

Using Data Mining Techniques to Support the Creation of Competence Ontologies

Sabrina Ziebarth, Nils Malzahn, H. Ulrich Hoppe
University of Duisburg-Essen, Department of Computational and Cognitive Science
{ziebarth, malzahn, hoppe}@collide.info

Abstract. Skills and competence requirements in the IT and media sector are changing dynamically at a high rate. This makes the “manual” adjustment of domain ontologies for this area hardly feasible, so that there is a need for automatic or semi-automatic techniques to update such knowledge bases. Several data mining techniques have been applied to different collections of job offers from this field to test and exemplify the feasibility of this approach. The results are encouraging and will be used to support decisions in professional development.

Keywords: intelligent technologies, data mining, ontologies, competences, lifelong and workplace learning, decision support, professional development

1. Introduction: Competence Ontologies

While there is a common agreement on the importance of competences, especially for education and human resources [1][2], there is a notable divergence with respect to a general definition. Erpenbeck and von Rosenstiel define competences as dispositions for self organized acting [3]. Definitions in the domain of human resources (like [1] or [2]) highlight the relevance of competences for the professional performance of individuals in business contexts and that they are measurable or at least observable.

In our approach, we focus on the use of “competence assignments” (either by oneself or by third parties), typically in response to “competence requirements” as they are formulated in job offers. Here, we simply consider competences as the central terms describing the abilities and skills potential applicants should possess. Thus no assumptions about the operationalisation or measurability of competences are needed. However, we assume that the terminology used is governed by semantic relationships to be captured in an ontology.

Ontologies are an important component in knowledge management and organization as well as in information retrieval. Ontologies aim at providing a shared and common understanding (cf. [4]) of a domain, which can be exchanged between people and heterogeneous application systems [5]. In this sense, ontologies support the externalisation of knowledge for knowledge management purposes. In the educational field, ontologies of theories and methods have been developed and applied to educational design [6].

Domain ontologies are normally built either from scratch or by composition from existing ontologies [5]. The process of ontology development from scratch can be compared with models from software engineering distinguishing phases of knowledge acquisition, conceptualization and formalization [5].

But there are some drawbacks in current ontology engineering practices, especially with respect to dealing with conceptual dynamics [7]. The domain of professional competences for jobs in the IT sector is inherently dynamic in that new competences arise and become important while others become irrelevant in a relatively short time span. Creating or updating an existing ontology needs time; release cycles for ontologies are typically six to twelve month at best [7]. Thus ontologies in domains of high conceptual dynamics tend to become obsolete because they cannot cope with novel concepts, which are often even the most interesting ones. Other problems are the (economic) costs of creating ontologies and their perspicuity for the end user.

While there are some approaches for including the end user into a collaborative ontology creation process [8][9], other approaches try to extract knowledge from artefacts of virtual communities automatically by data mining and information retrieval techniques [10]. The European TENCompetence project proposes a data fusion approach with similar goals [11]. Automatic generation and maintenance of ontologies may help to overcome the problems of conceptual dynamics and thus lower the cost. Following this approach, we describe first results from the ongoing German project KoPIWA on competence development in open innovation networks in the IT and media sector, which is funded by the BMBF (01FM07067-72).

2. Automatic Construction/Extraction of Competence Ontologies

Job offers reflect the requirements of specific job profiles and thus are adequate for analyzing the required competences and professional qualifications. Furthermore job offers and applications contain a certain vocabulary for describing competences which is shared by human resources and job applicants. Online job portals like Monster¹, StepStone² or JobScout24³ provide a rich amount of digitally available job offers, which may be used as corpus for extracting important competences and their relationships.

We have tested the applicability of certain data mining and information retrieval techniques for automatic knowledge extraction from job offers. These experiments are described in more detail in sections 2.2-2.3. The data set for 2.1 was taken from the job portal of the German Association of Digital Economy (BVDW), whereas 2.2 and 2.3 are based on around 3000 German job offers of the IT sector harvested from the job portal Monster. For the application of the data mining techniques the open source tool Rapid Miner [12] was used.

