-th feature of the i-th object. This representation of elements fits very well to the mathematical representation of CK in the CKR model discussed in section 3.2. In the CKR model, an object is a talent which is represented by variables (e.g. CK values that achieved through assessments for categories, sub-categories and sub-sub-categories in the model). Objects belonging to one cluster indicate talents with similar competitiveness to specific job position in the case study of this research. Similarly, clusters indicate for instance a group of under-qualified or over-qualified people.

To form the clusters of similar talents, one needs to define the dissimilarity, or equivalently similarity, concept between pairs of talents. Dissimilarities between talents of i and j are non-negative numbers of $d_{i,j}$ which are small when objects are near or similar to each other and get bigger when talents are far or more dissimilar to each other [Kaufman and Rousseeuw, 2009]. Typically, mathematical distances such as Euclidean, Manhattan or Minkowski are used to measure dissimilarity. The measured dissimilarities are then presented by a $n \times n$ matrix (Eq. 4.2) which is symmetric, i.e. $d_{i,j} = d_{j,i}$, and has zeros on its diagonal (each talent's dissimilarity to himself is zero). In this research, the Euclidean distance is used to measure dissimilarities between competences.

$$\begin{bmatrix} x_{1,1} & \dots & x_{1,f} & \dots & x_{1,p} \\ \vdots & & \vdots & & \vdots \\ x_{i,1} & \dots & x_{i,f} & \dots & x_{i,p} \\ \vdots & & \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,f} & \dots & x_{n,p} \end{bmatrix}_{n \times p} \tag{4.1}$$

$$\begin{bmatrix} 0 & d_{1,2} & d_{1,3} & \dots & d_{1,n} \\ d_{2,1} & 0 & d_{2,3} & \dots & d_{2,n} \\ d_{3,1} & d_{3,2} & 0 & \dots & d_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n,1} & d_{n,2} & d_{n,3} & \dots & 0 \end{bmatrix}_{n \times n} \tag{4.2}$$

As discussed earlier, in partitional clustering, the number of clusters,i.e. k, is typically specified. In addition, the intermediate solutions are tested against evaluation criteria to see if the clustering results are satisfactorily good and accept the results as final clustering solution. Although the partitional clustering algorithms find a k-cluster within the given data by evaluating given criterion, the result might not be naturally or conceptually correct. In theory, finding the solution of a clustering problem is not difficult, as one can test all partitioning of the data into k clusters and select the one which optimizes the given criterion [Jain and Dubes, 1988]. To this aim, the first difficulty is the suitable mathematical presentation of the criterion which might be quite complex and is dependent on the nature of the given data and its interpretation.

The next difficulty is the exponential growth in the number of possible ways of partitioning a dataset with n elements into k clusters which is so expensive even for relatively small numbers of n and k. Let $S(n,k)$ denote the number of clustering of a dataset of length n into k clusters which principally indicates the number of partitioning a set into nonempty subsets in which the orders of elements in partitions or the order of partitions themselves are unimportant. Assuming that all clusters of length $n-1$ have already been listed, a clustering for n objects can be obtained in either of the following ways [Jain and Dubes, 1988]:

1. The n-th object can be added as a singleton cluster to the list with exactly $k-1$ clusters.
2. The n-th object can be added to one cluster member of the list with exactly k clusters.

Therefore, with the above description $S(n,k)$ can be written as the following difference equation:

$$S(n,k) = S(n-1,k-1) + k\,S(n-1,k) \tag{4.3}$$

in which

$$S(n,1) = 1,\ S(n,n) = 1,\ S(n,k) = 0; k > n$$

are the boundary conditions. The solution of (Eq. 4.3) requires that values of $\{S(j,p)\}$ are known for the set of $\{(j,p) : 1 \le j \le n-2,\ 1 \le p \le k\}$. The solutions of (Eq. 4.3) are called Stirling numbers of the second kind [Gradshteyn and Ryzhik, 2007; Jensen, 1969] and are given by:

$$S(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\binom{k}{i}i^n \tag{4.4}$$

in which $\binom{k}{i} = \frac{k!}{i!\,(k-i)!}$ is the binomial coefficient. Moreover, the number of partitioning of n elements into non-empty subsets is called a Bell number and is denoted by B_n [Weisstein, 2015; Conway and Guy, 2012]. Therefore,

$$B_n = \sum_{k=0}^{n} S(n,k) \tag{4.5}$$

Based on (Eq. 4.4) and (Eq. 4.5), it is clear that the number of partitioning of a set of length n, which should be investigated for an optimum cluster, is growing exponentially. For instance, due to the fact that k is not known in the case study of this research, one needs to investigate $B_{200} \approx 6.24748 \times 10^{275}$ partitions for a set of $n = 200$ employees which one of all of their possible partitioning is the cluster of interest. It is quite enormous even for such small n. As briefly stated before, finding an optimum cluster is associated with a criterion which should be met. Such criteria are dependent on the application domain and nature of the data as well as the aim of clustering [Jain and Dubes, 1988; Tan et al., 2005].

In partitional clustering methods, the criterion is to form clusters in which elements in one cluster are more similar to each other than the elements of other clusters. Moreover, in this way, each cluster can be presented by a prototype which represents that cluster. Two of the most prominent prototype-based clustering techniques are k-means and k-medoids clustering algorithms [Kaufman and Rousseeuw, 2009]. These two algorithms try to form the clusters by investigating the most probable partitions which contributes to the criterion, avoiding to check all possible partitions. K-means is used for clustering of for instance 200 talents into k clusters. The *Silhouette* is a useful tool to determine the value of k in the *k-means* and *k-medoids* algorithms. Both of them minimize the squared error. The Silhouette method is useful to find the correct number of clustering, i.e. K, in the clustering algorithm. Clustering algorithms are heuristic and the initial choice of the cluster centers affect the results and the whole performance and the final results. Therefore, to check the adequacy of a clustering result, one can use the silhouette approach which graphically assesses the goodness of a clustering result. How silhouette method works is discussed in the following as the existing dataset is investigates in the next sections.

In k-means algorithm, the aim is to define a prototype or a central point in the data which is referred to as the *centroid*. The centroid is typically the mean of the elements of one cluster in multidimensional space. Such a centroid is not necessarily accompanies to the original types of the data points, e.g. if ordinal or categorical data are going to be clustered the concept of the centroid is not typically useful in practice. In contrast k-medoids clustering algorithms regard the prototype as a central point in the original data which is most representative for its cluster, thus the medoid is one of the original data points by definition and the k-medoids algorithms are more practical [Kaufman and Rousseeuw, 2009; Jain and Dubes, 1988; Tan et al., 2005].

In k-means algorithm, k initial centroids are selected where k is the number of clusters which is specified by the user. The choice of initial centroids are typically

done randomly but also some other strategies are used in practice [Tan et al., 2005]. In the next step each point is assigned to the closest centroid and the set of all points which are assigned to one centroid forms one cluster around that centroid. After such assignments, the centroids of clusters are updated and then the points are reassigned to the newly updated centroids. The procedure is then repeated until either the centroids do not change or equivalently there is no change in the set of the points of each cluster [Tan et al., 2005]. The choice of the initial centroids is reported to have effects of the resulting clusters as the algorithm will reach a different local optimum point with each choice. Moreover, the algorithm is also sensitive to outliers.

In k-medoids algorithm the aim is to form the clusters around those data points which are more representative for their clusters. The k-medoids algorithm can be used for both continuous and discrete data specially when there is categorical or nominal variables are present. Since in the case study of the talent analytics, competence measurements are discrete values, the k-medoid algorithm is used. In this regard, the PAM (Partitioning Around Medoids) algorithm which is presented in [Kaufman and Rousseeuw, 2009] is introduced. One may argue that K-Means may make sense of it due to the fact that they are not categorical values. But the fact is that in addition to clustering of discrete values, the most competent person (talent) in each cluster should be also found which best represents its cluster he belongs to.

Moreover, let U be an $n \times k$ matrix whose (i, j) element is $u_j(x_i)$, that is the membership coefficient. The u_{ij} coefficients are either 1 or 0 which respectively states that x_i belongs to cluster C_j or not. If $u_{ij} = 1$ then $u_{il} = 0; l \neq j$ indicating that x_i solely belongs to C_j and not other clusters. In short these two conditions are expressed by:

$$u_{ij} \in \{0, 1\}, \; j = 1, \ldots, k \tag{4.6}$$

and

$$\sum_{j=1}^{k} u_{ij} = 1 \tag{4.7}$$

Assume that Θ denotes the set of medoids for all clusters and I_Θ is the set of their indices in the set of initial data points of $X = \{x_1, x_2, \ldots, x_n\}$. Moreover, $I_{X-\Theta}$ is the set of indices in X which are not medoids. The following cost function can be used to assess the quality of the a clustering using Θ as the set of medoids [Theodoridis and Koutroumbas, 2009; Theodoridis et al., 2010]:

$$J(\Theta, U) = \sum_{i \in I_{X-\Theta}} \sum_{j \in I_\Theta} u_{ij} d(x_i, x_j) \tag{4.8}$$

in which

$$u_{ij} = \begin{cases} 1 & \text{, if } d(x_i, x_j) = \min\limits_{q \in I_\Theta} d(x_i, x_q) \\ 0 & \text{, otherwise} \end{cases} \quad i = 1, 2, \ldots, n \tag{4.9}$$

Minimizing (Eq. 4.8) is equivalent of finding a set of medoids Θ which best represents set of data points X. The aim of the PAM algorithm is to minimize $J(\Theta, U)$ provided that medoids are themselves members of X.

To present the PAM algorithm, few more concepts are needed. Assuming two sets of medoids Θ and Θ' each with m elements, they are being called as *neighbors* if they share $m-1$ elements. This way, the number of elements for $\Theta \subset X$ with n elements is $n(n-m)$. Moreover, Θ_{ij} denotes the neighbor of Θ which its $x_i, i \in I_\Theta$ element is replaced by $x_j, j \in I_{\theta - X}$. The ΔJ_{ij} denotes the difference in the cost function when instead of Θ, Θ_{ij} is employed, i.e. $\Delta J_{ij} = J(\Theta_{ij}, U_{ij}) - J(\Theta, U)$.

In the PAM algorithm, first the set of medoids Θ are initialized. Typically a random selection of points in X is used as initial values for Θ. For all $n(n-m)$ neighbors of the set Θ, i.e. Θ_{ij}; $i \in I_\Theta, j \in I_{X-\Theta}$, PAM selects Θ_{qr}; $q \in I_\Theta, r \in I_{X-\Theta}$ with $\Delta J_{qr} = \min_{i,j} \Delta J_{ij}$. It means that PAM selects q,r such that the difference in quality is minimal. If $\Delta J_{qr} < 0$ then replacing medoids of x_i and x_j contributes to minimizing (Eq. 4.8) and thus Θ is replaced by Θ_{qr} and the procedure is repeated. In the case that $\Delta J_{qr} \geq 0$, a local minimum has reached and the algorithm stops reporting the optimum value found for Θ. Using the optimum Θ, all elements of $x \in X - \Theta$ are then assigned to their nearest medoid.

To compute ΔJ_{ij} which was the difference in the cost function of J by replacing $x_i \in \Theta$ by $x_j \in X - \Theta$ in Θ, the ΔJ_{ij} is written as:

$$\Delta J_{ij} = \sum_{h \in I_{X-\Theta}} C_{hij} \tag{4.10}$$

in which C_{hij} is the difference in the cost function when all $x_h \in X - \Theta$ are moved from its old cluster to a new one as the result of replacing x_i by x_j. To compute C_{hij} the following cases might happen:

1. x_h belongs to cluster presented by x_i. Let $x_{h_2} \in \Theta$ denotes the second closest to x_h representative.

   (a) If $d(x_h, x_j) \geq d(x_h, x_{h_2})$ then by replacing x_i by x_j in Θ, x_h will be represented by x_{h_2} and therefore:

   $$C_{hij} = d(x_h, x_{h_2}) - d(x_h, x_i) \geq 0$$

   (b) If $d(x_h, x_j) \leq d(x_h, x_{h_2})$ then by replacing x_i by x_j in Θ, x_h will be represented by x_j and therefore:

   $$C_{hij} = d(x_h, x_j) - d(x_h, x_i)$$

   In this case C_{hij} can be either negative, zero or positive.

2. x_h does not belong to the cluster presented by x_i. Let x_{h_1} be the closest to x_h medoid.

(a) If $d(x_h, x_{h_1}) \leq d(x_h, x_j)$, then x_h will still be presented by x_{h_1} and therefore:

$$C_{hij} = 0$$

(b) If $d(x_h, x_{h_1}) > d(x_h, x_j)$, then:

$$C_{hij} = d(x_h, x_j) - d(x_h, x_{h_1}) < 0$$

Before closing this part, it should be noted that although the PAM algorithm is widely used in practice, other k-medoids algorithms are also available e.g. see [Park et al., 2006; Park and Jun, 2009].

As stated earlier, finding the correct number of cluster, i.e. k, in k-means or k-medoids algorithm is not an easy task and typically different values of k are tried out. Even worse, almost any clustering algorithm will find a cluster within the given data even though no real natural structure might be present [Tan et al., 2005]. Therefore it is necessary that the data and the computed clusters be investigated thoroughly. In this regard, it is necessary to see if there is non-random structure in the data, determine the correct number of clusters to be computed and to check if the results are compatible.

For the data in the Euclidean space one approach to test the quality of a clustering is to investigate the sum of squared errors (SSE) which is suitable for k-means and k-medoids methods [Tan et al., 2005]. In SSE, the error in the assignment of each data point to its cluster representative, i.e. the associated medoid, is computed and the sum of all these errors is considered as SSE. Comparing two different sets of clusters which are obtained by two runs of a k-medoids algorithm can be done using SSE. The cluster set which has less SSE is considered to be superior since the lower value of SSE indicates that the clusters' representatives, i.e. medoids, are better representing that clustering set. The SSE can be defined as:

$$\text{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} d\left(c_i, x\right)^2 \tag{4.11}$$

in which k is the number of clusters, C_i is the i-th cluster and c_i is the medoid of C_i.

To check the suitability of a clustering set, a very popular method is the silhouettes method [Rousseeuw, 1987]. Silhouettes graphically assess the quality of a clustering set by forming a silhouette for each cluster. A silhouette represents points in its cluster indicating which points lay well in the cluster and which ones are not properly classified. Silhouettes can be applied to the results of different clustering techniques as they require just the resulting clustering set of the algorithm as well as the proximities between objects. To compute silhouettes, for each object or point i in the dataset, $s(i)$ form the dissimilarities (see Eq-(Eq. 4.2)) is computed.

Let A be the cluster in which object i is assigned to and let C be any cluster which is different from A. The $a(i)$ denotes the average of dissimilarity of i to

other members of A provided that A has other members than i, formally:

$$a(i) = \text{average dissimilarity of object } i \text{ to all other objects of } A$$

The $d(i, C)$ denotes the average dissimilarity of i to members of C ($C \neq A$):

$$d(i, C) = \text{average dissimilarity of object } i \text{ to all members of } C$$

Let $b(i)$ be the smallest of $d(i, C)$ which can be obtained for all $C \neq A$, i.e. :

$$b(i) = \min_{C \neq A} d(i, C)$$

The cluster in computation of $b(i)$ which delivers the minimum for $d(i, C)$ is denoted by B and is called the *neighbor* of object i. The cluster B is the second best choice for i-th object. I other words, if A is disregard, then B is the best candidate cluster for i. Since B can be computed when there is at least two clusters, the silhouette method is applicable to $k \geq 2$ clusters. The silhouette of i, i.e. $s(i)$ can be computed by:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & \text{if } a(i) > b(i) \end{cases} \tag{4.12}$$

which can be written more compactly as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4.13}$$

In the case that A is a singleton cluster, it is not clear how $a(i)$ is computed therefore in this case $s(i) = 0$ is set. Considering (Eq. 4.12), it is clear that for each i:

$$-1 \leq s(i) \leq 1$$

When $s(i)$ is large, i.e. close to 1, (Eq. 4.12) indicates that *within dissimilarity* of $a(i)$ is much smaller than *between dissimilarity* of $b(i)$. Thus, it can be concluded that i-th object is *well clustered* and with little doubt i is assigned to a good cluster since the second best choice for i, i.e. cluster B, is not nearly as close as the actual choice of A. Conversely when $s(i)$ is close to -1, then $a(i)$ is much larger than $b(i)$. Therefore, i is on average much closer to B than A indicating that i is *misclassified* and should most likely be assigned to B. When $s(i)$ is close to zero, then $a(i)$ and $b(i)$ are approximately equal and its is not clear that object i should be assigned to either A or B. This is the *intermediate case* and the i-th object is almost equally far from both clusters.

To check the suitability of a cluster C, *cluster average silhouette width* (CASW) can be used which is defined by [Rousseeuw, 1987]:

$$\text{CASW}(C) = \frac{1}{|C|} \sum_{i \in C} s(i) \tag{4.14}$$

in which $|C|$ is the number of objects in C. Similarly, considering a clustering set of length k for a dataset of X with n objects, the *data average silhouette width* (DASW) is defined by:

$$\text{DASW} = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{4.15}$$

The DASW of k clusters is also denoted by $\bar{s}(k)$ and can be used to select the best possible value for the number of clusters, i.e. k, in a given dataset. To this end, $\bar{s}(k)$ for all possible values of k is computed in which takes the one which maximizes $\bar{s}(k)$ [Rousseeuw, 1987; Kaufman and Rousseeuw, 2009]. The possible values for clustering a data of length n are $k = 2, 3, \ldots, n-1$. The maximum value of $\bar{s}(k)$ is called the *silhouette coefficient* (SC) and is computed by:

$$\text{SC} = \max_{k} \bar{s}(k) \tag{4.16}$$

In the case study of this research, the sample dataset consists of 64 different competences at the third level in which each competence is measured on a discrete scale in the interval of $I = [1, 10]$ i.e. the competences are represented by integers of $\{1, 2, \ldots, 10\}$. As mentioned earlier, since the sample space consists of 200 talents' real competence data, the possible number of partitioning the dataset can be computed using (Eq. 4.5) which is enormous $B_{200} \approx 6.24748 \times 10^{275}$. In order to find a clusters in the CK dataset of this research, k-medoids algorithm is employed. Since the results of clustering using k-means or k-medoids algorithms is dependent on the selection of the initial points for centroids or medoids, they might produce different results with each run [Kaufman and Rousseeuw, 2009; Tan et al., 2005; Theodoridis and Koutroumbas, 2009; Xu and Wunsch, 2008]. Moreover, any clustering algorithm will generate a cluster set which might not be accurate or conceptually correct. The specification of an appropriate number of clusters in the algorithm is also quite difficult.

The first try of the k-medoid algorithm on the collected dataset for different values of k resulted no good clustering. The reason for that was there are too many dimensions for data points according to the CKR model, as 64 sub-subcategories are captured in the level 3. As stated earlier, the choice of initial values for centroids and medoids respectively in k-means and k-medoids algorithms will affect the final outcome since the algorithm will find a local optimum for the clustering problem. None of the clustering results of CK dataset using different values of k were successful to find a suitable clustering within CK dataset. As an example, Figure 4.1 shows the silhouette coefficients of the data points as well as CASW values for the clustering results of the k-medoids algorithm with $k = 5$ clusters. The case of $k = 5$ is just as an example to describe it in more details, otherwise it has been evaluated for $k = 1$ up to $k = 199$. As shown, CASW values for cluster of $C_1, \ldots, C_5$ as well as DASW are near zero indicating that the results of the k-medoids algorithm is not satisfactory.

To study the performance of the k-medoids algorithms in finding clusters in the CK datasets, its performance is investigated when other initial guesses for the

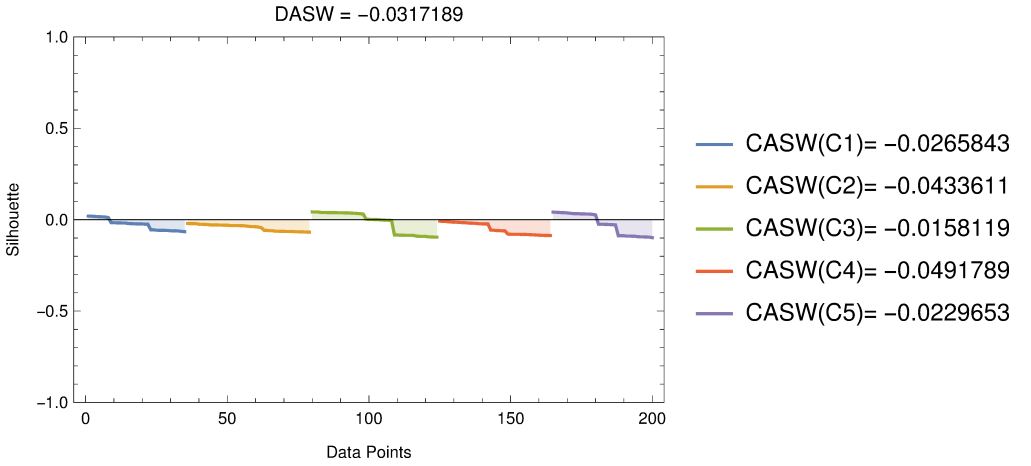


Figure 4.1: The performance analysis of the results of the k-medoids algorithm using the silhouette coefficients ($k = 5$). The data is unsorted in this figure to show that clustering results of CK dataset were not successful to find a suitable clustering within CK dataset. The CASW values for cluster of $C_1, \ldots, C_5$ as well as DASW are near zero indicating that the results of the k-medoids algorithm is not satisfactory. In the next steps, similar methods will be applied to sorted dataset, which is described in the following.

medoids are taken. Figures 4.2 and 4.3 show the average performance of 10 runs of the k-medoids algorithm using different values of initial guesses for $k = 2, \ldots, 199$. Figure 4.2 shows the average of SSE (see Eq-(Eq. 4.11)) for different values of k. As can be seen, the SSE values decrese as k increases suggesting that each point might be a singleton cluster. As it is clear, the SSE is always 0, when each point is a singleton cluster. Additionally as shown in Figure 4.3, the DASW values (see Eq-(Eq. 4.15)) increase as k increases providing more support to this assumption that each data point might be a singleton cluster, however as showed in the following, this assumption is not true.

To find a reasonable clustering for the CK dataset, other properties of CK dataset are studied. Figure 4.4 shows the mean-variance plot of the competences of each employee. As can be seen, data points form three visible patterns in their mean of competences. Figure 4.6 shows the mean-plot of competences for each employee in which employees are sorted based on their mean of their competences. Apparently there are three distinct groups which have means of their competences approximately at 2, 5.5 and 9.

In a first look at Figure 4.6, a sharp move between clusters maybe questioned. But the creation of 2-dimensional (mean-variance) plot makes it more clear to see the behavior and distribution of the data (Figure 4.5). As showed in this plot, for the first cluster (blue dots) there are data with a mean of around 3 and variance of 2.5, meaning that this data is located around 5.5 on average (even more). The same is checked for the second cluster. There are some data with the mean of

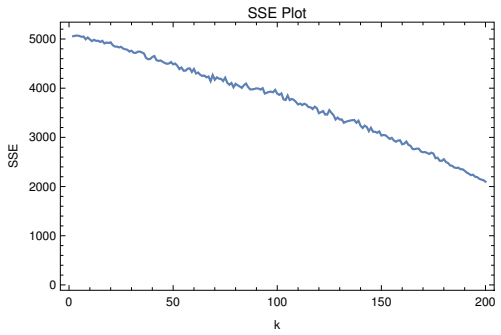


Figure 4.2: Average of SSE for 10 runs of k-medoids algorithm for $k = 2, \ldots, 199$.