2.1. The relevance of domain-specific vs. general competences

Current research on competences for the IT and media sector is centered on general competences and social skills [13][14]. Usually, domain-specific skills are not elaborated in adequate detail [15], although technical and methodical skills are very important for recruitment and training purposes in the IT and media sector. Using a data mining approach, here esp. a decision tree, we have been able to demonstrate that the role domain specific skills and competence is by far more important to identify the

¹ <http://www.monster.de>

² <http://www.stepstone.de>

³ <http://www.jobscout24.de>

specificities of a job profile than general competences. This should have an impact on strategic decisions for IT recruiting.

This first result was achieved by an analysis of a set of 152 electronic job offers from the job portal of the BVDW. For detecting the discriminatory terms of the job offer classification, a decision tree based on the ID3 algorithm [16] was created. The decision tree contains seven professional, two non-professional competences and no certified qualifications. The professional competences have a higher position in the tree, which indicates their importance as discriminatory aspects of job profiles.

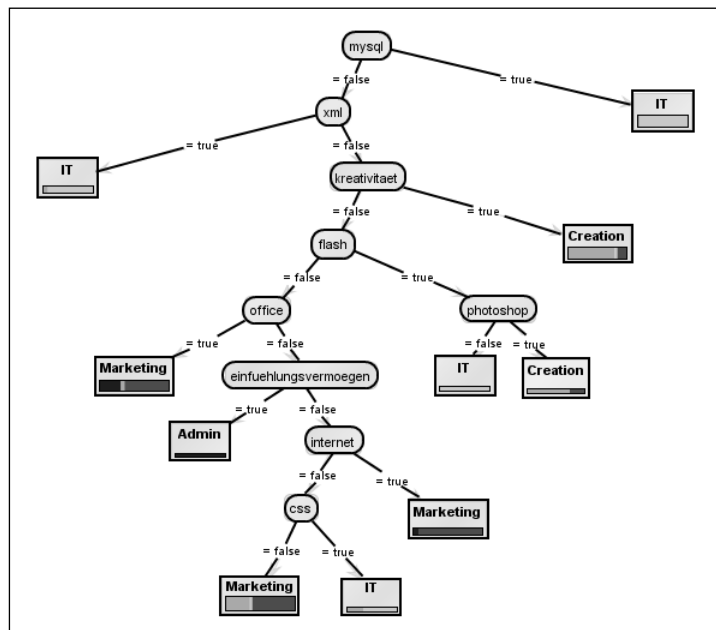


Figure 1: ID3-Tree of discriminatory terms for the job groups Admin, Creation, IT and Marketing.

2.2. Identification of profiles by clustering

Clustering techniques are used in data mining to decompose example sets into clusters of similar examples. Clustering of job offers has proved the existence of clusters of similar jobs in the corpus of job offers and provided an insight into the competences discriminating the clusters. We have clustered several samples of the example set containing 200 job offers each with the clustering algorithms k-Means [17], k-Medoids and FarthestFirst [18] using different parameters and resulting in 310 clusters. While the first order clusterings showed promising results, we tried to improve the robustness and quality by clustering the clusters [19][20]. For that purpose the centroids of the clusters were calculated as the average tf/idf values of the clusters and used as input for an additional clustering with the X-Means algorithm, which generally detects the best number of clusters in a given interval [21]. The tf/idf value is a measure for the importance of a term in document [17]. The 10 most important competences of the resulting clustering are displayed in table 1.

Cluster 0 concentrates competences for IT management and consulting, cluster 1 competences for commerce and support, cluster 2 for SAP consulting and cluster 3 for software engineering and development. The clustering results show, that it is possible to automatically create meaningful job clusters based on job offers.

Cluster 0	Cluster 1	Cluster 2	Cluster 3
IT, Management, Security, Consultant, Business, Project, Field of Activity, Service, Customer, Process	Merchant, IT, Service, Support , Administration, Ability to work under pressure, Completed Vocational Training, Office, Personnel Service, Installation	SAP SD Logistic, Consulting, Management, International , Project, System, BW, Merchant	Development, Java, Software, Software Development, Engineering, Business, Reporting, IT, Technology, Management

Table 1: Results from the clustering of clusters.

2.3. Detection of relations among competences using association rules and latent semantic indexing

The joint use of competences in different job offers indicates relationships between competences, which can be detected by association rule methods like the Apriori algorithm. The Apriori algorithm searches for prevalent term co-occurrences and deduces association rules meeting a given threshold for support and confidence [17]. The application of Apriori resulted in strong relationships between two groups of terms: The first consists of "merchant", "personnel service" and "accountancy" and the second one of "SAP BW", "BI", "Business Intelligence", "Netweaver" and "academic studies". These relationships seem quite plausible.