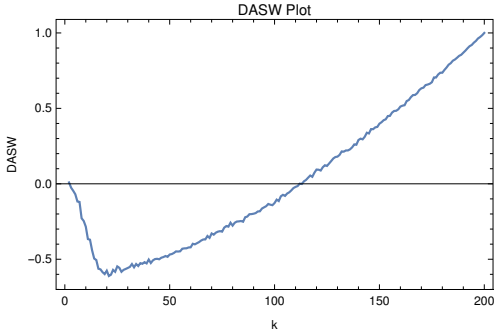


Figure 4.3: Average of DASW for 10 runs of k-medoids algorithm for $k = 2, \ldots, 199$.

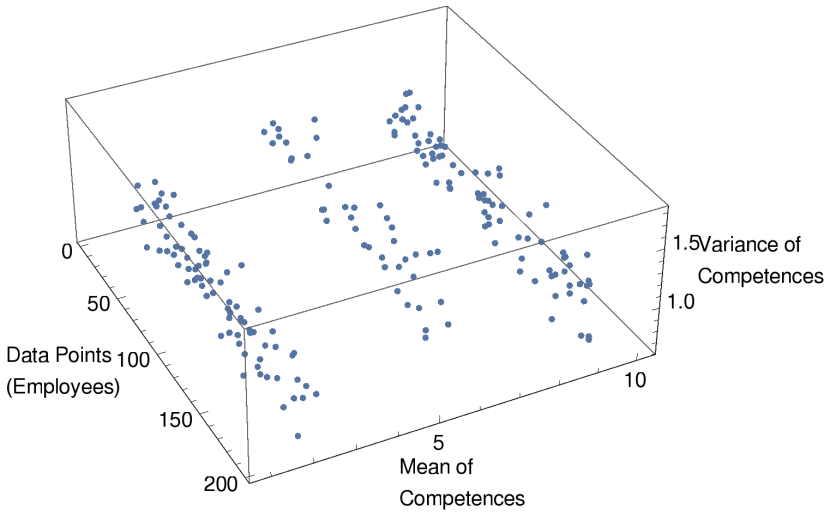


Figure 4.4: Mean-Variance plot of competences for the employees competences.

about 4 and variance of about 1, meaning that the data can be ranged up to 5 on average (even more). This plot shows that the data is already intermixed in higher dimensional space, so why the current plot is too sharp? This is because of computed mean of 64 competences for one person in which visualization of 64 dimensions to see the real behavior of the data is impossible.

To find the reasonable clustering within the CK dataset, the k-medoids algorithm is fed by the sorted list of the dataset in which the data were sorted based on their mean of their competences. In other words, first took the mean of competences for each employee and then sorted the list of employees based on their mean of their competences. Figures 4.7 and 4.8 receptively present the SSE and DASW plots of CK datasets k-medoids algorithm are forced to find $k = 2, \ldots, 20$ clusters. As shown in Figure 4.7 the SSE values drop sharply when $k = 3$ clusters are considered and then increases when more clusters are forced

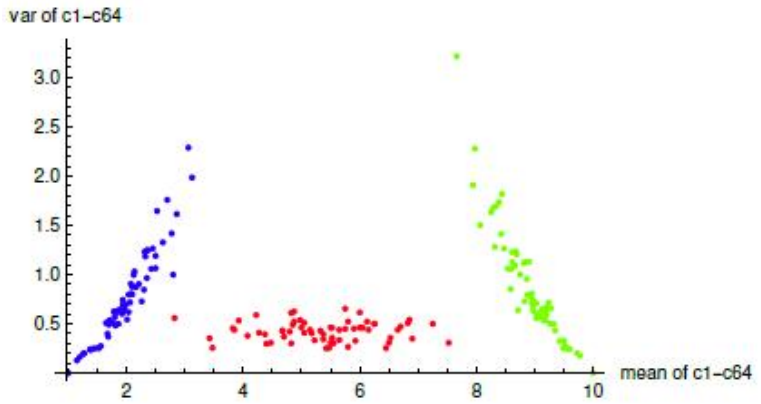


Figure 4.5: 2-dimensional (mean-variance) plot of original data

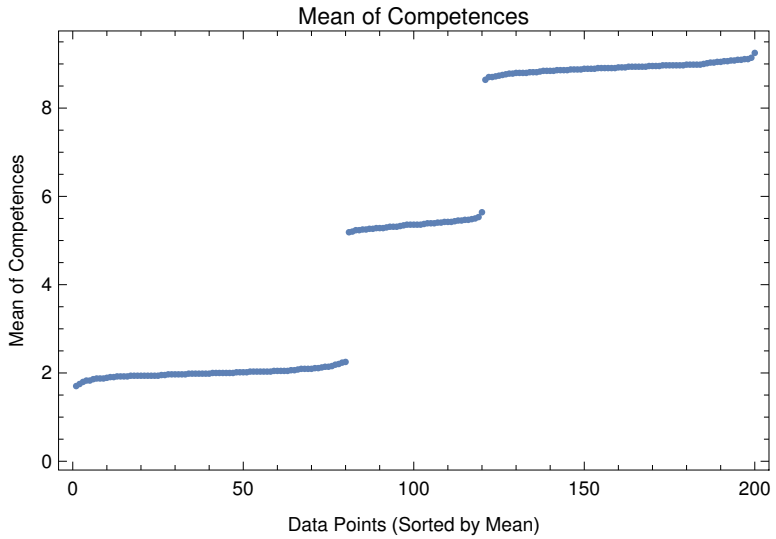


Figure 4.6: Mean-plot of the competences of each employee (Data points are sorted based on their mean).

to be found be the algorithm. Similarly as shown in Figure 4.8, the DASW is at its peak when $k = 3$ clusters are considered and then the DASW values drop considerably when more clusters are to be found. For $k = 3$ clusters, the SSE is computed at 1910.34 which is even lower than 2098.67 when even $k = 199$ clusters were considered as algorithm applied to unsorted data (see Figure 4.2). The same is true for DASW, as it is computed at 0.67365 but was near or below zero when the k-medoids algorithm was applied to unsorted data (see Figure 4.3). The SSE is much smaller because the clustering algorithm is not trapped in the local minimum when the initial cluster centers are blindly chosen. When they are chosen intelligently based on the means, then the local min is escaped which is caused by the random choice of initial cluster centers.

The sorting of the data has effects of the clustering results. As it is mentioned earlier, the total cases of possible clustering to investigate were around $B_{200} \approx 6.24748 \times 10^{275}$, which is not practically possible to investigate all these possibilities. Moreover, SSE and DASW indicators have to be considered to check the adequacy of the clustering results which earlier showed to not working (Figures 4.7 and 4.8) when initial centers were chosen blindly (randomly) by clustering algorithms themselves. As the data is sorted based on their means, the initial choice of cluster centers will be on each of the three categories (categories in means) and therefore in the following iterations the clustering algorithm will find the right cluster (k=3). This is due to big dimensions (64) as well as strong correlations in the data. Therefore, the algorithms have to be assisted with the initial choice of the cluster centers. In fact, the only way they work is to supervise the choice of initial cluster centers (using the means approach).

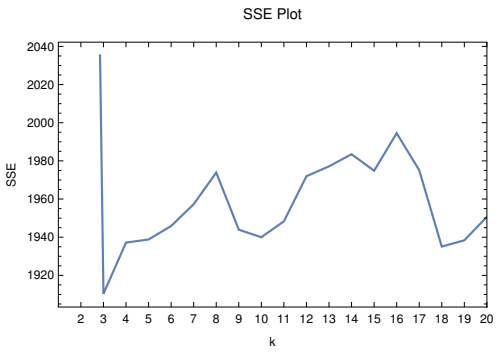


Figure 4.7: SSE plot of k-medoids algorithm applied to mean-sorted data points for $k = 2, \ldots, 20$.

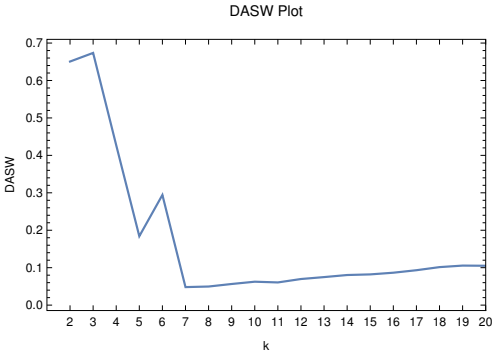


Figure 4.8: DASW plot of k-medoids algorithm applied to mean-sorted data points for $k = 2, \ldots, 20$.

Finally to additionally prove the suitability of $k = 3$ clusters to categorize the competence data, the silhouette plot of the k-medoids clustering is investigated when $k = 2, 3, \ldots, 20$ clusters are employed. Figure 4.9 shows the silhouette values of CK data points as well as the CASW values of the three computed clusters. As shown, all CASE values as well as DASW are near to 1 indicating that the computed clusters when $k = 3$ results suitable clusters and data points are assigned to their correct groups. As a result, the CK data points can be categorized into three clusters appropriately.

One might ask how human competences are related to each other and do they show similar or different behavior. To answer this question, the Pearson's correlation coefficient [Neter et al., 1996; NIST, 2013] is considered (see the Equation (Eq. 4.17)). The Pearson's correlation coefficient between two variables of X and Y are indicated by ρ which takes real values in the interval of $I = [-1, 1]$. The values near to 1 indicate that there is a linear correlation between X and

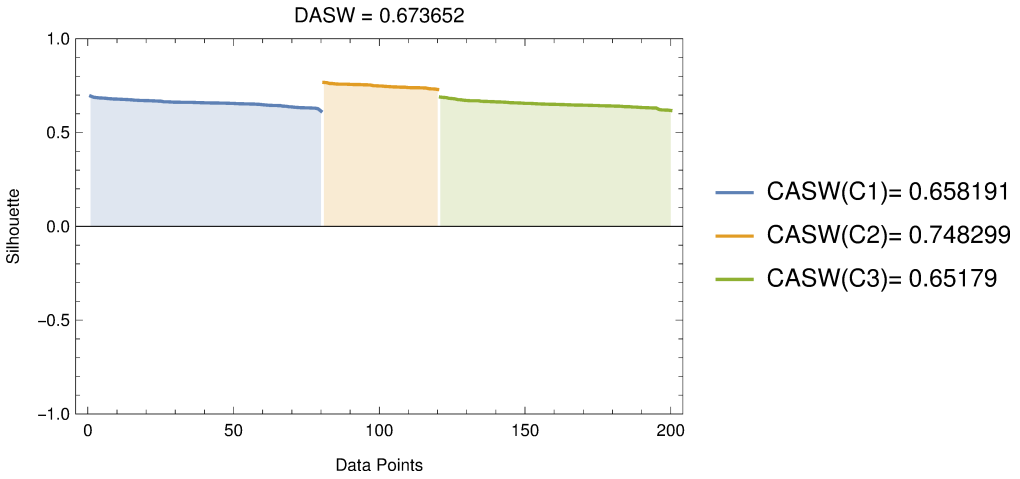


Figure 4.9: Silhouette Value of the CK data and CASW values of the three computed clusters. The data is sorted according to the clusters.

Y as they are located nearly as a line with positive slope and when one of them increases the other one increases as well. The values near -1 indicates that the relationship between X and Y are negative and as one increases the other decreases. Values near zero indicate that there is no linear relationship between X and Y, although there might be some none-linear relationships or no relationship at all.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.17}$$

Figure 4.10 shows the correlation between the four $l1$ competences in the list of 64 competences in the CKR model, namely Professional (C_1), Innovative (C_2), Personal (C_3) and Social (C_4) CK categories. As shown, the Pearson's correlation coefficient for each two pair of these competences are near to 1 (more that 0.9) indicating the these four competences are linearly correlated (around the line of $y = x$) and both increase or decrease simultaneously. As a result, less value of ρ indicates that the competences are less correlated. As an example, in the selected chart, with the growth of C_{23} upto 4, C_{21} remains always 1. This shows that they are less correlated and with the increase in C_{23}, any significant changes in the C_{21} don't happen. Such less correlation is also clear with the light color in a correlation matrix in Figure 4.11. Additionally, the pairwise correlation coefficient between each two kind of competences in the dataset is computed. Furthermore, a correlations between competences in Figure 4.10 can also be seen as an evidence that assessors cannot really distinguish between most of competences in the CKR model. Thus, they provide similar ratings for most of them.

Figure 4.11 shows the correlation matrix of the competences. The intenser colors show higher values for competences while colors with less intensity show

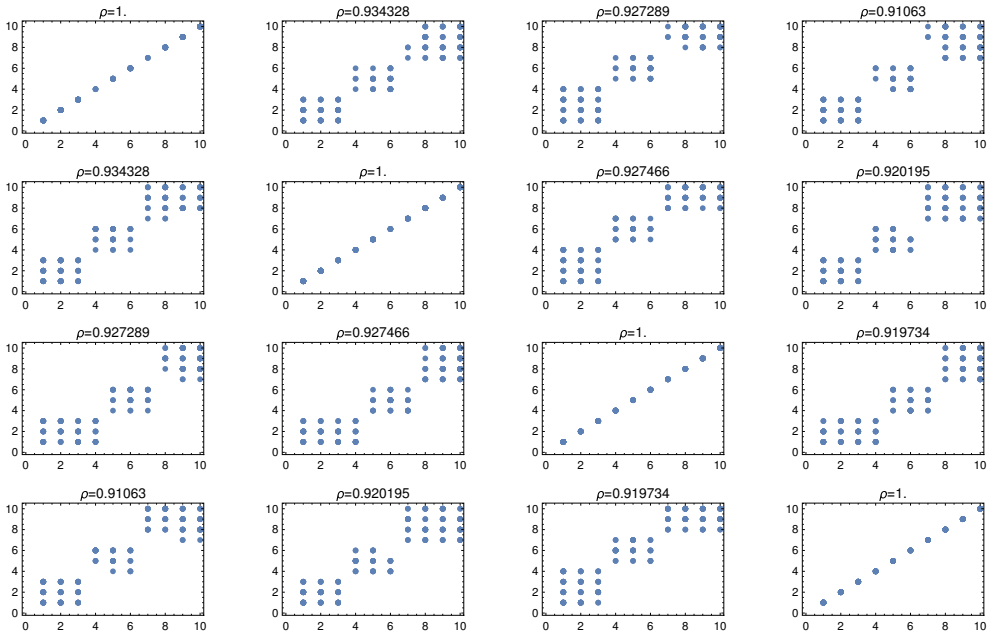


Figure 4.10: Correlation plot of the Professional (C_1), Innovative (C_2), Personal (C_3) and Social (C_4) competences. Each row and column of four Plots represents one competence category, meaning that for instance the first row is Professional Competences category (C_1). Similarly, the first column indicates the Professional Competences category (C_1). As it is clear from this figure each competence category is fully correlated with itself. In this figure, the x-axis of each plot indicates the competence value of its associated row and y-axis shows the competence value of its associated column. Colorful demonstration of correlations between level $l1$ competence categories is showed in Figure 4.11

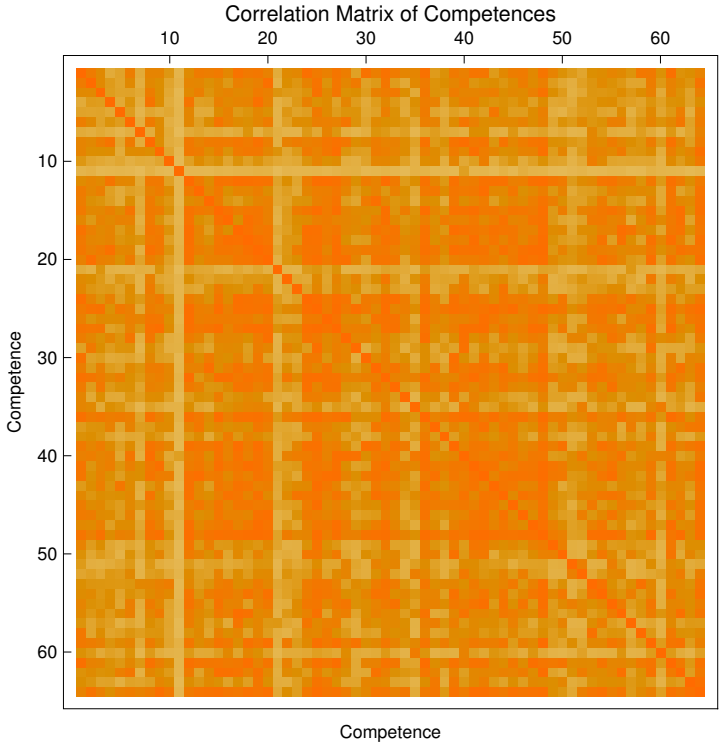


Figure 4.11: Correlation matrix of competences.

lower correlations. The minimum and maximum correlation coefficient are computed at 0.8426 and 0.9539 respectively which shows that all pairwise collection of competences are linearly correlated and increase in one results increase in the other. Moreover, this additionally supports the clustering approach mentioned earlier since it is not possible that one competence increase while the others are decreasing, therefore the clustering strategy to sort the data points based on the mean of the competences is meaningful.

4.2 Mathematical Models and simulation of Competences

As discussed in section 4.1, the clustering algorithm could identify three discrete clusters within the CK data. In order to properly investigate the big data approach of handling competence data (see Chapter 5), there should be reasonable amount of data for evaluating the performance of implemented machinery. Since primary dataset of this research is not big enough (200 employers) to be used for big data implementation, a reasonable amount of the data has to be found or regenerated that truly resemble the properties in the original dataset. In this regard, this section tries to identify a suitable statistical model that can formulate

	Cluster-1	Cluster-2	Cluster-3
No. of data points	80	40	80

Table 4.1: Clustering Information

the competence data in the original sample set. The mathematical models are then used in order to simulate the properties of the original CK data and generate big bulk of artificial competence data how resemble the same statistical property in the original CK dataset.

In the fist step, the results of clustering algorithms discussed earlier are investigated. Table 4.1 shows the amount of data points in each of three clusters. In order to simulate each cluster, the statistical properties of each of 64 competences in each cluster are analyzed with investigation on their histograms. The histogram of the each competence suggested that the data might be uniformly distributed [Johnson et al., 1994a,b] within the range of observed competences. To formally test this hypothesis, each competence dataset of each cluster has been fitted to the uniform distribution. In this regard, the null and alternative hypotheses are defined, i.e. $\mathcal{H}_0$ and $\mathcal{H}_a$, as follows:

$\mathcal{H}_0$: The dataset obeys the uniform distribution.

$\mathcal{H}_a$: The dataset does not obey the uniform distribution.

To fit a distribution and estimate its parameters two commonly used methods are the *method of moments* and the *maximum-likelihood estimation method* (MLE) [Wackerly et al., 2007; Bohm and Zech, 2010]. The method of moments tries to estimate the parameters of the distribution using the observed moments of the sample set. In this method, the population moments are equated to the sample moments in order to solve the resulting equations to find the parameters of the distribution. In the MLE method, parameters of the model is computed by maximizing the logarithm of the likelihood function. These two methods are discussed in [Johnson et al., 1994a,b]. The MLE method is used in this section. The computations are done using the Wolfram Mathematica® 9.0.1 computational engine [Wolfram Research Inc., 2014].

In the analysis, the significance level of 0.05 is taken for the test of initial hypothesis of the research and the Pearson's chi-square test is employed [NIST, 2013]. First, the parameters of the uniform distribution are estimated and then the p-value of the Pearson's chi-square statistic are computed in order to decide whether to reject $\mathcal{H}_0$ in favor of $\mathcal{H}_a$ or not. Figure 4.12 shows the p-value plot of the analysis. In the plot, the colors are coded as black cells which denotes the cases where the p-value in less than supposed significance level, i.e. $\mathcal{H}_0$ was rejected. When the calculated p-value was above the significance level, i.e. $\mathcal{H}_0$ was not rejected, the cell is colored. The more intense the color of the cell, the higher the associated p-value. As shown, the $\mathcal{H}_0$ hypothesis was rejected at 11

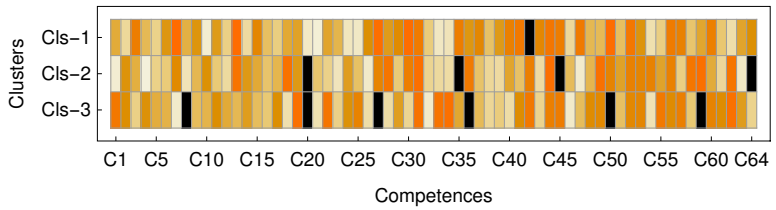


Figure 4.12: p-value plot of the Pearson's chi-Square test at the significance level of 0.05 for the uniform distribution.

out of 192 cases which yields the success rate of 94.27%. If the significance level is changed to 0.01, the success rate will be at 97.92%. Figure 4.13 shows the p-value plot when the significance level is at 0.01. In both cases, the uniform distribution satisfactorily can model the 64 competences of each cluster. The conclusion is that the uniform distribution can be used to simulate enough artificial competence data needed for the test and evaluations of the big data implementation in the Chapter 5. Figure 4.14 shows the histogram of the estimated parameters of the uniform distribution for the three clusters found earlier. The yellow color depicts the estimated first parameter of the uniform distribution while the blue color depicts the estimated second parameter of the distribution.

In order to simulate artificial competence data, the uniform distribution is used. In this regard, random numbers of the uniform distribution should be properly generated. Random number generation methods deal with producing sequences of independent and identically distributed (iid) numbers of the uniform distribution of $\mathrm{U}[0,1]$. Using a simple linear transformation, the generated numbers will follow the general form of the uniform distribution of $\mathrm{U}[a,b]$.

Random number generation methods principally are deterministic programs with finite set of states, including an initial state, (called a seed) and a mapping (transient function) which maps those states to themselves. The states correspond to a finite set of output symbols that the program produces. The role of the transient function is to create the next state based on the previous state [Banks, 1998]. Since the states are finite, the output of the program is also finite and the generator repeats itself with a period. Due to this reason these methods are

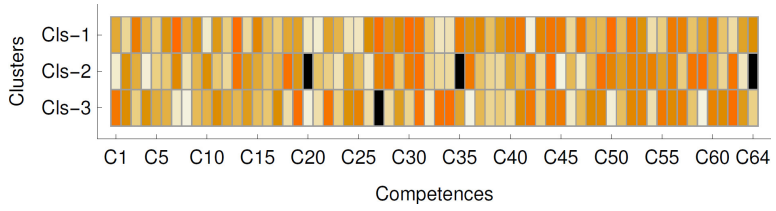


Figure 4.13: p-value plot of the Pearson's chi-Square test at the significance level of 0.01 for the uniform distribution.

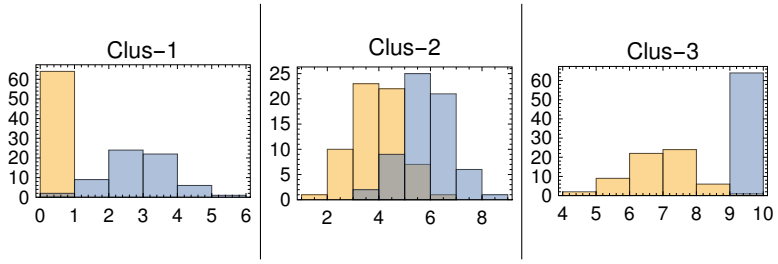


Figure 4.14: Histogram of the estimated parameters of the uniform distribution $\mathrm{U}(\alpha, \beta)$ for each cluster (Yellow: Histograms of α, Blue: Histogram of β). The Histograms are for 64 competences. Te x-axis indicates the value of competence categories and y-axis is competences.

usually referred to as "pseudo random number generators" since they do not really create iid random numbers, but they try to meet the statistical requirements for truly iid uniform random numbers as much as possible. These requirements are, uniformity and independence of the generated numbers as well as greater repetition periods [Banks et al., 2010; Banks, 1998].

Generating truly random numbers are still in the focus of many researchers and in this regard, many random number generators have been introduced so far. Two of the most frequently used random number generators are *linear congruential method* and *combined linear congruential method* [Knuth, 1998]. However, it should be mentioned that more robust methods exist and can be employed as well [L'Ecuyer, 1994].

4.3 Data Streaming and Retrieval from Digital Sources (Web)

In order to obtain the greatest available real computer science academic career competence data, various web-based systems (i.e. data sources) such as Google Scholar, IEEE Xplore, ACM Digital Library, ArXiv, CiteSeer, DBLP and AMiner have been investigated with regard to the suitability and availability of their datasets. The most important competence data in this regard is bibliographic data and information (metrics) about scientists' publications. Based on the feasibility study of stated datasets, the web-based bibliographic data is retrieved from AMiner[1] and DBLP[2]. The DBLP datasets consists mainly meta data about authors and their publications. Furthermore, AMiner provides further information about citations and references. Retrieved DBLP data are in the XML format, but AMiner uses text files to form the information about publications.

[1] `https://aminer.org/billboard/citation`, retrieved 28.07.2015

[2] `http://dblp.uni-trier.de/faq/How+can+I+download+the+whole+dblp+dataset`, retrieved: 28.07.2015

A type of each publication is discovered from its bibtex definitions (e.g. “book” or “article”). Most of the keys provided in the XML representation of a publication from DBLP database are easy to extract and integrate, since they use standard bibliographic terminologies such as “pages”, “year”, “author” and “title”. In addition to the publications, the DBLP XMLs consists of entries about scientists as well. An entry of one specific person is showed in the Listing 4.1.

```
< inproceedings mdate =" 2005 -06 -15 " key =" conf / metmbs / FathiWG04 ">
  <author >Madjid Fathi </ author >
  <author >Ursula Wellen </ author >
  <author >Hamid Garmestani </ author >
  <title >Software Support for Classifications of MRI Images .</ title >
  <pages >499 -502 </ pages >
  <year >2004 </ year >
  <crossref >conf / metmbs /2004 </ crossref >
  <booktitle >METMBS </ booktitle >
  <url >db/ conf / metmbs / metmbs2004 . html # FathiWG04 </ url >
</ inproceedings >

<www mdate =" 2007 -05 -24 " key =" homepages /c/ StefaniaCostache ">
  <author >Stefania Costache </ author >
  <author >Stefania Ghita </ author >
  <title >Home Page </ title >
  <url >http: // www . l3s .de /~ costache /</ url >
</www >
```

Listing 4.1: A sample representation of a publication in the DBLP database

As a summary, the retrieved dataset from the web in this research consists of 2,890,342 publications and 1,533,708 talents data. In total, 2,146,341 of the publications were contained in AMiner’s dataset. The size of resulted DB is 3.95 Giga Bytes. It was clear from the retrieved datasets that it is only a part of real AMiner DB, because many links between publications were missing in the dataset.

The process of streaming social media data using tools such as Twitter streaming API has been tested in the frame of this thesis as it is showed in Figure 4.15 [Bohlouli et al., 2015b]. But this step is not integrated in the data analytics described in chapter 5 and is considered as the future work in section 7.2. In the case streaming the data from Twitter, as a first step a table is created in the HBase database. Using OAuth process, it connects to the Twitter and starts data retrieval of individual tweets. The results of the Twitter streaming are converted to a suitable format and stored in a HBase table. In addition, it is being checked if another Tweet is required to be streamed or not. The determination condition of the streaming is depends on the configuration of already defined information in the Data job. Once the import has been completed, it closes the connection to the Twitter network again and creates an import log (Figure 4.15).

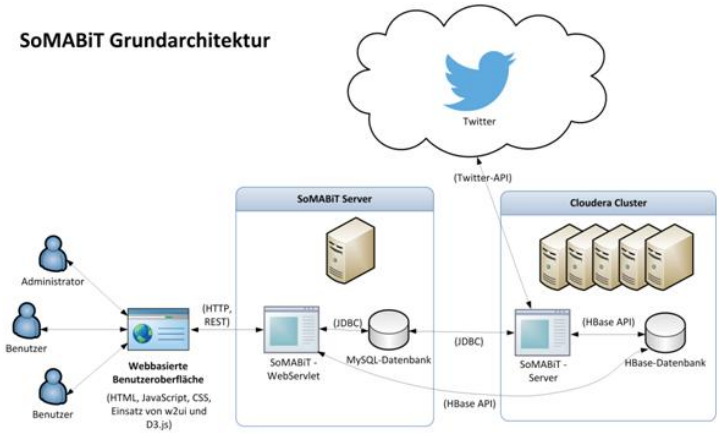


Figure 4.15: Streaming the data from social networks using tools such as Twitter Streaming API [Bohlouli et al., 2015b; Dalter, 2014]

4.4 Conclusion of the Chapter

The main focus of this research is on the matching of large scale talent data to already identified competence gaps in enterprises. To this aim, scalable algorithms have been developed in which should be tested with real big datasets in order to test the scalability and efficiency of them. Such datasets should be real datasets on the one hand that demonstrate real world cases and constitute a large volumes of the data on the other hand that could test scalability of the solution and its efficiency in this regard. Furthermore, the talent and competence data is one of the most sensitive data that cannot be achieved from industries easily, specially in large volumes. The only real dataset which was available from the beginning of this research was an anonymized competence data of 200 employees according to the CKR model as original competence datasets.

A solution towards preparing proper and sufficient dataset is regeneration and simulation of original datasets with the same behavior. To this aim, statistical behavior of the data is analyzed in order to find the best-fit statistical distribution. As a first, the original dataset has been clustered using k-means algorithm and the optimum number of clusters for original data has been achieved. The input data to the k-means algorithm is the level $l3$ competence data with results in grouping of talents with similar level of the CK. The conclusion according to the discussion in Section 4.1 is that the original dataset originates total number of three clusters. The total number of optimum clusters has to be also checked after regeneration of the data in order to ensure that regenerated data holds similar behavior as original dataset.

Besides of scaling up the data using statistical analysis, streaming and data retrieval from digital sources such as web mining and social media streaming delivers huge volumes of the competence data as well. For streaming of academic

computer science career data from web based digital sources, total number of 7 well known systems have been tested, in which only two of them grant an access to the source of their data. These bibliographic datasets are retrieved from DBLP and AMiner. In addition, streaming the data from Twitter has been tested, but not integrated in the system. The use and analysis of the social media streamed data is further discussed as future work in chapter 7. In general, the retrieved dataset consists of about 3 million publications' data in addition to the 1,5 million talent (i.e. computer scientist) data. This amount of publications data is reduced to about 2,1 million records after preprocessing of the data which is discussed in section 5.2.1. The resulted DB has a total size of 3.95 Giga Bytes. This streamed data from digital sources (section 4.3) is integrated to the regenerated data (4.2) in preprocessing of the data (section 5.2.1).

Chapter 5

Scalable Data Analysis and Clustering

» We can have facts without thinking but we cannot have thinking without facts. «

– John Dewey

The main goals of this work as discussed earlier are to (1) assess competences of talents and represent them as TPs (2) match already identified competence gaps named as JPs to the TPs, and (3) provide Competence Development Recommendations (CDRs) for under-qualified job seekers (TPs). Identification of the competence gaps is already discussed in section 3.2 which has formed the RCK matrix, equation (Eq. 3.5). For a scalable matching of TPs with JPs and recommendations based on the competence goals, existence of large scale HR data is assumed in this chapter. This large scale HR data has been achieved from regeneration and scaling up of retrieved and pre-processed small data as discussed in chapter 4. In order to process such large scale data volumes, one has to utilize big data analytics due to the fact that traditional solutions are incapable of processing such large datasets.

Furthermore, providing goal specific training programs in order to improve one or more specific competence(s) improves competitiveness of under-qualified talents. Such goal specific training programs are for instance professional on-the-job-trainings, VET programs, webinars or workshops. To this aim, one needs to first answer to the question of "Who needs what further trainings for which goal?". An assumption here is to match CDPs with specific competence goal(s). Definition and Identification of competence goals using AHP method is discussed in section 5.3.1. The use of MR and Hadoop ecosystem in matching of large scale TPs and JPs supports scalable processing and analytics for large enterprises and job centers. This topic becomes more beneficial when an enterprise uses social network analysis for retrieving and collecting the data.

As a part of input datasets, regenerated 15 million TPs based on statistical analysis of 200 talent data is used in this chapter to test and evaluate the proposed approach. In addition, 75,000 CDPs have been generated using scripts which is 100% artificial data. The data associated with the CK of talents (TPs) in the case study of this research have been retrieved from DBLP and AMiner. In order to define some JPs in the test and evaluation of the results, various job announcements in academic career in the computer science area have been analyzed. CDPs are for instance courses, workshops, seminars, on-the-job-training

programs, VET and any other source that could improve competences of talents. In order to understand details of all discussed algorithms in this chapter, reading section 5.1 is strongly recommended.

5.1 Hybrid Clustering and Matching Approach

A comprehensive and high level overview of the hybrid approach and relationships between different components, profile types as well as scientific algorithms is provided in Figure 5.1. As showed in this figure, three profile types (datasets) have been supposed as inputs: (1) TP, (2) JP, and (3) CDP. Based on these profile types, the approach is divided into three matching problems as of:

1. Matching of TPs with JPs to find the best talent for an already opened job position. This process is defined as "person-job-fit" method and contributes to the RQ 1 (Skill mismatch) challenge discussed in section 1.2.1.
2. Matching of TPs with CDPs for providing recommendations (CDRs) to under-qualified job seekers and increase their competitiveness for future similar JPs. This contributes to the IC 5 (VET recommendations) discussed in section 1.2.3.
3. Matching of JPs with CDPs aiming at job knowledge discovery and identification of relationships between required competences in job descriptions and effects of providing trainings in this regard. This will result in identification and assessment of the most important required trainings for various job categories. This matching method is described as a future work in section 7.2.

The use of assessment methods such as 360-degree feedback and self-assessment described in section 3.2 provide requirements of TPs. Particularly, different fields need various assessment methods as well as assessment metrics for efficiently measure domain specific competences. These domain specific metrics such as SCF or AIS (stated in sections 5.2.2 and 5.2.3) have to be integrated as a part of the TP. Proposed SCF and AIS metrics are associated with job related skills subcategory ($C_{1,3}$) in the CKR model. These metrics aim to assess and evaluate domain specific CK of talents through qualification measures such as citation counts of their scientific publications in a computer science academic career. Due to the importance of such metrics in supporting recruitment decisions with providing detailed knowledge about professional CK of talents, they receive higher weights in the RCK matrix in section 3.2.

It is obvious that proposed metrics are domain specific and cannot be generalized or reused in other domains. Domain experts in other fields and case studies should define similar metrics in their field of expertise in order to get full functionality of the CKR model and matching methods of this work. Extending these metrics to further domains such as nursing or politics contributes to the

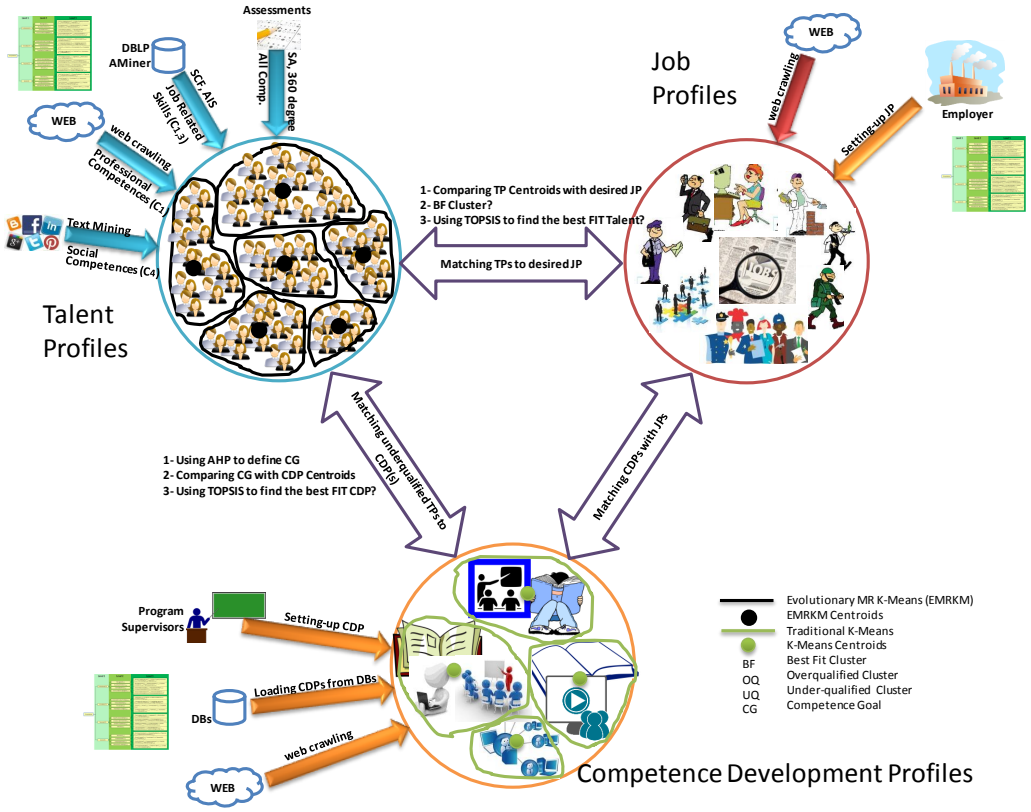


Figure 5.1: A high-level overview of the concept

generalization of the concept which is also discussed as a future work in section 7.2. In addition, developing further data retrieval methods and/or social media streaming techniques could enrich TPs and improve the accuracy of talents' CK data. This may even need utilization of further methods and analytics such as text mining.

Consequently, SCF and AIS metrics from retrieved data as well as assessment results (sections 3.2.1 and 3.2.2) deliver 200 TPs. These 200 TPs have been statistically analyzed and regenerated to 15 million TPs using uniform distribution as discussed in section 4.2. In total, this regenerated 15 million TPs together with associated documents as well as datasets while preprocessing is about 1,5 terabytes. Due to the large volume of TPs and associated analytics, the matching method of TPs and JPs is the most data-intensive parts of the approach which demands efficient scalable algorithms. In this regard, any abstraction or filtering of the data in order to reduce the complexity and volume of analytics will speed-up the matching of TPs with JPs.

Similarly, configuration of the CDPs is also on the basis of the CKR model. Supervisors and people who are responsible for training programs should setup

associated CDPs of their program. As an example, supervisor of a training program (seminar or course) in the field of "Parallel Programming Systems and Models" or "Entrepreneurship and Innovation Management" should clearly define which competences in the CKR model will be affected (and how much) by participating in those specific programs.

Enterprises, employers and job givers should setup description of JPs and define their required competences according to the CKR model. In this regard, the weights of all competences from the CKR model should be defined in the job description which results in the RCK matrix. In general, the most data-intensive part of processing is an analysis and matching of TPs to a desired job. In order to speedup this process, TPs have to be clustered and filtered to reduce the data intensity in order to concentrate on talents with the most close expertise to the desired JP. In this regard and due to the large volume of TP data, the use of MR and big data provided parallelism makes sense of it.

Through utilization of Evolutionary MapReduce K-Means (EMRKM), the 15 million TP data is broken down into different clusters. Any of these clusters consists of talent data with similar competences. It is not clear which cluster contains well-qualified ones (best-fit). To this aim, using Euclidean distance between centroids of each TP cluster and desired JP, the best-fit TP cluster can be identified. As a result, under-qualified clusters can be considered for matching with CDPs towards CDRs and best-fit cluster can be supposed for matching with desired JP towards recruitment decision. The matchings between best-fit TP clusters with desired JP as well as between under-qualified TP clusters with CDP clusters are on the basis of TOPSIS method.

A total number of 75,000 artificial CDP data has been generated using scripts which is not on the basis of any real data. In general, a configuration of the CDP should clearly define which competences will be affected by participation in that training program. Using traditional K-Means, CDPs are clustered, so a cluster of CDPs consists all training programs (CDPs) aiming at improvement of one or more specific competences. Due to not very large volume of CDPs data in this research, they have been clustered with traditional K-Means algorithm, but EMRKM can also be used in the case of higher volumes in the real world case studies. Clustering results in grouping of similar CDPs without any semantic interpretation of the clusters.

On the other hand, CAs and proposed metrics show clearly competence gaps of talents, specially for under-qualified ones. In order to improve the competitiveness of an under-qualified talent, he has to prioritize his competence gaps. This is done using AHP algorithm which is well suited for prioritization goals. The result is called "competence goal". Through calculating the Euclidean distance of a competence goal with the centroids of CDP clusters, a cluster of most related competence improvement solutions is identified. The competence goal is considered as a positive ideal in the TOPSIS algorithm and accordingly the best available competence improvement solution inside the selected CDP cluster can be discovered. Further details of the AHP to prioritize competence gaps, EMRKM

to cluster large volume of TP data, TOPSIS to find (1) the best TP to the desired JP and (2) the best CDP to the competence goal of under-qualified talent are described in the following sections.

5.2 Scalable Matching and Clustering of Talent and Job Profiles

Matching of talent and job profiles is in fact prioritization of job seekers in accordance with job descriptions. This matching process uses clustering algorithms in order to group similar talents and then find the best-fit TP cluster. Proposed Evolutionary MapReduce K-Means (EMRKM) clustering as well as TOPSIS algorithms are used to this aim. According to Garcia and Naldi, evolutionary K-Means shows better results and performance in comparison with traditional K-means [Garcia and Naldi, 2014]. The basic idea of the proposed EMRKM is inspired from the proposed algorithm in [Garcia and Naldi, 2014]. In particular, the goal of using EMRKM is to group similarly qualified talents into the same clusters and filter none-suitable talents towards speeding up the processing of large scale talent data.

In fact, clustering algorithm groups the talents without giving any further knowledge about interpretation of clusters. An euclidean distance between target JP and centroids of clusters has to be computed in order to select the best-fit cluster and exclude none-relevant ones from the computing. In the frame of best-fit cluster, TOPSIS searches and prioritizes the most competent talent according to the job description. In the case study of this work, parts of the dataset is prepared from the web data (bibliographic data) and need to be preprocessed for abstracting the data.

5.2.1 Pre-Processing of the Streamed Bibliographic Data

As described earlier, a bibliographic data is used to discover job qualification measures of computer scientists. The first issue in this regard is to discover thir research areas from their publications. Therefore, the first difficulty associated with streamed bibliographic data is identification of talents' research areas (referred as competence category, C_g) from their publications. These identified areas appear in TP and affect competence measures about them (e.g. scientometric measures such as h-index). As a solution, publication titles of talents are analyzed in order to extract their main research field from the titles of their publications. To this aim, Natural Language Toolkit (NLTK) provided as a Python package[1] is used. As a first, titles of the publications associated to a talent τ are summarized as strings. All republished entries with the same titles (i.e. multiple entries) of the same publication in the dataset are cleaned. A list of all words appeared in the publication titles of a talent τ are generated as his potential research domains

((Eq. 5.1)).

$$C''_\tau = \{w|w \text{ exists in the of title of talent } \tau\text{'s publications}\} \tag{5.1}$$

The Part-of-Speech Tagging (POST) method puts further tags to all words in the C'' list in order to filter and abstract the list. There are ready packages in programming languages such as a RDRPOSTragger package in Java and Python to reach this goal. Consequently, the RDRPOSTragger package is used to process the C''_τ list and exclude insignificant words from the list. The RDRPOSTragger package is a robust rule-based toolkit for POST and morphological tagging and supports 13 languages. It shows the speed of 5K words/seconds processing in the Python and 90K in the Java [Nguyen et al., 2014].

While abstracting the words in the list using the POST, a Penn Treebank tags retrain tagging models for English [Marcus et al., 1993]. Penn Treebank is a pre-existing tagging corpus that is developed at the Pennsylvania and provides the linguistic structure to the texts through tagging. The Penn Treebank consists of 4,5 million English words of American English and can be easily used for filtering of the words extracted from titles in order to keep only specific types of words such as singular or plural nouns. Since substantive words are to be considered as representatives of the competence category of a talent τ, they are once filtered using the following Penn Treebank tags (Eq. 5.2).

$$C'_\tau = \{w|w \in C''_\tau \wedge POS_w \in \{NN, NNS, NNP, NNPS\}\} \tag{5.2}$$

which NN, NNS, NNP and $NNPS$ denote singular or mass noun, plural noun, singular proper noun and plural proper noun, respectively and according to the Penn Treebank tags [Marcus et al., 1993]. All abstracted words listed in the C'_τ should be lemmatized. From linguistic point of view, the lemmatization is morphological and algorithmic method of detecting the lemma of the words in the list, removing inflectional endings and grouping altogether. A lemma is canonical form of the words [Skorkovská, 2012]. The lemmatization phase uses a WordNet lexical database which provides morphosemantic links in the Cross-POS as well [Fellbaum, 2005].

All words listed in the C''_τ do not represent the talent's research areas, because this list consists of all extracted words from titles including insignificant ones such as pronouns (e.g. "the", "towards", "and"). An additional processing to accurately remove insignificant words results in more abstracted list with the most relevant words to the competence (research) areas of a talent. In this regard, a list of stopwords referred as L_S is defined in German, English, Spanish, French, Russian and Portugese using the *stopwords* module of NLTK in Python [Bohlouli et al., 2015b]. In Natural Language Processing (NLP), stopwords such as "the", "about", "almost", or "any" refer to the words with less significance and vast amount of unnecessary information. As a result, they should be normally excluded

[1]Natural Language Toolkit 3.0 Documentation , accessed via `http://www.nltk.org` on 01 March 2016.

or removed in text processing or associated search queries. The set C'_τ is filtered based on *stopwords* and further criteria defined in (Eq. 5.3).

$$C_\tau = \{w | w \in C'_\tau \wedge w \notin L_S \wedge length(w) > 2 \wedge count(w) > 2\} \tag{5.3}$$

in which w consists of alphabetically characters. The result is C_τ which supposed to represent research areas of talents, but it is in fact most common words in the titles. This is sufficient for the goals of this work, but needs to be further researched to discover real research areas of talents from title and/or content of their publications which is discussed as a future work in section 7.2. The final step in the pre-processing is to create a new dictionary object and write all results in NoSQL DB (i.g. MongoDB in this work). To this aim, the remained words are supposed as keys and the total number of their occurrences is written as values in the DB.

Existing competence measures for scientists namely Bibliometrics or Scientometrics such as h-index or i10-index doesn't reflect active competence level of a talent in specific field. For instance, a person who was an expert in the "grid computing" and has recently changed his areas of interest to "data science" may be identified as an expert in "data science" as well based on stated measures. Because he has received higher h-index, i10-index or citation record for his publications in the "Grid Computing" in spite that he doesn't have any promising records in the "data science" as a newcomer in this area. In addition, some scientists may have been active in the past and have lost their motivation or reduced their scientific activities in the recent years due to reasons such as retirement or having new jobs, but their earlier published papers still receive very high citations and they can be seen as active competent scientists. Based on the current bibliometric measures, they are still being considered as active experts in the area.

As another example, imagine that there is a scientist who works in the "Grid Computing" area since 10 years ago and received i10-index of 15. In contrast, there is another scientist as a newcomer in this area since two years ago and has received an i10-index of 10. It is quite clear that a second person is more and more competent scientist than the first one, but the current bibliometric measures consider the first one as the most hard-worker. In fact, existing metrics provide the general and total measures about one person and lack the real-time (actual) measures of talents' activities and competences. In addition, they do not separate fields from each other, meaning that if a talent changed his research field three time in his career, results of all field are mixed together.

In addition to those problems, the growth of a person should be measured in comparison to the growth of the field. There may be some very competent scientists working in research areas that are not very popular and accordingly do not receive higher bibliometric measures. At the same time, there may be some scientists with average expertise in a popular research field(s) and receive higher bibliometric measures than the first case. In fact, they are not so good as the first scientists. But the current existing metrics show the reverse results. It is

too difficult to prioritize people for specific job definition based on such measures. The h-index particularly results in permanent values, meaning that a value, once reached, doesn't decline over time. This makes it impossible to get the picture of the talent's performance over time. In order to consider the time perspective as well, the scientific career of a person should be always measured in real-time and based on his activities through the time. In fact, the variable of time should be important and considered in the measurements.

5.2.2 Computing Scientific Competence Factor of Talents

A new competence measurement metric for academic career called SCF is defined in this research to indicate current competences and activeness of a talent and his development over time. In fact, it identifies a scientific contribution of a talent to specific field. According to this competence measure, a competent talent is a knowledgeable person in his field and steadily improves his knowledge. This competence measure considers the citation count, and analyses how a talent performed in comparison to his field of the research. In fact, it compares the competence development of a talent in comparison with the development of his respective field. In this way, it is also possible to compare talents who work in different fields. This measure is a snapshot-like view on the current performance of a talent, in contrast to the h-index, always analyses a time range of the last three years.

Definition of the SCF has been inspired from average acceleration and velocity formulas in physics stated in the Equations 5.4 and 5.5. It observes current values, but also takes the "citation lag" into consideration.

$$\overline{a} = \frac{\Delta v}{\Delta t} = \frac{v_2 - v_1}{t_2 - t_1} \tag{5.4}$$

$$\overline{v} = \frac{\Delta x}{\Delta t} = \frac{x_2 - x_1}{t_2 - t_1} \tag{5.5}$$

The SCF of a talent τ in a field f at the time t is defined in the (Eq. 5.6). The SCF, equation (Eq. 5.6), is then normalized through dividing it by the absolute value of the field's acceleration. The formula for the talent's and field's acceleration are being computed using the equations 5.7 and 5.8 with slight modification of using achieved citation counts instead of a covered distance.

$$SCF_{\tau,f,t} = \frac{a_{\tau,f,t} - a_{f,t}}{|a_{f,t}|}, t \geq 2 \tag{5.6}$$

$$a_{\tau,f,t} = c_{\tau,f,t} - 2c_{\tau,f,t-1} + c_{\tau,f,t-2} \tag{5.7}$$

$$a_{f,t} = c_{f,t} - 2c_{f,t-1} + c_{f,t-2} \tag{5.8}$$

where, $c_{\tau,f,t}$ is only the citation count of a talent τ belonging to the field f in the year t and $a_{\tau,f,t}$ is an acceleration of a talent τ in the field f at the time t. Similarly,

$a_{f,t}$ indicates the acceleration of the field f for the equivalent time intervals. It is clear that talents' competence development and their career growth (success) are computed in comparison to the growth of their respective field. So, if the field is not an active area, but the scientist if good enough, he will get promising metrics. The highlight of this formula is that the non-relevant publications or associated ones to the other fields are excluded from calculations. In addition, $c_{f,t}$ is the citation count of all publications in the field f in the year t. All publications of the field f in the year t are known in the DB. Accordingly, their citation count is also recorded in the DB. The total citations of a field ($c_{f,t}$) in the year t is sum of the citations of all related publications in the year t.

Considering the total citation counts in the time intervals (years) of $t-1$ and $t-2$ originates from the physical acceleration formula and favors the effects of a "citation lag". The best interpretation of the SCF can be achieved by evaluating and comparing the results of more than three years. This supports giving an opportunity to absolute newcomers to the science to develop and show their talent in growing up in the scientific career. As stated earlier, further similar metrics have to be developed or adopted to other case studies and areas due to the fact that SCF and AIS are dedicated to the scientific careers.

t	1997	1998	1999	2000	2001	2002	2003	2004
Talent A	2396	635	1227	245	618	730	176	495
Talent B	1258	1173	1667	785	728	896	634	1539
Talent C	7821	577	144	159	367	150	89	146
Talent D	377	278	301	724	382	302	438	734
Talent E	878	1154	18	2097	1063	661	690	870
Total field's citation	15769	6320	5759	6934	6339	5602	7336	7588

Table 5.1: Citations per year of authors in specific field between 1997 and 2004

In order to visualize the SCF with real facts and data, citation counts of top 20 scholars in one specific computer science field have been recorded according to the Google Scholars as showed in Table 5.1. These citation counts are visualized in Figure 5.2 as a form of stacked bar chart and shows the distribution of citations for those five talents between 1997 and 2004[2]. In the world of science, conferences and journals are indexed differently. Therefore, different indexes and scholar DBs consist of various information. One of the most completed DBs in this regard is Google Scholar which doesn't provide any access to its DB or any API to retrieve the data automatically. As a result, required test data of 200 talents had to be collected manually in order to ensure the accuracy of the metrics. But this is supposed to be an automatic process in the future work through crawling or

[2]The data providing for this analysis was extracted from Google Scholar to gain accurate results, using 20 authors active in the field.

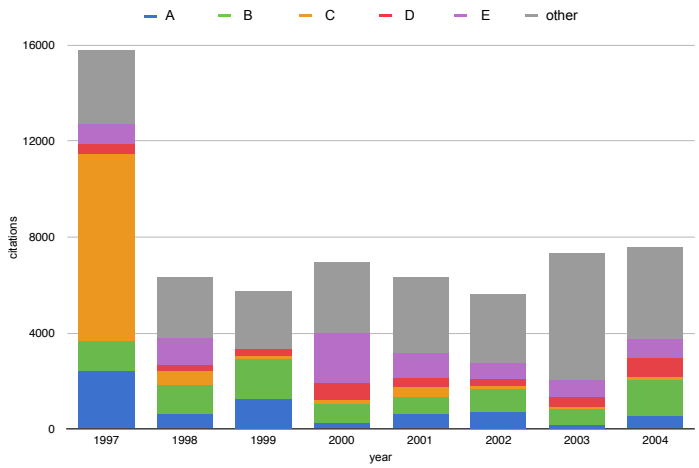


Figure 5.2: Visualization of the citation counts stated in Table 5.1 as stacked bar chart

provided APIs.

It should be noted that citations of a talent's publications is counted to the year the cited publication was published and not the year it has been cited. In the numerical example provided in this section, all other citations except those selected top 5 talents of the field are marked as "other". The sum of all five talents' citations and the other citations results in the total citations of the field. Therefore, the bar chart of a talent in Figure 5.2 represents the proportion of the person's citations on the total citations. The values in Table 5.2 are computed from the

t	1999	2000	2001	2002	2003	2004
A	-0.7353	-1.1907	1.7655	-0.8380	-1.2695	1.5891
B	-0.9349	-0.7926	1.4661	2.5845	-1.1740	1.7874
C	-0.2337	-0.7419	1.1090	-1.9930	-0.9369	1.0796
D	-0.9863	-0.7696	0.5678	2.8451	-0.9126	1.1080
E	-1.1589	0.8520	-0.7588	5.4507	-0.8256	1.1019

Table 5.2: Computed SCF for the data stated in Table 5.1 in specific field between 1999 and 2004

citation counts in Table 5.1. Accordingly, Figure 5.3 shows the visualization of SCF for top five talents of one specific area between the years 1999 and 2004 that have been computed in Table 5.2. Interpretation of this chart shows the importance of proposed SCF metric in assessing the job quality of scientists.

Newcomers need some time to develop their personal and professional com-

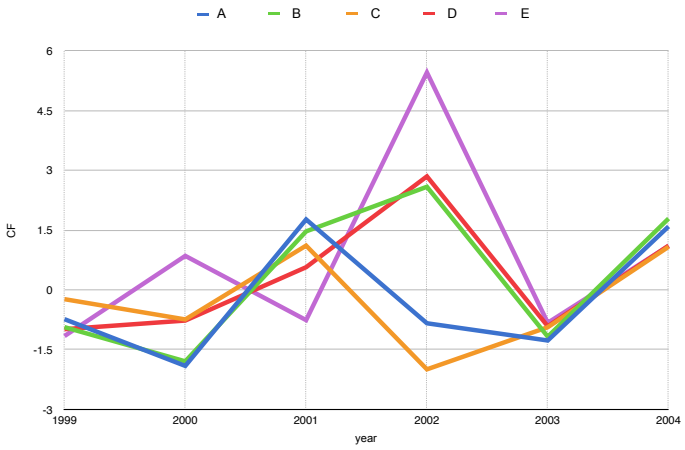


Figure 5.3: Visualization of SCF results computed in Table 5.2

petences in one specific field. To this aim, SCF formula provides two years gap to scientists for receiving some citation to their publications. It is also clear in Table 5.1 which the citation count starts since 1997, but SCF starts from 1999 in Table 5.2. This two years gap depends on equations 5.7 and 5.8, which indicates two years distance in the formula. In particular, the SCF has following important highlights and advantages:

- Considering the growth of a research in measuring career development of a talent: If a research field is not so popular and therefore a competent researcher with good quality publications doesn't receive too many citations similar to his research field, his SCF will not show him as incompetent person.
- Preventing well known talents of the field from dominating the whole field with just one or few of their existing publications: If a talent publishes one or more well qualified papers which they always receive high number of citations, but he doesn't stays active in the field, the SCF prevents him from being identified as an active competence person of the field. This is because his citations of those well qualified papers will be counted to the year of publications rather than citation years. As a result, in order to get stable positive SCF value, scientists have to always stay active in the field.
- A generic and comparable metric for all talents even from different areas: Existing metrics (e.g. h-index) of different hard working talents of two different areas, one very popular and another not popular, are not comparable. But comparative nature of the SCF considers the growth and popularity of the field as well. As a result, different scientists from different areas can be compared using SCF.

As it is clear from the state-of-the-art, those proposed highlights of the SCF are

existing critics to the currently available metrics such as h-index, i10-index and etc. In fact, they may result in some interesting statistics and also information about scientists, but not good candidates to measure the CK level and competence of the talents.

5.2.3 Active Influence Scientometric of Talents

In addition to the SCF, mapping of talent's competence development influence over time to the fields' growth is important as one of talent's scientific CK measures. To this aim, an AIS metric (Eq. 5.9) is developed which defines how a talent (i.e. scientist) influences his specific research field over time or in specific amount of years. Competences, and activities of talents are interpreted and identified based on their citation counts. Since the AIS index uses yearly basis calculations, the earliest time that this index can be measured for any given field is one year after releasing the first publication in the given field. The AIS is computed using the (Eq. 5.10).

$$AIS_{\tau,f,t} = \frac{\text{Average yearly influence of a talent } \tau \text{ in a field} f}{\text{Average yearly development of the field}} \tag{5.9}$$

$$AIS_{\tau,f,t} = \frac{\frac{C_{\tau,t}}{\delta t_\tau}}{\frac{C_{f,t}}{\delta t_f}} = \frac{C_{\tau,t} \times (t - t_{0,f})}{C_{f,t} \times (t - t_{0,\tau})} \tag{5.10}$$

where, $C_{\tau,t}$ is total number of citation counts of a talent τ in the year t. In addition, $t_{0,\tau}$ denotes the year of talent's first publication in the field f. In this equation, δt_τ denotes whole time period that a talent τ is active in the field f and δt_f indicates the whole time period that a field f has been started till now. It should be noted that the $C_{f,t}$ represents total citations of the field f since the first publication of the field in the year $t_{0,f}$. All citations associated to any publication are counted to their publication year, not to the citation year.

While the SCF can be seen as an snapshot of 3 years timespan, the AIS is an even "narrower" snapshot, only taking one specific year into account while bearing in mind the history of the field. To achieve a better understanding of how the AIS works, one should look at Figure 5.4 based on input values and calculations in Table 5.3 and Table 5.4. This figure shows the development of 13 authors in the field of "cloud computing". The citation counts used for this example were extracted from Google Scholar and can be found in Table 5.3.

A better example of one talent dominating a field is the situation of a researcher B in 2009. From the total 6899 citation counts of the field, talent B received 5969 citations. In other words person B received approximately 86.52% of the total citations, essentially becoming as fast growing as the field itself. This results in the lower AIS values for other researchers of the field at this time, since they are being compared to B. The year 2015 shows that the AIS enables others to take the leadership of the field. This measure provides an opportunity to the newcomers to clearly show up their competences and skills while growing up in

year	A	B	C	D	E	F	G	H	I	J	K	L	M	total (field)
2001	0	0	0	0	0	0	0	0	0	0	0	0	1	1
2002	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2003	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2004	0	0	0	0	0	0	0	0	0	0	0	0	3	3
2005	0	0	0	17	0	0	0	40	0	0	0	6	0	63
2006	0	0	0	2	0	0	0	0	0	0	0	12	0	14
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2008	0	1887	0	15	0	0	0	0	0	0	0	0	525	2427
2009	0	5969	0	0	0	0	12	0	918	0	0	0	0	6899
2010	0	2397	0	62	0	19	17	379	0	0	0	0	585	3459
2011	0	4181	0	13	0	0	0	148	2	0	12	0	138	4494
2012	0	1824	0	2	2	0	0	26	85	20	0	27	37	2023
2013	0	683	3	1	1	0	0	74	0	0	16	0	32	810
2014	0	345	0	1	0	6	0	0	1	0	0	0	31	393
2015	0	3	0	0	0	0	0	0	0	0	0	0	9	12

Table 5.3: Citations for the field of "cloud computing"

year	A	B	C	D	E	F	G	H	I	J	K	L	M
2001	0	0	0	0	0	0	0	0	0	0	0	0	0
2002	0	0	0	0	0	0	0	0	0	0	0	0	0
2003	0	0	0	0	0	0	0	0	0	0	0	0	0
2004	0	0	0	0	0	0	0	0	0	0	0	0	1
2005	0	0	0	0	0	0	0	0	0	0	0	0	0
2006	0	0	0	0.7143	0	0	0	0	0	0	0	4.2857	0
2007	0	0	0	0	0	0	0	0	0	0	0	0	0
2008	0	0	0	0.0144	0	0	0	0	0	0	0	0	0.2163
2009	0	6.9216	0	0	0	0	0	0	0	0	0	0	0
2010	0	3.1183	0	0.03226	0	0	0.04423	0.1972	0	0	0	0	0.1691
2011	0	3.1012	0	0.0048	0	0	0	0.0549	0.0022	0	0	0	0.0307
2012	0	2.4795	0	0.0016	0	0	0	0.0202	0.1541	0	0	0.0210	0.0183
2013	0	2.0237	0	0.0019	0.0148	0	0	0.1370	0	0	0.1185	0	0.0395
2014	0	1.9020	0	0.0037	0	0.0496	0	0	0.0066	0	0	0	0.0789
2015	0	0.5	0	0	0	0	0	0	0	0	0	0	0.75

Table 5.4: AIS for the field of "cloud computing". Due to the fact the $(t - t_{0,\tau})$ returns 0, the citations of talents' first year are not being evaluated in the AIS formula. For this reason, as it is seen in this table, despite the fact that talents D H and L have received citations in 2005, but their AIS is equal to 0, because they just entered to the field in 2005. This is being reflected in the citations of the year after

the field. In addition, it respects the field's growth, meaning that scientists of less demanding (being cited) fields don't look like less competent as more demanding fields. Consequently, the highlights of the AIS are as follows:

– Giving an opportunity to newcomers: Imagine that a scientist who is active in one specific field since 20 years ago received h-index of 15. At the same time, another scientist who has entered to this field since two years ago received

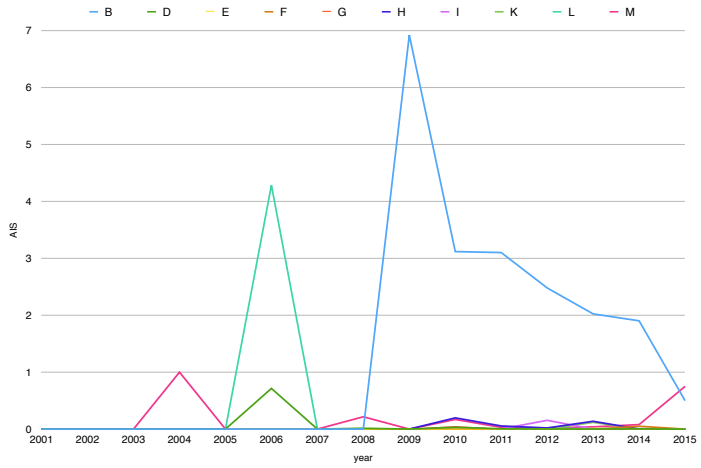


Figure 5.4: Visualization of the AIS results computed for the field of "Cloud Computing"

h-index of the 12. Both are active in just this field since beginning of their academic career. According to the h-index values, the first scientist seems to be more competent than the second one. But in fact, the second scientists in more competent researcher than the first one. The AIS clearly considers this point in the calculations and provides nearly real-time measurements.

– Considering time variant in the metrics: Current metrics suppose citation year rather than a publication year of articles in the calculations. Imagine the case of a scientist who published a paper in 2000 and received a citation count of 5,000 in 2016. He moved to the industry in 2014 and is not active in the science anymore. According to the current metrics, he is still competence in this field in the year 2016 which doesn't truly reflect the competence of a person.

5.2.4 Scalable Clustering of Talents based on Quality Measures

Associated computations in sections 5.2.2 and 5.2.3 are used to prepare real 200 talent data. As stated earlier, they are grouped as job related skills ($C_{1,3}$) in the CKR model and also get higher weights in assessments and making final recruitment decisions. It should be stressed that these values are associated with the time variant and depending on different years, they produce different SCF and AIS values. In the frame of this work, the latest achieved SCF and AIS values are stored in the TP, but one can consider for instance mean of all achieved values. This achieved real talent data is the basis of regenerating large scale artificial data (15 million talent data) in chapter 4. Traditional data analysis methods and algorithms are not capable of handling such large volumes. Consequently, scalable solutions based on the big data technology provide significant performance

improvement in analyzing and clustering of large scale datasets. Efforts made in chapter 4 resulted in assignment of the K-Means clustering algorithm for the case study of this research.

Fundamentals of K-Means algorithm have been described in earlier sections. It has been presented by MacQueen in 1967 [MacQueen, 1967]. A MR deployed version of the K-Means algorithm is available in the Apache Mahout and provides almost satisfactory performance improvements in comparison to the tradition algorithms, but still needs further improvements which are discussed in the following. Apache Mahout, as discussed in chapter 2, provides scalable machine learning and data mining algorithms based on the MR and Hadoop ecosystem, but doesn't cover all machine learning and clustering algorithms. The process of how K-Means clustering works as *Map* and *Reduce* jobs is showed in Figure 5.5. In addition, an algorithm of MR based K-Means Mapper and Reducer job is showed in the Listings 1 and 2.

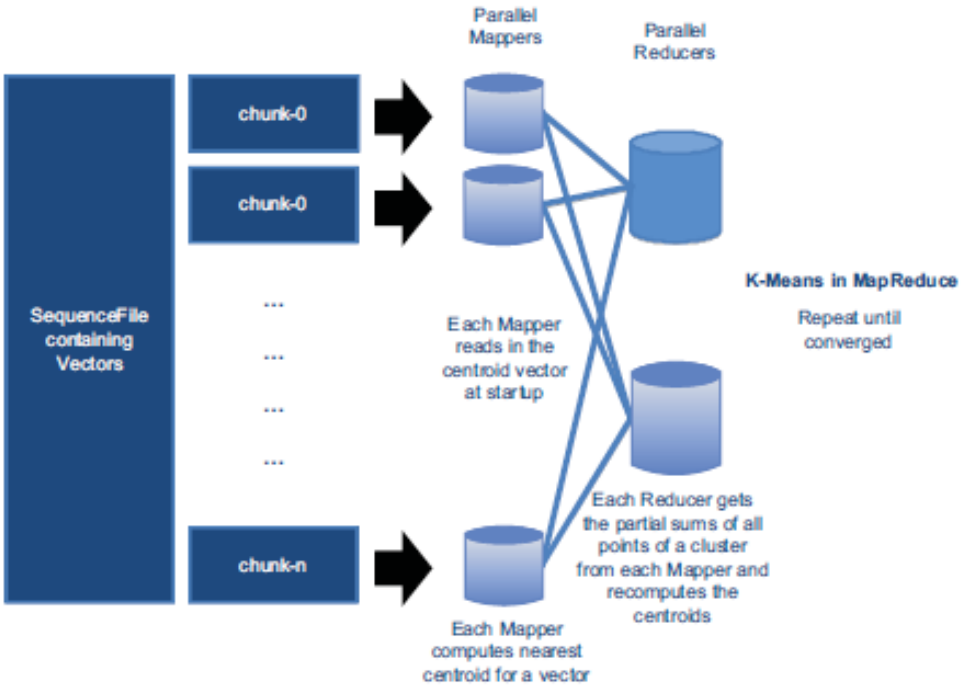


Figure 5.5: Running K-Means algorithm in MR showing Mappers and Reducers [Owen et al., 2011]

In order to efficiently design K-Means clustering with MR and also improve its performance, it is imperative to understand details of mappers and reducers in the K-Means design. According to [Owen et al., 2011], clustering algorithms consist of CPU-bound operations such as vector serialization or distance computation

as well as I/O-bound operations like transmitting *centroids* to *Reducers* over the network. MR based K-Means clustering algorithm runs in parallel through multiple mappers and reducers. Owen et al. suggest to decrease the number of clusters (k), if possible, in order to increase the performance of MR based clustering. He argues that "Clusters are usually represented as large, dense vectors, each of which consumes considerable storage. If the clustering job is trying to find a large number of clusters (k), these vectors are sent across the network from the Mappers to the Reducers" [Owen et al., 2011]. This will decrease transferring more information via the network and as a result the performance of I/O bound operations will be improved.

Algorithm 1 Pseudo-code of a Map job for MapRedcue based K-Means [Zhao et al., 2009]

```
1: procedure K-means Map
2:     minDistance ← DoubleMAX
3:     index ← -1
4:     for i ≤ length(centers) do
5:         dist ← ComputeDist(instance, centers[i])
6:         if dist < minDistance then
7:             minDistance ← dist
8:             index ← i
9:     key ← index
10:    Construct value as a string comprise of the values of different dimensions
11:    return < key, value >
```

Zhao et al. suggested well designed algorithms for *maps* and *reduces* of the MR deployed K-Means [Zhao et al., 2009]. These algorithms clearly define how reducers recalculate centroids of clusters in the K-Means with collected clustering results from mappers. These algorithms of *map* and *reduce* jobs of the K-Means algorithm are showed in the listings 1 and 2. The K-Means algorithm developed in this research is called Evolutionary MapReduce K-Means (EMRKM) and uses MR to support distributed large scale computing and data clustering. The basic idea of the EMRKM is inspired from Scalable Fast Evolutionary Algorithm for Clustering (SF-EAC) [Oliveira and Nald, 2015; Naldi and Campello, 2014]. It should be noted that EMRKM shows better results in the quality and speed comparing it with the SF-EAC algorithm. This issue is discussed further in the following and chapter 6.

In general, an evolutionary K-Means draws inspiration of Darwinian natural selection. According to the Darwin's theory, K-Means can be simply understood as **survival of the fittest**. The natural selection usually prefers those individuals that fit the environmental conditions the best. Evolutionary computing, given a population of individuals, results randomly in the creation of some candidate

Algorithm 2 Pseudo-code of a Reduce job for MapRedcue based K-Means [Zhao et al., 2009]

```
1: procedure K-means Reduce
2:     Initialize one array record the sum of value of each dimensions of the
       samples contained in the same cluster, e.g. the samples in the list V;
3:     Initialize a counter NUM as 0 to record the sum of sample number in the
       same cluster;
4:     while V.hasNext() do
5:         Construct the sample instance from V.next()
6:         Add the values of different dimensions of instance to the array
7:         NUM + = num
8:     Divide the entries of the array by NUM to get the new centroids
9:     Construct value as a string comprise of the centroids
10:    return < key, value >
```

solutions. The candidate solutions can be evaluated by means of functions, usually named a *fitness function*, to get a fitness value, the higher the better [Eiben and Smith, 2003].

According to the fitness values, some candidates are selected as seeds for giving birth to the next generation by carrying out crossover or mutation to them. Crossover (also called recombination) is an operator used for two or more chosen candidates (the parents) and generates one or more children. Mutation operator is applied to only one candidate and produces a new candidate. The whole process will be run iteratively until a "good enough" candidate is found or other limitations are reached [Eiben and Smith, 2003].

Eiben and Smith summarized the following elements for evolutionary algorithms in which are basics of the EMRKM as well: (1) Representation, (2) Fitness Function, (3) Population, (4) Initialization, (5) Selection, (6) Variation Operators, and (7) Termination Condition [Eiben and Smith, 2003]. As discussed repetitively in this work about the K-Means clustering, it has inherent drawbacks as well. For instance, the quality of K-Means clustering depends heavily on the initial centroids and needs the number of clusters, k, to be defined as an input. This is usually unknown or hard to predict in real applications, specially when the amount of data is huge. Oliveira and Nald claimed that SF-EAC results in an ideal set of clusters with convenient running time and power while multiple runs of the K-Means [Oliveira and Nald, 2015]. In the frame of this work, the SF-EAC has been redeveloped and tested with regenerated CKR model data in order to fairly compare improved achievements of the EMRKM.

A partition consists of whole dataset and results in one alternative clustering solution which is independent of solutions in all other partitions. Each partition is represented by its all cluster centroids, because once the centroids of a partition are determined, all points can be assigned according to the centroids. Accordingly,

the total number of partitions (genotype) is referred as population size and the higher population size, finding the optimum solution is faster. Because by having greater population size in the EMRKM, more values of the k will be checked in each generation and consequently the best value of the k can be found sooner. It particularly solves the problem of the need for defining the k at the beginning of the clustering. Different potential values of the k are being proofed in different partitions and the best result according to the silhouette values is selected. The only need in this regard is a maximum and minimum of the k.

EMRKM examines different possible values of the k from defined range in different partitions. The main concern in the EMRKM is to run an algorithm for multiple times with various ks and centroids and choose the best solution based on quality (silhouette) and fitness measures. In the EMRKM implementation, there is a class named "cluster" which consists of following elements: (1) a cluster ID, (2) a partition ID, (3) a Centroid, (4) a Cardinality, (5) a convergence, and (6) a fitness value. A cluster and partition ID which are integer values represent the ID of an existing cluster and its associated partition, accordingly. A centroid as it is clear from its name indicates the specifications of the final centroid of the cluster. The cardinality value records the total number of elements in the cluster. A boolean convergence value becomes true, when the cluster becomes converged according to defined convergence threshold. The convergence threshold is defined 0.01 in this work.

Furthermore, there is a class named "Talent" in the EMRKM implementation which consists of the following important factors: (1) Talent ID, (2) partition ID, (3) a matrix of competence values, (4) combiner count, (5) Distance A, (6) Distance B, and (7) simplified silhouette value. Talent ID and partition ID are integer numbers and similar to definitions in the cluster class. A matrix of competence values is a string which consists of talents' 84 competence values according to the CKR model which are achieved through assessments and retrieval from digital sources and cover all levels of the competence tree. A "distance A" value computes a distance of an element with a centroid of belonging cluster. Furthermore, a "distance B" value computes a distance of an element with a centroid of nearest cluster. Finally, a simplified silhouette value is computed using (Eq. 4.13) and is in a range of $(-1, +1]$. The closer to $+1$, the quality of clustering is better.

An array of ***centroids*** stores the number of clusters for every genotype. Note that, the number of clusters in different genotypes is randomly chosen in an interval of K_{min} and K_{max}. The algorithm consists of different iterations called generations in each partition (or genotype). The termination condition of the algorithm is to either reach a total number of defined generations or a clustering in any of the genotypes is converged. The convergence condition is to reach a specific defined value of simplified silhouette value. This simplified silhouette value is normally based on former experiences.

As it is showed in Table 5.5, each genotype produces two new children (genotypes) resulted from applying both of Mutation Operators (MO). As an example, GT05 and GT06 are children of GT01 which have been resulted from MO1 and

MO2, accordingly. For transmission from one generation to the new one, four different genotypes are selected using roulette wheel strategy. For instance, in a transmission from generation 01 to the 02 in the provided numerical example of Table 5.5, four genotypes (GT7, GT08, GT02 and GT06) are selected from the list of 9 candidate genotypes (GT2, GT05, GT06, GT07, GT08, GT09, GT10, GT11 and GT12). Candidate genotypes in each generation are those reproduced child genotypes as well a parent genotype with maximum silhouette value.

Table 5.5: An example of evolutionary K-Means iterations (generation) and their associated simplified silhouette computation for each genotype

	Genotype		Genotype		Genotype		Genotype	
Generation 01	GT01=0.3391		GT02=0.8210		GT03=0.6421		GT04=0.4193	
	GT05=0.3310	GT06=0.425	GT07=0.7901	GT08=0.7821	GT09=0.4523	GT10=0.5345	GT11=0.8023	GT12=0.2363
Generation 02	GT07=0.7901		GT08=0.7821		GT02=0.8210		GT06=0.425	
	GT13=0.3421	GT14=0.8421	GT15=0.7687	GT16=0.6734	GT17=0.3421	GT18=0.8323	GT19=0.1198	GT20=0.8362
Generation 03	GT20=0.8362		GT13=0.3421		GT16=0.6734		GT02=0.8210	
	GT21=0.8935	GT22=0.6365	GT23=0.5763	GT24=0.8965	GT25=0.3421	GT26=0.9256	GT27=0.4529	GT28=0.8329

The first population is initialized by randomly selecting n points from the data set, where n is equal to the sum of all elements in the ***centroids*** array. In each generation, every genotype is fine-tuned by the k-means algorithm, a maximum number of iterations t and convergence are adopted as the stopping conditions. The genotype with the highest fitness is directly copied into the next generation, then one has to select some genotypes from the remaining according to the roulette wheel strategy (proportional selection) to be mutated. Oliveira and Nald defined two mutation operators, namely eliminate (MO1) and split (MO2) [Oliveira and Nald, 2015]. One of these mutation operators is being applied randomly to the current generation of genotypes and results in new generation. Oliveira and Nald analyzed the simplified silhouette values of resulted genotypes to see whether the mutation operators changed a clustering positively or negatively. In the case of negative effect in the clustering, other mutation operator is being selected for the next generation.

One of the important highlights of developed evolutionary K-Means in this work in comparison with the [Oliveira and Nald, 2015] is applying both mutation operators in each generation of genotypes and choose the better result. This improves the quality of the clustering and also the speed of reaching to the best solution very significantly. Performance measures and evaluation of results are given in chapter 6. Furthermore, another important highlight of the proposed evolutionary algorithm in this work is its mutation operators. In addition to the eliminate and split operators, a merge mutation operator is also applied to genotypes in each generation and the best result is selected between results of those three mutation operators. Table 5.5 shows how silhouette value affects the process of iterating different generations of an evolutionary K-Means. In this table, green colored partitions are converged and red colored ones remain without changes in the next generation. One highlight of the EMRKM in comparison

to the similar algorithms like the one from [Oliveira and Nald, 2015; Naldi and Campello, 2014] is its clustering results becomes more accurate and close to final solution in each generation (iteration) of it.

5.2.5 Matching Clustered Talent Profiles with the Job Profile

Matching of TPs with target JP contributes to the person-job-fit challenge as discussed in chapter 1. Imagine that an enterprise has already identified a need to recruit a new employee for specific needs and released a job position announcement (i.e. JP). The JP consists of all required CK (RCK matrix) and is represented in a form of matrix (Eq. 3.5) as showed in section 3.2. This once more repeated in the equation (Eq. 5.11).

$$RCK = \begin{bmatrix} r_1 & \dots & r_n \end{bmatrix}_{1\times n} \tag{5.11}$$

$$ACK = \begin{bmatrix} c_1 & \dots & c_n \end{bmatrix}_{1\times n} \tag{5.12}$$

where $n = 84$ represents the total number of competence categories and r_i represents the required values (level) of the competence categories in the CKR model which $0 \leq r_i \leq 1$; $i = 1, ..., n$ and $\sum_{i=1}^{n} r_i = 1$. It should be once more stressed that r_i doesn't address the weight, but the value of the competence which is required for the job. As an example, in description of the "proficient knowledge in English language", the word "proficient" is being addressed as r_i. In this way, those values are comparable with acquired competence values of talents achieved through assessment methods.

$$d_{r,c}^2 = \sum_{i=1}^{n} (r_i - c_i)^2 \tag{5.13}$$

Given that SCF, AIS as well as further assessment methods and metrics result in the ACK matrix of a talent (i.e. TP) and TPs are then clustered using EMRKM algorithm, the problem is to find the the most relevant TP cluster and consequently prioritize talents of this cluster based on desired JP. To this aim, the squared euclidean distance of the target JP with the centroids of all TP clusters is computed using (Eq. 5.13). Centroids of the TP clusters are in fact competence data of the specific talent which has been represented as RCK matrix. Each TP is represented as ACK which has been discussed in section 3.2 and showed once more in the equation (Eq. 5.12). The result of squared euclidean distance computation is selection of a TP cluster with the shortest distance to already defined job position (JP). The selection of specific talent to target job is still not clear.

Notably, finding the best talent inside one specific TP cluster and in accordance with the desired job description is on the basis of the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method. In this method, a decision matrix like the one showed in the (Eq. 5.14) defines requirements of the

problem. The $m \times n$ decision matrix consists of the m alternatives (i.e. total number of talents in the selected cluster) and n criteria (i.e. total number of competence categories of the CKR model).

$$D_{p\times n} = \begin{array}{c} \\ \tau_1 \\ \vdots \\ \tau_i \\ \vdots \\ \tau_p \end{array} \begin{array}{c} \begin{array}{ccccc} c_1 & \cdots & c_j & \cdots & c_n \end{array} \\ \begin{pmatrix} c_{11} & \cdots & c_{1j} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{i1} & \cdots & c_{ij} & \cdots & c_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{p1} & \cdots & c_{pj} & \cdots & c_{pn} \end{pmatrix} \end{array} \tag{5.14}$$

where n is the total number of competence categories in the CKR model ($n = 84$) and c_{ij} represents achieved competence value of talent i for the competence category j in the CKR model. In addition, the total number of existing talents in the selected cluster is considered to be p.

Hwang and Yoon suggest to normalize elements of the decision matrix by vector normalization method as showed in the (Eq. 5.15) [Hwang and Yoon, 2012]. The square of denominator in (Eq. 5.15) avoids having a divide by zero error. The normalized decision matrix using (Eq. 5.15) is denoted as $D^*_{p\times n}$.

$$c^*_{ij} = \frac{c_{ij}}{\sqrt{\sum_{i=1}^{p} d^2_{ij}}}; j = 1, ..., n \tag{5.15}$$

where d^2_{ij} is the outcome of equation (Eq. 5.13). Furthermore, a weighted normalized decision matrix (5.17), $\hat{D}_{p\times n}$, is achieved through multiplying normalized decision matrix, $D^*_{p\times n}$, by already predefined JP matrix as showed in (Eq. 5.16). In fact, each row in the JP matrix is the RCK showed in (Eq. 5.11).

$$JP_{n\times n} = \begin{bmatrix} r_1 & \cdots & r_1 \\ \vdots & \ddots & \vdots \\ r_n & \cdots & r_n \end{bmatrix}_{n\times n} \tag{5.16}$$

$$\hat{D}_{p\times n} = D^*_{p\times n} \times JP_{n\times n} \tag{5.17}$$

Two "virtual alternatives" are constructed from the weighted normalized decision matrix as of (1) "positive-ideal" alternative (A^+) and (2) "negative-ideal" Alternative (A^-). Those positive and negative-ideal alternatives demonstrate respective best and worst criteria expressions and are computed using equations 5.18 and 5.19 [Hwang and Yoon, 2012], accordingly.

$$A^+ = \left\{ \max_i \left(\hat{c}_{ij} \right) \right\} = \left\{ \hat{c}^+_1, \cdots, \hat{c}^+_p \right\} \tag{5.18}$$

$$A^- = \left\{ \min_i \left(\hat{c}_{ij} \right) \right\} = \left\{ \hat{c}^-_1, \cdots, \hat{c}^-_p \right\} \tag{5.19}$$

where, $\hat{c}_{ij}$ is an element of normalized weighted decision matrix, $\hat{c}_i^+$ and $\hat{c}_i^-$ denote criteria with positive and negative impacts, accordingly.

The similarity of all alternatives, talents in the selected TP cluster (τ), to the best and worst alternative is calculated using the Euclidean distance of each talent, τ, with the best (A_i^+) and the worst (A_i^-) alternatives as described in the equations 5.20 and 5.21 [Hwang and Yoon, 2012].

$$S_{i^+} = \sqrt{\sum_{j=1}^{n} \left(\hat{c}_{ij} - \hat{c}_j^+\right)^2};\ i = 1, \ldots, p \tag{5.20}$$

$$S_{i^-} = \sqrt{\sum_{j=1}^{n} \left(\hat{c}_{ij} - \hat{c}_j^-\right)^2};\ i = 1, \ldots, p \tag{5.21}$$

As a final step, all achieved distance indexes are normalized using (Eq. 5.22) in order to ensure an index between $[0, 1]$.

$$C_{i^+} = \frac{S_{i^-}}{S_{i^+} + S_{i^-}},\ 0 \leq C_{i^+} \leq 1;\ i = 1, \ldots, p \tag{5.22}$$

As a result, the distance index becomes 1 when the alternative is equal to the "positive-ideal". The distance index of 0 means that a talent is equal to the "negative-ideal". Achieved results allow to rank talents of the selected TP cluster on the basis of the distance index. Accordingly, a talent with the highest index is the best fitting candidate to already announced job position.

5.3 Matching Identified Gaps and Development Profiles

Calculation of talents' CK level by metrics such as SCF and AIS and clustering and sorting them based on job knowledge level of talents result in an identification of their competence gaps (lacks), specially for those under-qualified ones and those who want to achieve required competences of specific job. These competence gaps should be fulfilled by participation in programs such as training or workshops. As a result, the main concern here is to analyze competence gaps of talents, proof currently existing CDPs and recommend the best fit ones to their competence goal(s). Furthermore, enterprises may plan to recruit a talent who is not the best fit to desired JP and accordingly provide him some further trainings (CDP) to fit him to the target job position. A CDP is for instance on-the-job-training, course, seminar, workshop, webinar, internship or any other activity that may improve one or more specific competence(s). Each profile (CDP) specifies which competence can be improved and in which level by participating in or using this program. A matching problem here is to find the best CDP to already defined competence goal.

5.3.1 Identification of Competence Gaps (Goals)

In order to improve the competence level of talents through matching their competence gaps (goals) with CDPs such as on-the-job-training, their competence goal should be first defined. In this regard, an AHP[3] method is used to weight and prioritize one or few competence(s) from the CKR model. The AHP algorithm provides pairwise comparisons between competences in order to identify the most important competence goal [Saaty, 1988]. In this way, an employee or a talent should prioritize competences in a pairwise comparisons that he wants to get training and improve them.

An important difficulty issued for the use of AHP algorithm in prioritization is its large number of required pairwise comparisons, in the case of many alternatives. For instance, a total number of 105 pairwise comparisons are required for prioritizing 15 competence goals, which is nearly impossible. But due to the fact that a number of desired competences to be improved through trainings at the same time are not too much in the reality, this difficulty can be ignored. In addition, there are some new techniques and methods for reducing the number of pairwise comparisons in the AHP algorithm that can be used as future work of this research.

Assume that in accordance with the proposed metrics in sections 5.2.2 and 5.2.3, it has been concluded that a talent should improve his knowledge in one of the following areas (or competences) of the "Cloud Computing" field in order to improve his scientific competitiveness:

- Distributed System Models and Technologies (DSM)
- Parallel Programming Systems and Models (PPS)
- Workflow Systems (WS)
- Virtualization Technologies(VZ)

To this aim, he prioritized and weighted existing choices on the basis of the AHP algorithm and reached to an exemplary evaluation matrix of (Eq. 5.23). Further details of the AHP algorithm and how pairwise comparisons should work and which steps to be taken to reach an evaluation matrix are discussed in the [Saaty, 1988].

$$E_{n\times n} = \begin{array}{c} \\ DSM \\ PPS \\ WS \\ VZ \end{array} \begin{array}{c} \begin{array}{cccc} DSM & PPS & WS & VZ \end{array} \\ \begin{pmatrix} 1 & 3 & 7 & 9 \\ \frac{1}{3} & 1 & 2 & 5 \\ \frac{1}{7} & \frac{1}{2} & 1 & 3 \\ \frac{1}{9} & \frac{1}{5} & \frac{1}{3} & 1 \end{pmatrix}_{4\times 4} \end{array} \tag{5.23}$$

[3]The AHP method can also be used in prioritization of required competences while setting up the JP. The only difficulty in this regard is the total number of pairwise comparisons in the AHP method. Since there are 84 dimensions in the CKR model in this work, it is not feasible to use AHP in prioritization of required CK. For case studies with less competences, the use of AHP is recommended.

where the elements of such evaluation matrix are integer values in the range of $[1, 9]$, called Saaty-scales with special interpretations which have been described in Table 5.6. For instance, displayed weights in the matrix (Eq. 5.23) means that a DSM competence has slightly higher importance than the PPS competence whereas DSM provides a much higher importance than VZ competence.

Values	Definition or Interpretation of the Value
1	Equal Importance
3	Moderate importance of one over another
5	Essential or strong importance
7	Very strong importance
9	Extreme importance
2,4,6,8	Intermediate values between the two adjacent judgments
$\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}, \frac{1}{9}$	Reciprocal Values

Table 5.6: Definition and interpretation of evaluation matrix values while pairwise comparisons in the AHP algorithm [Saaty, 1988]

In order to be prepared for a significant judgment through AHP, the sum of each column (Eq. 5.24), normalization of elements in each column, total of normalized rows (Eq. 5.25), normalized principal eigenvector (Eq. 5.26) are calculated [Saaty, 1988].

$$S_j = \sum_{i=1}^{n} c_{ij},\ j = 1, ..., n \tag{5.24}$$

where, S_j denotes sum of elements in the column j of the evaluation matrix and c_{ij} represents the competence value in the row i and column j of an evaluation matrix ($E_{n\times n}$) and n is the total number of prioritized competences.

$$c'_i = \sum_{j=1}^{n} \frac{c_{ij}}{S_j},\ i = 1, ..., n \tag{5.25}$$

$$\widehat{e}_i = \frac{c'_i}{n},\ i = 1, ..., n \tag{5.26}$$

$$c''_i = \sum_{j=1}^{n} c_{ij} \times \widehat{e}_i \tag{5.27}$$

The results of applying these equations in the exemplary Matrix (5.23) are summarized in Table 5.7. In fact, a normalized eigenvector ($\widehat{e}_i$) provides the final weights of the competences. Final result is indicated by c''_i, meaning that the highest value of c''_i indicates the most prioritized one. Meanwhile, the c''_i can be used for sorting (prioritizing) all alternatives based on their values. As

an example and according to the calculated eigenvector values in Table 5.7, *Distributed System Models and Enabling Technologies (DSM)* competence has the most of importance from talent's point of view and is defined as competence goal (identified competence gap). Any deviation of the weights from original competence goals expressed by the decision makers (i.e talents) depending on the inconsistency of pairwise comparisons. As a result and as suggested by Saaty, an inconsistency ratio ensures the quality of available information. Assignment and recommendation of the improvement potentials (CDPs) of desired competence goals is the next step.

Table 5.7: Resulted Values after normalization and eigenvector calculation in the AHP with Equations 5.24 to 5.27

Competence	DSM	PPS	WS	VZ	DSM	PPS	WS	VZ	c_i'	$\widehat{e_i}$	c_i''
DSM	1	3	7	9	0.63	0.64	0.68	0.5	2.45	0.61	2.56
PPS	0.33	1	2	5	0.21	0.21	0.19	0.28	0.89	0.22	0.91
WS	0.14	0.5	1	3	0.09	0.1	0.1	0.17	0.46	0.12	0.47
VZ	0.11	0.2	0.33	1	0.07	0.04	0.03	0.06	0.20	0.05	0.20
s_j	1.58	4.7	10.33	18	1	1	1	1	4.76	1	

5.3.2 Recommending Competence Improvement Solutions through Matching

Similar to other existing profile types such as JP and TP, a mathematical representation of the competence goal (CG) is based on the CKR model as showed in the (Eq. 5.28). The $n = 84$ in (Eq. 5.28) defines total number of competence categories according to the CKR model, and w_i indicates the weight (importance) of competence goal(s) to be improved, where $0 \leq w_i \leq 1$; $i = 1, ..., n$. In this way, multiple competences can also be defined by talents considering the fact that some programs affect (improve) more than one competence.

$$CG = \begin{bmatrix} w_1 & \dots & w_n \end{bmatrix}_{1 \times n} \tag{5.28}$$

Any recommendation of competence improvement solution requires matching of CDPs with the competence goal. The matching method relies on the clustering of CDPs and finding the closest CDP cluster to already defined competence goal. A similar method as described for matching of TPs with the desired JP as described in section 5.2 is used here. Therefore, target matching method of this section is described only with using a numerical example which avoids from repeating formulas and details described in section 5.2. Due to the fact that there are not too many generated CDPs in this research, a traditional k-means algorithm is used to cluster them. After clustering the CDPs, the closest CDP

cluster to already defined competence goal has to be found, whereas calculating an euclidean distance (Eq. 5.13) of defined competence goal with centroids of all CDP clusters facilitates this concern.

Similar to the matching of talents from already selected TP cluster with the target JP, selected best fit CDP cluster is the basis of the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method for identification of the best fitting CDP from the pool of existing ones. A decision matrix is like 5.14 and defines the requirements of a problem. In particular, a decision matrix of the $D_{m\times n}$ consists of the m CDPs and n competences. The positive and negative-ideal alternatives in this part demonstrate respective best and worst criteria expressions (e.g. competence goals) and are computed similarly using equations 5.18 and 5.19, accordingly. A given numerical example in the following clarifies more details of the approach.

Supervisors or tutors of competence improvement programs specify target competences that could be improved by participation in their leading solution. Suppose that a decision matrix summarized in Table 5.8 is achieved through stated methods. In addition to the decision matrix, a talent has defined his competence goal as $CG_{DSM,PPS,WS,VZ} = (0.41, 0.25, 0.19, 0.15)$ using AHP method in section 5.3.1. For the ease of clarification in this numerical example, a total dimension of 4 is considered for CG instead of 84 in the reality. Table 5.8 consists of weighted (multiplied by the competence goal) row-wise normalized decision matrix. The positive and negative virtual ideals can be formed as in Table 5.9. Using these ideals, even criteria such as costs or time to complete the course can be considered. In this case, the positive ideal should have the minimum costs and negative ideal involves the maximum costs.

Table 5.8: Decision Matrix and its weighted normalized result

Course \Tags	DSM	PPS	WS	VZ	DSM	PPS	WS	VZ
1	0.61	0.22	0.12	0.05	0.265	0.102	0.052	0.039
2	0.40	0.23	0.22	0.15	0.174	0.107	0.095	0.116
3	0.31	0.42	0.17	0.10	0.135	0.195	0.074	0.077
4	0.51	0.12	0.32	0.05	0.222	0.056	0.138	0.039

Table 5.9: Virtual Ideals

Ideal	DSM	PPS	WS	VZ
A^+	0.265	0.195	0.138	0.116
A^-	0.135	0.056	0.052	0.039

The Euclidean distance to the positive and negative ideals for this example are calculated in Table 5.10. A distance index is resulted from two virtual ideal indexes and support of the (Eq. 5.20), (Eq. 5.21) and (Eq. 5.22) for ranking of CDPs. According to these calculations, it can be concluded that the solution 3 fits the best to the talent's defined competence goal(s) and improves his competences significantly more than other ones.

Table 5.10: Distance Index and Ranking

Alternative	S_{i+}	S_{i-}	C_{i+}	Ranking
1	0.15	0.14	0.48	2
2	0.14	0.11	0.44	3
3	0.15	0.15	0.50	1
4	0.17	0.12	0.41	4

5.4 Conclusion of the Chapter

The main issue in this chapter refers to the matching algorithms. The matching targets two different perspectives: (1) matching of the TPs to already identified competence gaps in enterprises (JPs) and (2) matching of CDPs to TPs in order to provide competence improvement recommendations. The artificial data that is regenerated and retrieved from the web consists of 15 million talent data as well as 75,000 CDPs. The CDPs consist of courses, seminars, books, on-the-job-training and VET programs. In the case if a talent feels less competitiveness in one specific field and wants to improve it, such a recommendation facility supports him and works based on his preferences.

One of the problems associated with the case study of this research is that there are not real metrics to evaluate competences of scientists. Current existing metrics are good enough to get perspective about a person in general, but not about his real scientific competence. To this aim, proposed SCF and AIS metrics consider the variable of the time and accurately calculate the total number of citation counts as well as separate publications and associated statistics of one person in different fields. As a result, one person is proofed in his respective field rather than whole his scientific career with different research areas. Those metrics and further proposed assessment methods prepare requirements of the talent data according to the CKR model.

Regenerated domain specific CKR model data of talent is clustered using proposed EMRKM algorithm in order to filter the data and speed up the processing. This has been discussed in the next chapter through evaluation of the performance and accuracy. Due to the fact that the data has 84 dimensions, its visualization is

not possible. To have an insight about how clusters look like, the data has been visualized with respect to only two dimensions. This issue is already discussed in the next chapter.

Due to the fact that the CDP data is not associated with the large volumes of the data, it has been processed through traditional K-Means algorithm. This process provides recommendations for those under-qualified talents who aim to improve their competitiveness in one or more specific competence(s). Further visualization and analysis of the recommendation results based on the numerical example given in this chapter are provided in chapter 6. Evaluation results show that the proposed EMRKM algorithm is 47 times faster than traditional K-Means algorithm and 2.3 times faster than MR based K-Means algorithm provided in the Apache Mahout. The Apache Mahout K-Means has been tested with the same data and an associated table of the performance results is given in section 6.1.

Chapter 6

Evaluation of the Results

»If you tell people where to go, but not how to get there, you'll be amazed at the results. «

– George S. Patton

Evaluation of the results is in this work is strongly connected with preparation and production of large volumes of data, otherwise an efficiency of scalable algorithms cannot be proofed. The main goal of this chapter is to evaluate performance of proposed scalable clustering algorithms as well as proposed competence metrics in the case study (i.e. SCF and AIS). The evaluation covers an analysis of the results in terms of the processing time and accuracy. From the time perspective, the goal is to ensure that proposed EMRKM clustering algorithm is fast enough for large scale datasets. Consequently, it should be ensured that it provides correct and accurate results in which parallelization does not reduce accuracy of the clustering. In addition, an analysis of proposed metrics for the case study ensures efficiency of the concept and modeling of the domain specific CK.

The first and most important step in the test and evaluation of the proposed hybrid approach is to prepare an efficient test environment and configure required tools and services. To this aim, as detailed in Table 6.1, a virtual infrastructure consisting 5 virtual machines have been configured as a Hadoop cluster which consists of 1 namenode and 4 datanodes. The operating system used in all VMs is CentOS. Regenerated 15 million competence data has been first transfered to the HDFS. The Hadoop cluster is installed using Bigtop 1.1.0 which consists of the Hadoop 2.7.1 and runs OpenJDK 1.8.0. In addition Apache Mahout version 0.11.1 is also installed in the virtual infrastructure.

This chapter consists of two sections to covers performance analysis of the matching job and talent profiles in section 6.1 and matching of competence goal with CDPs in section 6.2.

6.1 Matching Job and Talent Profiles

As described earlier, proposed hybrid approach of matching talent and job profiles consists of three different steps. First, proposed competence metrics, SCF and AIS, are tested to see if they provide useful information about domain specific

Table 6.1: Specifications of virtual infrastructure used in the practical test and evaluation

	namenode	datanoe01	datanoe02	datanoe03	datanoe04
Number of cores	12	12	12	12	10
RAM	32 GB	24 GB	24 GB	24 GB	24 GB
System HDD	200 GB	200 GB	200 GB	200 GB	200 GB
HDFS	400 GB	400 GB	400 GB	400 GB	400 GB
OS (CentOS)	6.7	6.7	6.7	6.7	6.7

competences of talents and scientists. Furthermore, proposed clustering algorithm, EMRKM, is also tested with already regenerated 15 million talent data. Evaluation of the person-job-fit matching relies on the final step as of TOPSIS method. In most cases, the evaluation is done from two different perspectives: performance and accuracy. The performance test ensures efficiency and scalability of proposed method in terms of the time. Moreover, accuracy test proofs the quality of proposed method in finding the best-fit talent to already defined job description.

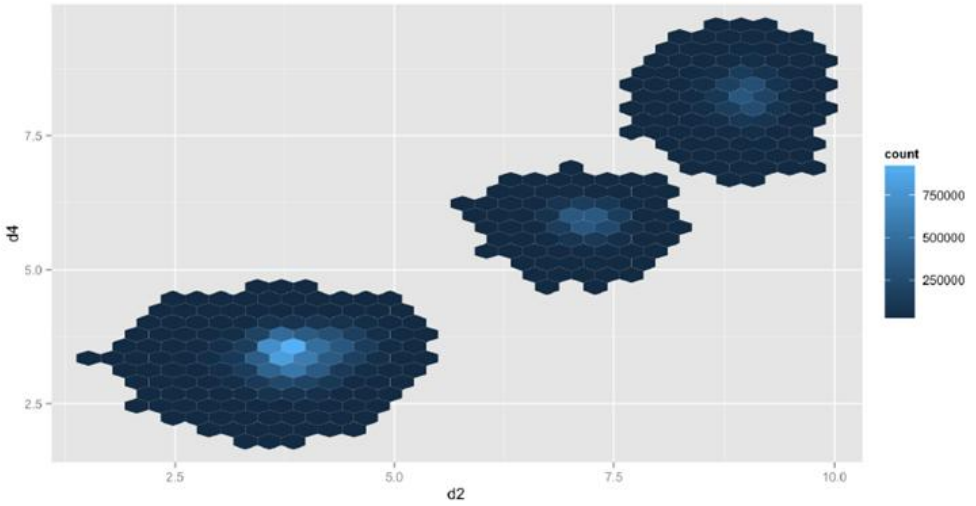


Figure 6.1: Visualization of the clustered 15 Million Talent Data using Multivariate Hexagonal Binning chart considering $C2$ ("Innovative" CK category) and $C4$ dimensions ("Social" CK category)

Competence metrics discussed in section 5.2 clearly show significant advantages in evaluation of one's job performance. For instance, being able to compare interdisciplinary scientists independent of each others as well as evaluating their performance over time is ensured through analysis of numeric results achieved from practical tests. Due to the quantifier in the acceleration formula (Eq. 5.4),

which weights the years in the timespan individually, a talent (scientist) is required to at least hold or improve citation counts year after year to retain his SCF metric. The AIS can be considered to be beneficial for newcomers, as the time a talent was active in one specific field is regarded in its calculation. This also means that more experienced authors get tougher penalties for not remaining active and well on a constant basis. Based on the assumption that one scientist acquires more knowledge in a field the longer he or she is active in the field, this seems fair.

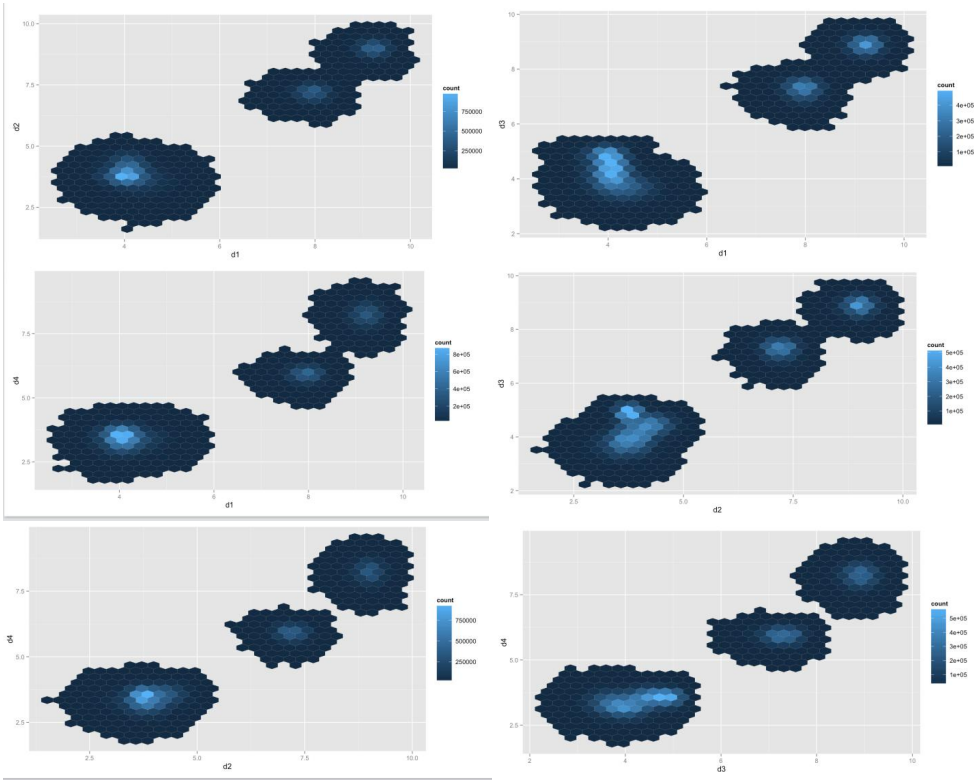


Figure 6.2: Visualization of clustered 15 Million Talent Data using Multivariate Hexagonal Binning chart, each time with considering two different dimensions of the level $l1$ competence categories, d1:Professional Competences, d2: Innovative Competences, d3: Personal Competences, d4: Social competences

Because of two reasons, the most important part to be tested is proposed EMRKM algorithm. First, the large scale 15 million talent data is statistically regenerated, so getting promising results specially in the clustering shows better statistical regeneration accuracy as well. Second, the tests should ensure quality of proposed EMRKM algorithm in terms of scalability and accuracy. The test data consists of 15 million talent data in one hand, and 84 dimensions on the

other hand which cannot be visualized. In order to graphically demonstrate the clustering results, two dimensions of the level $l1$ have been ignored to be able to visualize the data with Hexagonal Binning chart as showed in Figure 6.1. The Hexagonal Binning chart is an interesting visualization technique provided in the R language to visualize the density of data points in large scale datasets. Various colors of Hexagons represents different number of data points in the Hexagon [Carr et al., 2011]. The clustering results of 15 million points can be visualized using Hexagonal Binning.

Figure 6.1 shows clustering results according to the "Innovative ($C2$)" and "Personal ($C4$)" competence dimensions. Further visualization of the clustered data from different dimensions are showed in Figure 6.2. It should be stressed that this figure is just to show the clusters in this section and is not used anymore in the decision making processes. The original 200 talent data followed the $k = 3$ clustering results in already discussed clustering in chapter 4. As it is clear from figures 6.1 and 6.2, they follow the similar behavior and seem to be groups into three main large groups. The total number of clusters achieved as final result is $k = 9$.

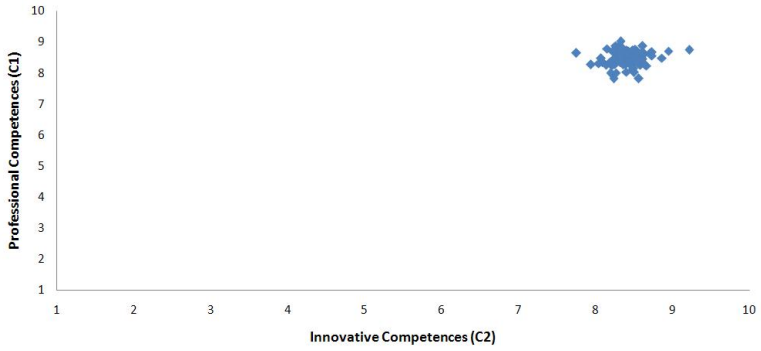


Figure 6.3: The Visualization of the selected TP cluster with the shortest Euclidean distance to the desired JP. This cluster consists of 80 talents.

For final evaluation of the matching TPs and desired JP, details of the final selected talent and Job Profile (JP) should be compared. Table 6.2 contains the numerical details of desired JP defined in the test phase. Due to the space limitation, values of the first 20 dimensions of the JP which are level $l1$ and $l2$ competences of the CKR model are provided in this table. The Euclidean distance between centroids of all TP clusters and this JP is computed.

The TP cluster with the shortest Euclidean distance represents the group of talents who are best fits to the desired job position. This cluster is showed in Figure 6.3 and for clear understanding of the details is further zoomed in Figure 6.4. The final selected cluster of the best-fit talents consists of 80 talent data. Inside this cluster and using TOPSIS, the best candidate can be found as it

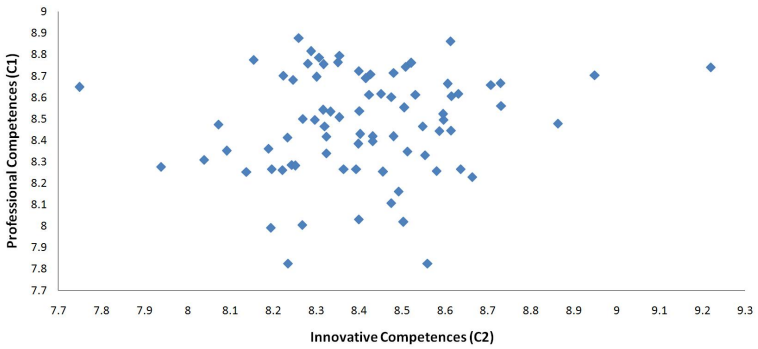


Figure 6.4: Zoomed overview of the selected TP cluster

is showed in Table 6.2.

Table 6.2: Comparing levels $l1$ and $l2$ Competence values of requested JP and Selected Best-fit Talent based on proposed hybrid approach

	c_1	c_2	c_3	c_4	$c_{1,1}$	$c_{1,2}$	$c_{1,3}$	$c_{1,4}$	$c_{2,1}$	$c_{2,2}$	$c_{2,3}$	$c_{2,4}$	$c_{3,1}$	$c_{3,2}$	$c_{3,3}$	$c_{3,4}$	$c_{4,1}$	$c_{4,2}$	$c_{4,3}$	$c_{4,4}$
Requested Job	8.581	8.257	8.216	8.400	8.544	8.771	8.571	8.437	8.506	8.414	8.002	8.105	7.849	8.133	8.351	8.530	7.895	8.227	8.409	9.069
Selected Talent	8.522	8.761	8.303	8.581	7.966	9.037	7.926	9.160	8.544	8.832	8.724	8.945	7.637	8.526	8.032	9.018	7.834	8.463	8.599	9.429

According to the results in Table 6.2, finally selected talent is the closed candidate to already desired job definitions. A careful analysis of required CK values in the JP and acquired CK values in the TP confirms the quality of selection and matching algorithm. It should be noted that the test of the 15 million talent data has been crashed in most of tries in testing on the single computer with traditional k-kmeans. It was just once successful with the clustering time of 57,528 seconds which is 47 times slower than proposed results through EMRKM in Table 6.3. The Apache Mahout is tested with the same data size of 15 million talents and resulted in the 2,3 times slower clustering time.

6.2 Recommending Competence Development Profiles

In order to test CDP recommendation method, results achieved from distributed Hadoop ecosystem have been compared with the results from a single instance cluster to check the efficiency of maps and reduces in the MR method. A recommendation method is deployed in the namenode and uses datanodes to read and write the inputs and outputs as well as MR jobs (processing operations). The performance evaluation in this section covers testing of the CDP clustering algorithms in terms of speed and accuracy. As described in section 5.3 and due to the nonavailability of real large scale CDP data, traditional and Apache Mahout based K-Means algorithm are used to cluster generated artificial CDP data. As a

Table 6.3: Comparing Clustering time of different algorithm including proposed EMRKM on on various scales of the data on the configured Hadoop virtual infrastructure

	SF-EAC				EMRKM			
	1st run	2nd run	3rd run	Average	1st run	2nd run	3rd run	Average
K	n.a.				n.a.			
Population Size	5				5			
Kmin	3				3			
Kmax	8				8			
Reference	0.850				0.850			
Maximum Generation	10				10			
Runtime for 200 T. data (s)	571	216	229	338.67	224	227	318	256.33
Runtime for 1,000 T. data (s)	220	222	320	254	225	218	230	224.33
Runtime for 15M T. data (s)	843	1636	1473	1317.33	1342	1283	1047	1224

result, their processing time is compared in this step.

In addition, it is key to ensure the scale of existing data in the test and also compare the effects of different volumes of the data in the processing time. In this regard, three different test scenarios consisting of 10,000, 50,000 and 100,000 course profiles have been generated. It should be noted that even the total number of 100,000 course profiles is not big enough to test real behavior of the system. An exception in the test with these data volumes is to reach significant improvement by the larger volumes of data. As a result, the MR processing should not be good enough for the case of 10,000 courses and should show better rates in the case of 100,000 course profiles. But this is still not the best solution for such small volume of the data. These 100,000 CDP data is artificially generated in this work using Java Script without any originality from real world case studies. It artificially simulates the similar behavior as real work CDPs.

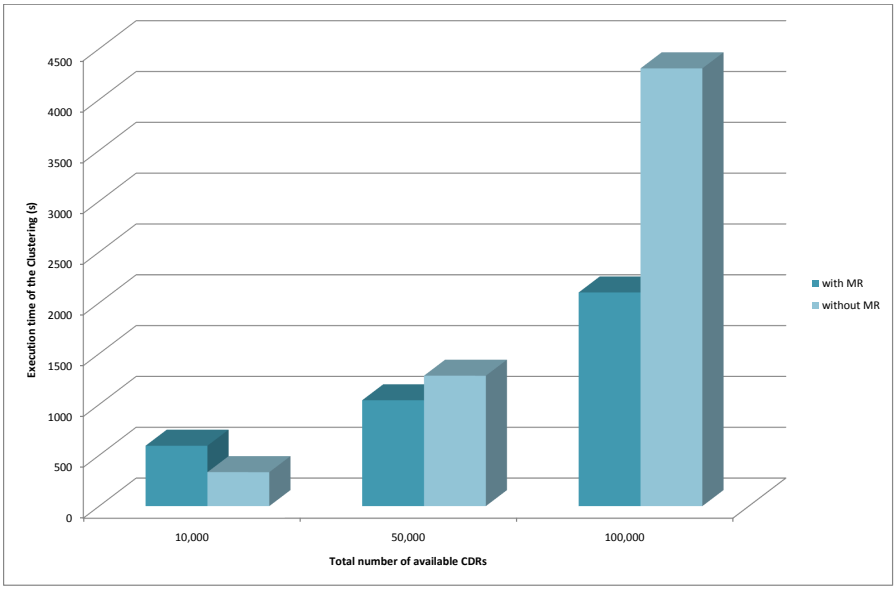
Table 6.4: Total Performance Measurements for Computing Operations in the Clustering of Competence Development Profiles

# of Courses	Data Size (TB)	Clustering with MR (hh:mm:ss)	Clustering without MR (hh:mm:ss)
10,000	0.2	00:09:54	00:05:34
50,000	0.85	00:17:21	00:21:23
100,000	1.5	00:35:02	01:11:47

Table 6.4 shows the performance evaluation in terms of the time (scalability) with artificial test data of those three scenarios. As it is clear from the table, clustering algorithm is tested with and without MR. In this way, the differences can be efficiently compared and discovered. It is clear from the stated processing times in the table that the MR clustering becomes faster and more efficient for large scale data. By scaling up the data from 10,000 to 100,000 CDPs (10 times more) the processing time becomes 5 times slower, in which the expectation should be 10 times more in the linear relationship between size and processing time.

It is also clear in Table 6.4 that not-MR algorithms are faster for small sized data rather than MR algorithms. Timings of both types of operations have been visualized in Figure 6.5.

Figure 6.5: Evaluation and comparison of the K-Means clustering time (seconds) of the CDPs with and without MR



In terms of the processing operations (i.e. clustering algorithms), a significant increase in the performance is achieved, specially by extending available cores/CPUs or memory. The conclusion is that parallelized algorithms show better performance just for large scale volumes of the data.

As stated earlier, in addition to the scalability test, there is also second perspective in testing of the proposed approach which is accuracy of clustering and recommendation algorithms. As showed in Figure 6.6, the recommendation of top 5 CDPs is based on already defined competence goal and using proposed method in section 5.3. The CDPs in the legend are presented in the recommended sequence. The CDP 3 consists the largest competence improvement according to the defined competence goal. The standardization of the values with normalization of the results avoids an implicit weighting based on economics of the scale. Through demonstration and application of the positive and negative ideal CDPs, the advantages of the CDP 3 becomes even more evident, since the largest increase in competences occur.

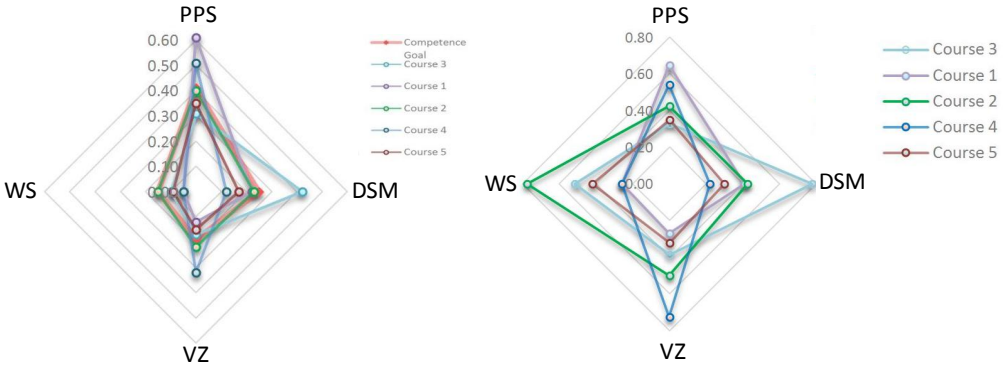


Figure 6.6: (A) Top 5 of the best recommendations for specific competence goal (B) Normalized values of Top 5 of the best Recommendations

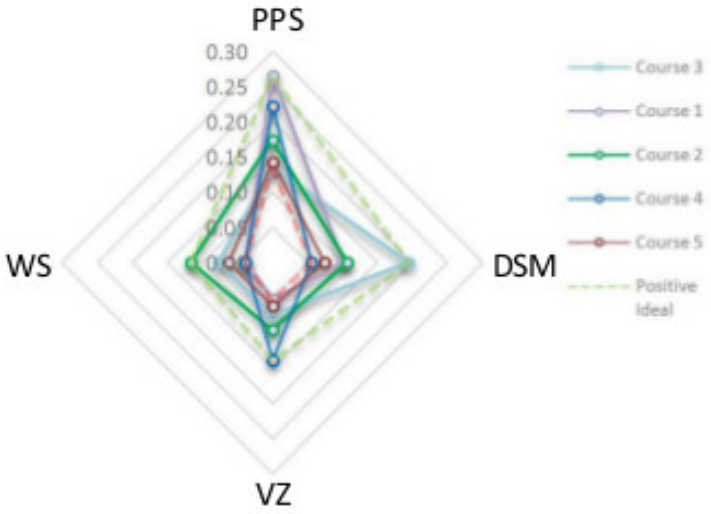


Figure 6.7: Weighted Normalized Top 5 of the best Recommendations

With the dataset of 100,000 CDPs, clustering recommendation method is tested once with MR and once more without MR. In each test scenario the processing time is recorded. This experience is repeated for 5 different competence goals (A-E). Table 6.4 shows the results of this test scenario which demonstrates significant improvement in the scalable clustering of the data across multi-node infrastructure with MR jobs. This test is repeated for 6 different competence goals being summarized as rows in Table 6.5. The achieved results show that the algorithms works similarly in term of runtime for all 6 different iterations with different competence goal definitions.

Table 6.5: Comparing K-Means clustering time (seconds) for 5 different competence goals (A-E) from the pool of 100,000 artificial CDP data with single and multiple-nodes cluster to identify the role of distributed (parallel) computing effects

	Multi-node Cluster					Single-node Cluster				
	A	B	C	D	E	A	B	C	D	E
1.	20.40	17.32	19.59	21.78	17.43	100.83	61.46	62.16	100.57	61.86
2.	17.82	16.89	19.59	19.71	14.25	97.48	60.14	70.63	103.04	62.25
3.	17.32	18.82	17.56	17.84	19.78	101.65	59.10	61.75	101.24	62.10
4.	23.70	17.09	16.63	18.61	16.21	100.31	64.01	65.18	100.10	63.09
5.	18.48	17.21	17.47	20.44	20.65	100.97	63.62	78.38	102.16	61.61
6.	19.54	17.47	18.17	19.68	17.66	100.25	61.67	67.62	101.42	63.76

Chapter 7

Discussion and Outlook

»The better way to predict the future is to create it.«

– Peter Drucker

This work is an interdisciplinary research covering multiple research directions and applications, specially big data analytics, competence management and job knowledge management as well as applied statistical analysis in the practice. Therefore, the main focus is not only limited to the computer sciences, but also on know-how towards disseminating the research results to real world applications such as job performance improvement and skill mis-match resolution through applied computer science methodologies. The current chapter provides a summary about research results, achievements, contribution to the knowledge, conclusion as well as identified future work potentials of this PhD work. The conclusion of research as well as further discussion on achieved outcomes is given in section 7.1. Moreover, section 7.2 provides a summary on potential future work. The future work forms continual research directions of this work for further research projects, student theses and products.

7.1 Conclusion and Discussion

The emergence of new technologies as well as strategic changes cause opening of new job positions and require further skills and CK in enterprises (e.g. competence gaps). Reassigning current employees or recruiting new candidates along with providing on-the-job-training to them solves this problem. But, there should be a clear understanding of what CK is available and what CK is required in an enterprise. In addition, enterprises should be aware of substantial methods and sources that could fill CK gaps. This can be achieved through efficient measurement methods and development of new job specific metrics. As a summary, a proper assessment method provides clear understanding of available CK and supports better DM, clustering and ranking of talents with respect to the requirements.

The current PhD thesis focuses on an efficient and scalable modeling, representation and analysis of HR competences using a hybrid approach deployed through big data analytics. Research results facilitate an efficient expert allocation process and CK improvement of employees (talents). The mathematical approach proposed in this dissertation provides adaptability to other case studies

or career areas as well. The HRM field has been selected as a case study of this research, since this area lacks the utilization of computerized DM methods and the HR data is also growing exponentially in this area. In addition, this domain is currently very important to enterprises from talent analytics and job performance improvement perspectives.

An intensive literature review in the frame of this research shows that early research efforts in the CKM area are mainly theoretical studies with traditional methods without utilization of an efficient and applied computer science, machine learning and mathematical algorithms. According to the literature analysis in chapter 2, relevant funded research projects lack utilization of big data in analyzing HRM data and consequently providing any scalable solution. This issue is even discussed in the recent literature as a main challenge of the HRM area. The focus of 86% of reviewed research projects is on one specific career domain with providing a solution which cannot be applied or extended to other areas. One reason of this issues is that these projects use mainly ontologies as a part of their competence management solution which results in developing fine-tuned domain specific results as target ontology of the domain.

According to the fundamental literature review about CKM area, any Competence Management System (CMS) should consist of at least competence (1) discovery, (2) assessment, (3) analysis, and (4) improvement recommendations. The competence discovery can be achieved through Competence Assessment (CA) as discussed in chapter 3. Similarly, competence analysis consists of competence gap identification, development of domain specific competence measures and metrics as well as competence matching and evaluation methods as discussed in chapter 5. The competence improvement recommendations can also be achieved through mathematical modeling of training programs followed by matching them to already identified competence gaps specially for under-qualified people as discussed in section 5.3.

The proposed CKR model in this work is inspired from European funded ComProFITS and CoMaVet research projects. The studies about professional, innovative and social competences are nearly identical with achievements in those projects. As an improvement to the competence models of those projects, a "personal" CK category is added in proposed CKR model. In order to identify and study the importance of the CKR model, the survey study has been conducted in this research. This survey study and its achievements after collecting the results from domain experts resulted in the fine tuning and slightly modification of the CKR model. The survey defines the weights of competences in the academic computer science career as RCK matrix which is partly summarized in Table 3.2.

In total, 186 participants from industry and academia have participated in this survey study. The participants are academics from computer science area and also industrials from HR or IT areas. According to the achieved results, 78% of HR experts believe that having a CKR model would improve the job knowledge analysis and performance improvement in their daily HRM processes. The feedback collected from IT experts resulted in specification of CK categories' weights for

computer science academic career as numeric values in the RCK matrix. In addition, 83% showed an interest to use new e-recruiting solutions that use scientific methods and machine learning algorithms in order to support them in their strategic and recruitment decision makings. This result identifies the importance of different assessment methods as well, meaning that self-assessment shouldn't have the same importance as multi-source assessment for specific competence.

The Competence Assessment (CA) consists of three different methods: (1) Multi-source assessment, (2) Self-assessment, and (3) Data retrieval from the web and other digital sources. Due to multilayer and hierarchical architecture of the CKR model, visualizing different levels provides different meanings and conclusions. In particular, visualization of the level $l1$ categories results in identification of collective competences of enterprises. In this way, enterprises could be aware of their general competence gaps (high level), but not in details. As a result, they could understand which competences (i.e. identified collective competence gaps) should be highlighted in their future job announcements. In order to clearly define the importance and weights of their required competences, they would need to have detailed visualization of level $l2$ subcategories, specially for those identified ones in the level $l1$ visualization. Finally, the level $l3$ sub-subcategories are proper for setting up assessments and also clustering and prioritization of employees.

According to statistical studies made in this research, the 200 real talent data is uniformly distributed. To test the uniform distribution, the fitness test has been employed in order to estimate its parameters as well. In testing the initial hypothesis as uniform distribution of the data, a significance level of 0.05 as well as Pearson's chi-square test are used. With this significance level, it shows the success rate of 94.27%. Similarly, significance level of 0.01 results in the success rate of 97.92 %. The conclusion based on significance level achievements is that the uniform distribution simulates enough artificial competence data needed for the test and evaluations of the big data implementation. Random numbers of uniform distribution have also been produced in this regard. Statistical analysis used in this chapter resulted in the regeneration of 15 million talent data. It represents sufficient data volume to test proposed algorithms in chapter 5 and proof the results in chapter 6.

Three types of competence related profiles are defined in this dissertation as of: (1) Talent Profile (TP), Job Profile (JP), and Competence Development Profile (CDP). These profiles associate with three different matching problems. Matching of TPs with the JP results in the identification of the best-fit talent to already defined open position. Accordingly, matching of CDPs with already identified competence gaps through assessments or matching of TPs with JPs concludes in recommending competence improvement solutions specially for under-qualified people. For prioritization of the competence goal, an AHP algorithm is used in this work. Important and key required trainings for specific job categories as well as domain specific job knowledge can be achieved through matching of the CDPs with JPs.

The most important part is matching of TPs to job positions (JPs) which

contributes to the RQ 1 (Skill mismatch) challenge in real world applications as well. In order to provide job knowledge qualitative evaluation metrics fr the case study of this work, two bibliographic factors of SCF and AIS are developed. Text mining techniques specially Part of Speech Tagging method, lemmatization as well as stpwords are used in order to prepare inputs of these two metrics. These text mining techniques analyze publication titles of the scientists. The main highlight of these two metrics in comparison to currently available ones such as h-index or i10-index is their intensive dependency on the activeness and competences of a scientist rather than being just quantitative measure. In particular, these metrics respect the activities and contribution of the person in his own field and also measure his competence and growth level with respect to the growth of his field of research. In this way, newcomers as well as scientists in not very popular fields will have the chance to show their scientific competences and compare them with well-known ones.

In order to efficiently prioritize job applicants and talents, an evolutionary MR based K-Means clustering algorithm is used. This algorithm shows better performance achievements in the clustering quality as well as computing power and speed. The algorithm uses a novel approach of simplified silhouette value measures in order to cover distribution of data across multiple nodes as mappers. This algorithm uses three different mutation operators as of elimination, merge and split and also checks in each generation all of those three operators and takes the best operator. The quality of clustering becomes significantly better and closer to final solution by each generation. In addition, the algorithm is faster than similar solutions, because by each iteration of the algorithms, the results converge faster. Proposed EMRKM algorithm showed 47 times faster speed in comparison to the traditional K-Means clustering as well as 2.3 faster speed than Apache Mahout provided K-Means clustering.

Computing an Euclidean distance between centroids of talent clusters and desired JP, under-qualified talents cluster can be easily identified. As a result, those under-qualified talents based on the desired job description may plan to improve their competitiveness which is referred as competence goal. A competence goal can be reached by participation in the programs and trainings such as VET, on-the-job training or seminars. Matching of the competence goal to the competence improvement potentials is handled by using traditional K-Means algorithm due to the fact that there was not large volumes of the course data in this research. Identification and prioritization of competence goal uses AHP method. As soon as the competence goal is defined and all competence development profiles are clustered, the best-fit CDP cluster to already identified competence gaol(s) is selected using an Euclidean distance of defined competence goal and centroids of CDP clusters. In the selected CDP cluster, TOPSIS method searches for the best-fit recommendation based on defined competence gaps. This method is not based on the MR and uses traditional methods due to not having large volumes of the data in the research.

As it is issued in chapter 5, analyzing the publication titles of scientists does

not extract accurately research areas of scientists, but it is sufficient for the focus of this research. In this regard, further research efforts and realization of better text mining methods as well as analyzing the content of publications towards topic detection are mandatory in order to identify accurately research areas of scientists. It should be noted that the best interpretation of the SCF can be achieved by evaluating and comparing the results of more than three years. This supports giving an opportunity to absolute newcomers to the science to develop and show their competence in growing up in the scientific career. An automatic data retrieval from the web should be further researched and integrated in the system. It is currently nearly impossible to integrate for instance Google Scholar results through the APIs because of the technical and data privacy concerns. In the case any released API from for instance Google Scholar in the future, their experiences and results can be easily integrated in this research.

Moreover, new metrics and competence assessment methods should be developed to other domains such as nurses and politicians. Providing generic metrics that cover many disciplines may improve the quality of achievements of this research. Additionally, an organizational issues of exams such as how to manage an on-line test in the self-assessment method are not addressed in this dissertation. It should also be noted that for the specific case study of this research, current analysis of publication titles doesn't exactly result in identification of the research areas of scientists. As described, new and novel topic detection and publication content analysis methods improves significantly the quality of identifying research areas of scientists. In the frame of this research, a total number of 75,000 CDP data has been prepared as artificial data using scripts without any use of real world data. As a final issue, any test of recommendation method with real world data ensures the quality of proposed hybrid approach.

7.2 Future Work

The contributions of this work open up new directions in the research and practice. Proposed algorithms and scientific methodology have been applied in the case study of computer science academic career. Profiling of competence development sources such as seminars, courses and VET programs delivers input data and information for recommendations towards identified competence gaps. An artificial data has been used in this regard. Competence assessment in the frame of current work relies on the 360-degree feedback and self-assessment methods. In order to retrieve competence data from digital sources such as the web, bibliographic data has been collected from DBLP and AMiner.

It is strongly recommended to integrate proposed algorithms and methodology to further case studies such as competence assessment in the automotive industry or education career. This ensures the generalization of the concept in a wide variety of sectors. To this aim, one should start from modification of the CKR model and assessment methods to one specific case study and collect outcomes of the analytics

specially from evolutionary K-Means clustering. A generic competence metrics such as SCF or AIS have to be developed in order to correctly assess talents' domain specific competences. For instance, SCF and AIS as Scientometrics which are discussed in chapter 5 do not model competences of nurses or politicians. One solution in this regard is to to employ social media streaming such as tweets of politicians in order to retrieve domain specific content associated to talents and use text mining methods such as sentiment analysis to compute interesting metrics [Bohlouli et al., 2015b]. Such methods should result in the metrics such as likes, dislikes or any other metric which shows the contextual and professional measures about the CK of talents.

Furthermore, the CKR model has to be specified for further sectors. The best is to conduct a survey study to involve domain experts and identify the key factors and domain specific competences. Such a survey study should respect geographical distribution of experts as well as their proficiency level. Careful analysis of such a survey results specifies the model with more professional and domain specific competence categories. This will also contribute in modeling of the domain specific competence metrics. Additionally, representation of the domain specific competences using proposed XML model will be benefited. In addition, experiences in testing Hadoop ecosystem showed that the I/O operations specially uploading the data to HDFS is too slow. Further studies and solutions in order to improve and speedup I/O operations in the Hadoop ecosystem is recommended.

Web data mining followed by text analytics (mining) is a key future work of this thesis. In this regard, streaming of proper social media, identification of relevant job knowledge data sources as well as knowledge discovery techniques through unstructured data are key issues. Text mining algorithms should ensure job knowledge discovery from text (KDT) from sources such as tweets and job descriptions from the web. Crawling for job and topic related data from the web and job (career) knowledge formation from job descriptions are main issues in this regard. This is not limited to the job descriptions, but also profiling of competence development sources. This requires a model in which autonomously extracts important competence related information from course descriptions and integrates them with representation techniques in the system.

As stated earlier, some parts of this work such as recommendation component have been tested with artificial datasets. A need to test and evaluate with real big data sources and streams is important to be classified as a future work. It is also recommended to test further clustering algorithms such as spectral clustering and compare the results. Integration and development of further assessment methods will improve current studies of the research from social science and practical perspectives. Text mining and social media integration may also needed in this regard. In addition and due to a generalized goal of this research as discussed in the conclusion, further conception and research for matching of service/product seekers with providers is recommended.

Further research on proposing novel text mining techniques to discover research fields of scientists from their publications is recommended as a future work of

this PhD. It is not limited to the text analytics, but also on some further DBs or APIs that restrieve such information from existing solutions such as Google Scholar or IEEE Xplore. Developing further data retrieval methods from web as well as social media streaming in order to enrich TPs and improve the accuracy of the CK data about talents in their profiles is also defined as future work of this research. In this regard, any proposed method should focus on the autonomous data retrieval method.

Further research in order to propose new methods to reduce the number of pairwise comparisons while prioritizing competence goals using AHP is an interesting research contribution to this work. In this way, the AHP method can also be used for prioritization of 84 competences while setting up the JPs. Currently, the HCV method is used. When the data retrieval and accordingly any decision support such as recruitment decisions are discussed, a confidence factor is a must. Such a factor indicates whether the judgments about specific talent and based on the size, quality and relevance of the data are feasible or not?! Any further efforts in developing such a confidence factor and similar methods is strongly recommended as future work of this research. Moreover, further research in clustering and analyzing JPs in order to identify the most common jobs, their configurations and required competences as well as considering their configurations in providing CDP recommendations to job seekers will improve their competitiveness.

Bibliography

Daniel J. Abadi. Column Stores for Wide and Sparse Data. In *CIDR 2007, Third Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 7-10, 2007, Online Proceedings*, pages 292–297, 2007.

Chunk Allen and Lon Pilot. HR-XML: Enabling Pervasive HR e-Business. In *XML Europe 2001, International Congress Centrum (ICC)*, 2001.

Mehrdad Amiri, Mostafa Zandieh, Roya Soltani, and Behnam Vahdani. A Hybrid Multi-criteria Decision-making Model for Firms Competence Evaluation. *Expert Systems with Applications*, 36(10):12314–12322, December 2009.

Janet L Bailey. Non-technical skills for success in a technical world. *International Journal of Business and Social Science*, 5(4):1–10, March 2014.

Peter Baladi. Knowledge and Competence Management: Ericsson Business Consulting. *Business Strategy Review*, 10(4):20–28, 1999.

Raja C. Bandaranayake. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. *Medical Teacher*, 30(9-10), 2008.

Kyle Banker. *MongoDB in Action*. Manning Publications Co., Shelter Island, NY, 2012.

Jerry Banks, editor. *Handbook of Simulation: Principles, Methodology, Advances, Application and Practice*. Wiley, 1998.

Jerry Banks, John S. Carson, Barry L. Nelson, and David M. Nicol. *Discrete-Event Systems Simulation*. Pearson, 5th edition, 2010.

Patrik Berander and Anneliese Amschler Andrews. Requirements Prioritization. In Aybüke Aurum and Claes Wohlin, editors, *Engineering and Managing Software Requirements*, pages 69–94. Berlin: Springer Verlag, 2005.

Josh Bersin. Big Data in HR: Building a Competitive Talent Analytics Function: The Four Stages of Maturity. Research report, Bersin by Deloitte, April 2012.

Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker, editors, *Data Analysis, Machine Learning and Applications*, pages 319–326. Springer-Verlag Berlin Heidelberg, 2008.

Sigrid Blömeke and Olga Zlatkin-Troitschanskaia, editors. *The German funding initiative "Modeling and Measuring Competencies in Higher Education": 23 research projects on engineering, economics and social sciences, education and generic skills of higher education students. (KoKoHs Working Papers, 3)*. Humboldt University & Johannes Gutenberg University, Berlin and Mainz, 2013.

M Bohlouli, F Merges, and M Fathi. A cloud-based conceptual framework for knowledge integration in distributed enterprises. In *Proceedings of International Conference on Electro/information Technology*, 2012a.

Mahdi Bohlouli and Morteza Analoui. Gid-HPA: Predicting resurce requirements of a job in the grid computing environment. *Academy World of Science, Engineering and Technology*, 21.

Mahdi Bohlouli, Patrick Uhr, Fabian Merges, Sanaz Mohammad Hassani, and Madjid Fathi. Practical approach of knowledge management in medical science. In *IKE*, pages 79–84, 2010.

Mahdi Bohlouli, Alexander Holland, and Madjid Fathi. Knowledge integration of collaborative product design using cloud computing infrastructure. In *Electro/Information Technology (EIT), 2011 IEEE International Conference on*, pages 1–8. IEEE, 2011.

Mahdi Bohlouli, Fazel Ansari, and Madjid Fathi. Design and realization of competence profiling tool for effective selection of professionals in maintenance management. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2195–2200. IEEE, 2012b.

Mahdi Bohlouli, Fazel Ansari, Yogesh Patel, Madjid Fathi, Miguel Loitxate, and Lefteris Angelis. Towards Analytical Evaluation of Professional Competences in Human Resource Management. In *the 39th Annual Conference of the IEEE Industrial Electronics Society (IECON-2013)*, Vienna, Austria, November 2013a.

Mahdi Bohlouli, Frank Schulz, Lefteris Angelis, David Pahor, Ivona Brandic, David Atlan, and Rosemary Tate. Towards an Integrated Platform for Big Data Analysis. In Madjid Fathi, editor, *Integration of Practice-oriented Knowledge Technology: Trends and Prospective*, pages 47–56. Springer Berlin Heidelberg, 2013b.

Mahdi Bohlouli, Fabian Merges, and Madjid Fathi. Knowledge integration of distributed enterprises using cloud based big data analytics. In *IEEE International Conference on Electro/Information Technology*, pages 612–617. IEEE, 2014.

Mahdi Bohlouli, Fazel Ansari, George Kakarontzas, and Lefteris Angelis. An adaptive model for competences assessment of it professionals. In *Integrated Systems: Innovations and Applications*, pages 91–110. Springer, 2015a.

Mahdi Bohlouli, Jens Dalter, Mareike Dornhoefer, Johannes Zenkert, and Madjid Fathi. Knowledge Discovery from Social Media using Big Data provided Sentiment Analysis (SoMABiT). *Journal of Information Science (IF=1.087)*, 41(6): 779–798, December 2015b.

Gerhard Bohm and Günter Zech. *Introduction to statistics and data analysis for physicists*. DESY, 2010.

André B. Bondi. Characteristics of scalability and their impact on performance. In *Proceedings of the second international workshop on Software andperformance-WOSP*, pages 195–203, New York, NY, USA, 2000. ACM.

Nick Boreham. A Theory of Collective Competence: Challenging The Neo-Liberal Individualisation of Performance at Work. *British Journal of Educational Studies*, 52(1):5–17, March 2004.

Josiah L. Carlson. *Redis in Action*. Manning Publications Co., 2013.

Dan Carr, Nicholas Lewin-Koh, and Martin Maechler. hexbin: Hexagonal binning routines. *R package version*, 1260, 2011.

Chartered Institute of Personnel & Development. Research Report: Talent Analytics and big data - the challenge for HR. Technical Report 6368, CIPD in partnership with Oracle Human Capital Management Cloud, November 2013.

Cheng-Tao Chu, Sang Kyun Kim, Yi-An Lin, YuanYuan Yu, Gary Bradski, Andrew Y. Ng, and Kunle Olukotun. Map-Reduce for Machine Learning on Multicore. In *Twentieth Annual Conference on. Neural Information Processing Systems (NIPS)*, 2006.

John H Conway and Richard Guy. *The book of numbers*. Springer Science & Business Media, 2012.

Jens Dalter. SoMABiT: Social Media Analysis using Big Data Technology. Master's thesis, Institue of Knowledge Based Systems and Knowledge Management, University of Siegen, September 2014.

Thomas H. Davenport, Jeanne Harris, and Jeremy Shapiro. Competing on Talent Analytics. *Harvard Business Review*, 88(10):52–58, October 2010.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *Sixth Symposium on Operating System Design and Implementation (OSDI)*, pages 137–149, San Francisco, CA, December 2004. USENIX Association.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1):107–113, January 2008.

Francoise Delamare Le Deist and Jonathan Winterton. What is Competence? *Human Resource Development International*, 8(1):27–46, March 2005.

Gary Dessler. *Human Resource Management*. Prentice Hall, 14 edition, Fabruray 2015.

Ching-Shen James Dong and Ananth Srinivasan. Agent-enabled Service-oriented Decision Support Systems. *Decision Support Systems*, 55(1):364–373, April 2013.

Fotis Draganidis and Gregoris Mentzas. Competency based management: a review of systems and approaches. *Information Management & Computer Security*, 14 (1):51–64, 2006.

Nicky Dries, Richard D. Cotton, Silvia Bagdadli, and Manoela Ziebell de Oliveira. HR Directors' Understanding of 'Talent': A Cross-Cultural Study. In Akram Al Ariss, editor, *Global Talent Management: Challenges, Strategies, and Opportunities*, pages 15–28. Springer International Publishing, 2014.

Edd Dumbill. The SMAQ stack for big data. In Mac Slocum, editor, *Big Data Now*, pages 16–29. O'Reilly Media, first edition, September 2011.

Patrick Dunleavy. *Authoring a PhD: How to plan, draft, write and finish a doctoral thesis or dissertation*. Palgrave Macmillan, 2003.

Agoston E Eiben and James E Smith. *Introduction to evolutionary computing*, volume 53. Springer, 2003.

Robert W. Eichinger and Michael M. Lombardo. Patterns of Rater Accuracy in 360-Degree Feedback. *Human Resource Planning Journal*, 27(4):23–25, 2004.

Tanja Joan Eiler. Competence Management for IT Companies. Technical report, Institue of Knowledge Based Systems and Knowledge Management, University of Siegen, January 2015.

Hesham El-Rewini and Mostafa Abd-El-Barr. *Advanced Computer Architecture and Parallel Processing (Wiley Series on Parallel and Distributed Computing)*. Wiley-Interscience, 2005.

Michelle R. Ennis. Competency models: a review of the literature and the role of the employment and training administration (ETA). Technical report, Office of Policy Development and Research Employment and Training Administration, US. Department of Labor, January 2008.

eQuest Big Data for Human Resources. Big Data: HR's Golden Opportunity Arrives. Technical report, eQuest Headquarters, 2010.

Eurpean Commission. The European Qualifications Framework for Lifelong Learning (EQF). Technical report, Office for Official Publications of the European Communities, Luxembourg, 2008.

Sarah S. Fallaw and Tracy M. Kantrowitz. 2013 Global Assessment Trends Report. Technical report, SHL Talent Measurement Solutions, 2013.

Christiane Fellbaum. WordNet and wordnets. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford, 2005. Elsevier.

Jackie Fenn and Hung LeHong. Hype Cycle for Emerging Technologies. Technical Report G00215650, Gartner, Inc, July 2011.

Sara Flisi, Valentina Goglio, Elena Claudia Meroni, Margarida Rodrigues, and Esperanza Vera-Toscano. Measuring Occupational Mismatch: Overeducation and Overskill in Europe—Evidence from PIAAC. *Social Indicators Research*, pages 1–39, 2016.

Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications*, volume 20. Siam, 2007.

Kemilly Dearo Garcia and Murilo Coelho Naldi. Multiple Parallel MapReduce k-Means Clustering with Validation and Selection. In *2014 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 432–437, October 2014.

Michael Gebel and Johannes Giesecke. Labor Market Flexibility and Inequality: The Changing Skill-Based Temporary Employment and Unemployment Risks in Europe. *Journal of Social Forces*, 90(1):17–39, 2011.

Sanjay Ghemawat and Leung Shun-Tak Gobioff, Howard. The Google file system. *Operating Systems Review*, 37(5):29–43, December 2003.

Thomas F. Gilbert. *Human competence: Engineering worthy performance.* McGraw-Hill, New York, 1978.

I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series and products.* Academic Press, 7th edition, 2007.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations Journal*, 11(1):10–18, 2009.

Joy F. Hazucha, Sarah A. Hezlett, and Robert J. Schneider. The impact of 360-degree feedback on management skills development. *Journal of Human Resource Management*, 32(2):325–351, 1993.

S. Holtzman. *Intelligent Decision Systems.* Addison-Wesley, Reading, MA, 1989.

Eli Hustad, Bjorn Erik Munkvold, and Brigitte Vigemyr Moll. Using IT for Strategic Competence Management: Potential Benefits and Challenges. In Timo Leino, Timo Saarinen, and Stefan Klein, editors, *ECIS*, pages 801–812, 2004.

Ching-Lai Hwang and Kwangsun Yoon. *Multiple attribute decision making: methods and applications a state-of-the-art survey*, volume 186. Springer Science & Business Media, 2012.

IEEE. IEEE Standard for Learning Technology - Data Model for Reusable Competency Definitions. *Std 1484.20.1-2007*, pages C1–26, January 2008.

Bob Ippolito. Drop ACID and think about Data, March 2009.

Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

Robert E Jensen. A dynamic programming algorithm for cluster analysis. *Operations Research*, 17(6):1034–1057, 1969.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous univariate distributions*, volume 2. Wiley, 2nd edition, 1994a.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous univariate distributions*, volume 1. Wiley, 2nd edition, 1994b.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

Amin Keshavarzi, Abolfazl T Haghighat, and Mahdi Bohlouli. Research challenges and prospective business impacts of cloud computing: a survey. In *Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2013 IEEE 7th International Conference on*, volume 2, pages 731–736. IEEE, 2013.

Barbara Kitchenham and Stuart Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report, 2007.

Donald E Knuth. *The Art of Computer Programing – Volume 2 / Seminumerical Algorithms*, volume 2. Addison-Wesley, 2nd edition, 1998.

Pierre L'Ecuyer. Uniform random number generation. *Annals of Operations Research*, 53(1):77–120, 1994.

Tobias Ley, Armin Ulbrich, Peter Scheir, Stefanie N. Lindstaedt, Barbara Kump, and Dietrich Albert. Modeling competencies for supporting work-integrated learning in knowledge work. *Journal of Knowledge Management*, 12(6):31–47, 2008.

Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

Rikard Lindgren, Ola Henfridsson, and Ulrike Schultze. Design Principles for Competence Management Systems: A Synthesis of an Action Research Study. *MIS Q.*, 28(3):435–472, September 2004.

Mike Loukides. What is data science? In Mac Slocum, editor, *Big Data Now*, pages 1–15. O'Reilly Media, first edition, September 2011.

Craig C. Lundberg. Planning the Executive Development Program. *California Management Review*, 15(1):pp. 10–15, 1972.

James B. MacQueen. Some Methods for Classification and Analysis of MultiVariate Observations. In L. M. Le Cam and J. Neyman, editors, *Procedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute (MGI), 2011.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.*, 19(2):313–330, June 1993.

David McClelland. Testing for competence rather than intelligence. *American Psychologist*, 1973.

Alok Mishra and Ibrahim Akman. Information Technology in Human Resource Management: An Empirical Assessment. *Public Personnel Management*, 39(3), 2010.

N Mittas, G Kakarontzas, M Bohlouli, L Angelis, I Stamelos, and M Fathi. Comprofits: A web-based platform for human resources competence assessment. In *Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on*, pages 1–6. IEEE, 2015.

Murilo Coelho Naldi and Ricardo JGB Campello. Comparison of distributed evolutionary k-means clustering algorithms. *Neurocomputing*, 163:78–93, 2015.

Murilo Coelho Naldi and Ricardo José Gabrielli Barreto Campello. Evolutionary k-means for distributed data sets. *Neurocomputing*, 127:30–42, 2014.

Murilo Coelho Naldi, Ricardo JGB Campello, Eduardo R Hruschka, and ACPLF Carvalho. Efficiency issues of evolutionary k-means. *Applied Soft Computing*, 11(2):1938–1952, 2011.

Nishant Neeraj. *Mastering Apache Cassandra*. O'Reilly Media, second edition, March 2015.

John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham. RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–20, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

NIST. *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST (National Institute of Standards and Technology) and SEMATECH, 2013.

Gilberto de Viana Oliveira and Murilo Coelho Nald. Scalable Fast Evolutionary k-Means Clustering. In *2015 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 74–79. IEEE, 2015.

Sean Owen, Robin Anil, Ted Dunning, and Ellen Friedman. *Mahout in Action*. Manning Publications Co., October 2011.

Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2):3336–3341, 2009.

Hae-Sang Park, Jong-Seok Lee, and Chi-Hyuck Jun. A K-means-like Algorithm for K-medoids Clustering and Its Performance. *Proceedings of ICCIE*, pages 102–117, 2006.

Jukka Piirto, Annika Johansson, and Helene Strandell, editors. *Key Figures on Europe: 2013 Digest of the Online Eurostat Yearbook*. Pocketbooks / Eurostat. Publications Office of the European Union, Luxembourg, 2013.

Coimbatore K Prahalad and Gary Hamel. The core competence of the corporation. *Harvard Business Review*, 68(3):79–91, 1990.

R Development Core Team. R: A Language and Environment for Statistical Computing. Technical report, the R Foundation for Statistical Computing, Vienna, Austria, 2011.

Marcel M. Robles. Executive Perceptions of the Top 10 Soft Skills Needed in Today's Workplace. *Business and Professional Communication Quarterly*, 75 (4):453–465, December 2012.

Piotr Ronkowski. Labour market policy expenditure and the structure of unemployment. Technical Report KS-SF-13-031-EN-N, European Commission, 2013.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Przemyslaw Rozewski. Discussion of the Competence Management Models for Education Context. In Gordan Jezic, Mario Kusek, Ngoc Thanh Nguyen, Robert J. Howlett, and Lakhmi C. Jain, editors, *KES-AMSTA*, volume 7327 of *Lecture Notes in Computer Science*, pages 604–613. Springer, 2012.

Przemyslaw Rozewski and Bartlomiej Malachowski. Competence Management in Knowledge-Based Organisation: Case Study Based on Higher Education Organisation. In Dimitris Karagiannis and Zhi Jin, editors, *Knowledge Science, Engineering and Management*, volume 5914 of *Lecture Notes in Computer Science*, pages 358–369. Springer Berlin Heidelberg, 2009.

Przemyslaw Rozewski, Emma Kusztina, Ryszard Tadeusiewicz, and Oleg Zaikin. Methods and Algorithms for Competence Management. In *Intelligent Open Learning Systems*, volume 22 of *Intelligent Systems Reference Library*, pages 151–176. Springer Berlin Heidelberg, 2011.

Philip Russom. Big Data Analytics. Technical report, TDWI Research, 2011.

Thomas L. Saaty. *Mathematical Models for Decision Support*, chapter What is the Analytic Hierarchy Process?, pages 109–121. Springer Berlin Heidelberg, Berlin, Heidelberg, 1988. ISBN 978-3-642-83555-1. doi: 10.1007/978-3-642-83555-1_5.

Jorgen Sandberg. Understanding Human Competence at work: an interpretative approach. *Academy of Management Journal*, 43(1):9–25, 2000.

Marc Seeger. Key-Value stores: a practical overview. Technical report, September 2009.

Richard J. Shavelson. *An Approach to Testing & Modeling Competence*. Sense Publishers, 2013.

Lucie Skorkovská. *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings*, chapter Application of Lemmatization and Summarization Methods in Topic Identification Module for Large Scale Language Modeling Data Filtering, pages 191–198. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

Peter Sloane, Kostas Mavromaras, Nigel O'Leary, Seamus McGuinness, and Philip J O'Connell. The Skill Matching Challenge: Analysing Skill Mismatch and Policy implications. Technical report, Publications office of the European Union, 2010.

Scott A. Snell, Donna Stueber, and David P. Follow Lepak. Virtual HR Departments: Getting Out of the Middle. Technical report, Ithaca, NY: Cornell University, School of Industrial and Labor Relations, Center for Advanced Human Resource Studies (CAHRS Working Paper #01-08), 2001.

Nicole Sprafke and Uta Wilkens. the 30th European Group for Organizational Studies (EGOS) Colloquium. In *Examining dynamic capabilities with an actor-centered measurement approach and instrument*, Rotterdam, Netherlands, July 2015.

Thomas Süsse. Lifecycle Management mit Planspielen interaktiv erfahren - Ansätze zur Konzeption eines Planspiels für Product-Service Systems, September 2013.

Thomas Süße and Uta Wilkens. Preparing Individuals for the Demands of PSS Work Environments through a Game-based Community Approach – Design and Evaluation of a Learning Scenario. In Hoda ElMaraghy, editor, *Proceedings of the 6th CIRP Conference on Industrial Product-Service Systems*, volume 16, pages 271–276. Elsevier Inc., 2014.

Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. Introduction to data mining. Pearson Publishing, 2005.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In *In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990–998. ACM SIGKDD, 2008.

Tina Teodorescu. Competence versus competency: What is the difference? *Journal of Performance Improvement*, 45(10):27–30, December 2006.

S. Theodoridis and K Koutroumbas. *Pattern Recognition*. Academic Press, fourth edition, 2009.

Sergios Theodoridis, Aggelos Pikrakis, Konstantinos Koutroumbas, and Dionisis Cavouras. *Introduction to Pattern Recognition: A Matlab Approach: A Matlab Approach*. Academic Press, 2010.

Philippe Tissot, editor. *Terminology of European education and training policy*. Office for Official Publications of the Europ. Communities, Luxembourg, 2008.

Dennis Wackerly, William Mendenhall, and Richard Scheaffer. *Mathematical statistics with applications*. Cengage Learning, 7th edition, 2007.

Pete Warden. *Big Data Glossary*. O'Reilly Media, 2011.

Eric W. Weisstein. Bell Number. MathWorld – A Wolfram Web Resource., 2015. URL `http://mathworld.wolfram.com/BellNumber.html`.

Robert W. White. Motivation reconsidered: The concept of competence. *Psychological Review*, 66(5):297–333, 1959.

Tom White. *Hadoop: The Definitive Guide*. O'Reilly Media, May 2009.

Wolfram Research Inc. *Matematica 10 Language and System Documentation Center*, version 10 edition, 2014.

Rui Xu and Don Wunsch. *Clustering*. John Wiley & Sons, 2008.

Lihui Yang. Developing an Intelligent Questionnaire System for On-going Competence Management Platform. Master's thesis, Institue of Knowledge Based Systems and Knowledge Management, University of Siegen, February 2015.

Weizhong Zhao, Huifang Ma, and Qing He. Parallel K-Means Clustering Based on MapReduce. In *Proceedings of the 1st International Conference on Cloud Computing*, CloudCom 2009, pages 674–679, Berlin, Heidelberg, 2009. Springer-Verlag.

Appendix A

Summary of the Literature Analysis

Table A.1: Analysis and classification of the literature based on the focus of this dissertation and its directions

Description	Literature
1- Fundamental Literature: Fundamental domain specific references that have been used as sources of definitions, and setting up general and core issues of this research	[White, 1959] [Lundberg, 1972] [McClelland, 1973] [Gilbert, 1978] [Prahalad and Hamel, 1990] [Baladi, 1999] [Sandberg, 2000] [Lindgren et al., 2004] [Hustad et al., 2004] [Lindgren et al., 2004] [Delamare Le Deist and Winterton, 2005] [Teodorescu, 2006] [Eurpean Commission, 2008] [Ennis, 2008] [Ley et al., 2008] [Robles, 2012] [Bailey, 2014] [Dessler, 2015] [Banks, 1998] [Banks et al., 2010] [Berander and Andrews, 2005] [Bohm and Zech, 2010] [Boreham, 2004]
2- Proof of the concept Literature: Publications that have realized research road-maps in the main disciplines of this PhD through survey studies or similar methods in order to support argumentation for novelty of this thesis	[Lindgren et al., 2004] [Draganidis and Mentzas, 2006] [Ennis, 2008] [Tissot, 2008] [Rozewski and Malachowski, 2009] [Rozewski et al., 2011] [Bailey, 2014] [Hustad et al., 2004] [Ley et al., 2008] [Rozewski and Malachowski, 2009]
3- Technological Literature: References and guidelines that are used in for development, setting-up and examining technologies	[Abadi, 2007] [Allen and Pilot, 2001] [Bandaranayake, 2008] [Banker, 2012] [Berthold et al., 2008] [Bohlouli et al., 2013b] [Bondi, 2000] [Carlson, 2013] [Chu et al., 2006] [Conway and Guy, 2012] [Dalter, 2014]
4- Gap Analysis Literature: Related research works and projects that lack the use of specific methods or concept. In fact, this PhD contributes to identified lacks of these works.	[Bersin, 2012] [Bohlouli et al., 2013a] [Blömeke and Zlatkin-Troitschanskaia, 2013] [Bohlouli et al., 2015b] [Shavelson, 2013] [Naldi et al., 2011] [Oliveira and Nald, 2015] [Naldi and Campello, 2015]
5- Related Disagreed Work: All works that have implemented or realized similar focus of this PhD with different methods and algorithms. In particular, this dissertation disagrees with their method and provides better research results and improved added values	[Rozewski, 2012] [Shavelson, 2013] [Amiri et al., 2009] [Ley et al., 2008] [Lindgren et al., 2004] [Rozewski, 2012] [Rozewski and Malachowski, 2009] [Süße and Wilkens, 2014] [Zhao et al., 2009] [Süsse, 2013]

Appendix B

List of Supervised Theses

The following is the list of selected supervised theses in the frame of this PhD from 2012 till 2016 at the institute of knowledge based systems and knowledge management. The list consists of student projects, seminar reports as well as bachelor and master theses.

1. Yogesh Patel. Developing a Software Tool for Competency Profiling of Professionals. Tech. rep., Institute of Knowledge Based Systems (KBS), University of Siegen, Germany, November 2012.

2. Serik Bekdjanov. CMaaS: Competence Management as a Service. Master's thesis, Institute of Knowledge Based Systems (KBS), University of Siegen, Germany, October 2013.

3. Jens Dalter. SoMABiT: Social Media Analysis using Big Data Technology. Master's thesis, Institute of Knowledge Based Systems and Knowledge Management,University of Siegen, September 2014.

4. Tanja Joan Eiler. Competence Management for IT Companies. Technical report, Institute of Knowledge Based Systems and Knowledge Management, University of Siegen, January 2015.

5. Jonathan Peter Hermann, Fabian Peter Sunnus. Developing a System to Analyze Scientific Competences from Bibliographic Data using NoSQL, Bachelor's thesis, Institute of Knowledge Based Systems and Knowledge Management,University of Siegen, July 2015.

6. Martin Schrage. Realization of a Recommender System for Improving Competence Goals using Big Data Technology, Master's thesis, Institute of Knowledge Based Systems and Knowledge Management,University of Siegen, April 2016.

7. Zhonghua He. Clustering of Large Job Specific Data Using Hadoop Based on R Analysis, Master's thesis, Institute of Knowledge Based Systems and Knowledge Management,University of Siegen, June 2016.

List of Abbreviations

AHP Arithmetic Hierarchy Processing. xxi, 26, 46, 93, 96, 114–118, 133, 134, 137

AIS Active Influence Scientometric (AIS). xviii, 61, 94, 95, 101, 104–106, 112, 114, 119, 121, 123, 134, 136

BMBF Bundesministerium für Bildung und Forschung. 17, 27, 32

CA Competence Assessment. 18, 20, 21, 24–26, 30, 31, 52, 61, 63, 68, 96, 132, 133

CDP Competence Development Profile (CDP). xviii, xxii, 93–97, 114, 116–119, 121, 125–129, 133–135, 137

CDR Competence Development Recommendations (CDR). 93, 94, 96

CKM Career Knowledge Management. 2–4, 8, 12, 13, 43, 44, 47, 132

CKR model Career Knowledge Reference Model. xvii, xxi, 14, 15, 21, 22, 26, 30, 31, 44, 46–48, 51–63, 65–69, 72, 73, 80, 85, 92, 94–96, 106, 109, 110, 112–115, 117, 119, 124, 132, 133, 135, 136

CK Career Knowledge. xvii, xviii, xxi, 1–6, 8–16, 19, 21, 22, 25, 30, 43–45, 47, 51–67, 69, 71–73, 80–82, 84–87, 92–95, 104, 112, 114, 121, 122, 125, 131, 132, 136, 137

CMS Competence Management System. 8, 22, 23, 47, 132

CM Competence Management. xxi, 2, 7, 15, 17, 18, 20, 22, 23, 25–30, 32, 47, 55

ComProFITS Competence Profiling Framework for IT sector in Spain. 54, 64, 132

DB Database. 33–40, 53, 90–92, 99, 101

DFG Deutsche Forschungsgemeinschaft. 17, 27, 28, 31, 32

DM Decision Making. 3, 4, 9, 14, 45, 46, 131, 132

DSS Decision Support System. 3, 8, 41, 43, 46

DS Decision Support. 3, 8, 9, 12–14, 48

EC European Commission. 3–5, 19, 21, 47

EDR Employee Development Review. 6, 7, 44, 65

EMRKM Evolutionary MapReduce K-Means Algorithm. 96, 97, 108–112, 119, 121–123, 125, 134

EU European Union. 17, 27

HCV Hierarchical Cumulative Voting. 14, 61–64, 67, 68, 137

HDFS Hadoop Distributed File System. 34, 35, 40

HRM Human Resource Management. 1–6, 8, 12, 13, 30, 41, 47, 55, 69, 132

HR Human Resource. 1–15, 18, 19, 22, 24, 26, 27, 34, 45–48, 51, 53, 68, 71, 72, 93, 131, 132

IC Industrial Challenge. 4, 9–12, 15, 25, 44–46, 48, 52, 56, 71, 94

IT Information Technology. 3, 5, 7, 20, 23, 26, 53, 132

JP Job Profile. xviii, xxii, 9, 10, 14, 52, 55–61, 63, 65, 93–97, 112–114, 117, 119, 124, 125, 133, 134, 137

KSA Knowledge, Skills and Abilities. 20, 21, 23

ACK Acquired Career Knowledge (ACK) Matrix. 52, 53, 61–65, 112

RCK Required Career Knowledge (RCK) Matrix. xxi, 52, 53, 56, 62, 63, 65, 67–69, 93, 94, 96, 112, 132, 133

MR MapReduce. xvii, xviii, xxii, 1, 34–37, 39–43, 71, 93, 96, 107, 108, 119, 125–129, 134

PMML Predictive Model Markup Language. 42, 43, 47

IEEE RCD IEEE Reusable Competency Definitions (IEEE RCD). 24, 30

RDBMS Relational Database Management System. 33, 35–40, 44

RQ Research Question. 4, 9–12, 15, 26, 43, 45, 46, 48, 69, 71, 94, 134

SCF Scientific Competence Factor (SCF). xviii, xxi, 61, 94, 95, 100–104, 106, 112, 114, 119, 121, 123, 134–136

TA Talent Analytics. 4, 6, 41, 71

TOPSIS Technique for Order of Preference by Similarity to Ideal Solution. 14, 46, 96, 97, 112, 117, 122, 124, 134

TP Talent Profile. xviii, 52, 53, 93–97, 106, 112–114, 117–119, 124, 125, 133, 137

VET Vocational Education and Training. 13, 23, 26, 28–30, 32, 44, 93, 94, 119, 134, 135

VM Virtual Machine (VM). 34, 121

XML Extensible Markup Language. 23, 25, 30, 55, 90