Another result of analyses and knowledge extraction process is the support of human resources management by reflection and recommendation mechanisms. When we applied the Tertius algorithm [22] to detect negative dependencies we found e.g.:

- engineering = true \Rightarrow merchant = false
- merchant = true \Rightarrow creative = false
- enthusiasm=true \Rightarrow creative = false

Remarkable rules are those ones, which cannot be explained easily, like the third rule. They hint at missing patterns in the heads of the one, who create the job offers.

But not only are the "apparent" relationships of competences of interest, but also the ones veiled by lexical ambiguities (synonymy, polysemy). These latent relationships can be identified by latent semantic indexing (LSI) [23]. This method uses singular-value decomposition to arrange the semantic space in a way which highlights important patterns and neglects nonrelevant ones. LSI was used to create relationships between competences, job offers among themselves and between competences and job offers. Figure 2 shows an extraction of competence relationships generated with LSI.

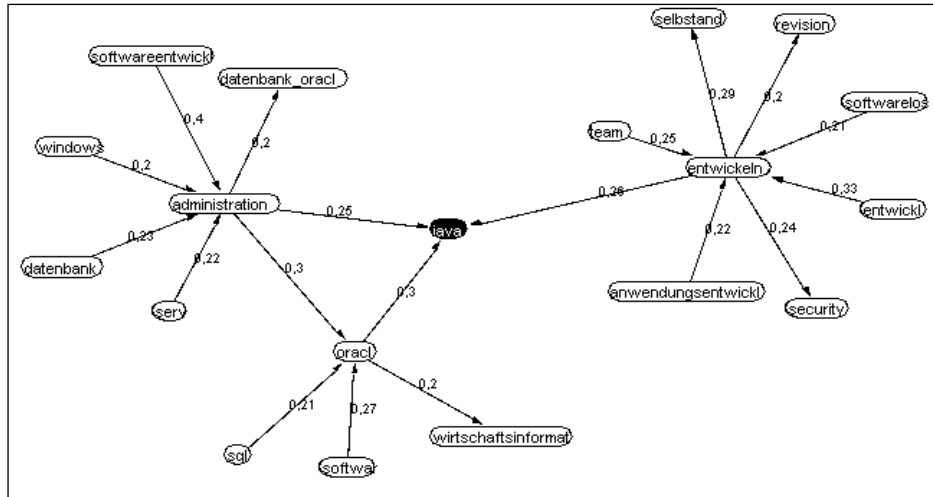


Figure 2: Extraction of competence relationships generated with LSI.

3. Prospect: Support for individual professional development

Based on empirically rooted competence ontologies, we have developed a cost/benefit model, in which these are used to support the planning of professional competence development (cf. [24]). Figure 3 shows a sketch of this development model. From top left to bottom right, the model consists of the market level (1), the competence level (2) and the social network level (3).

On the market level skills are extracted from job offers. They are weighted based on the number of occurrences per time slice, i.e. if these skills are mentioned in more recent job offers their weight is higher. On the competence level the skill nodes from the market level are combined and set into relation with each other and the deeper level competences. The numbers on the relations represent the estimated effort that needs to be spent to acquire a competence, if the competence on the starting point of the arrow has already been learned. The social network level includes the competence nodes as well. Arrows from the competences to the actor mean that this actor likes to acquire this particular competence. Links from the actor to the competence means that this specific actor has learned this competence. Links between actors shows who thinks that one is attracted by the other. Attraction in this context means that they either know each other or that one actor takes interest in the other.

The social network level takes into account that professionals are part of a social network that represents a part of their social capital. Thus changes caused by decisions for new learning and career opportunities. This may happen e.g. if the professionals try to orientate themselves towards new technologies which are not part of the portfolio of their current employer or more general of the team they are currently working in.

The competence level is based on an ontology of competences, which is partly built and updated by the methods described section 2 and partly engineered by domain experts. Apart from describing the competences in the specific domain, the ontology can be used to provide insight into relationships between particular competences. We are especially interested in providing links between competences representing the

estimated effort to acquire the specific competence if the individual already possesses one or more of the competences that are linked to it. Another important aspect to be modeled by the ontology is the explicit differentiation between surface level competences (e.g. skills) and deep level “generic” competences.

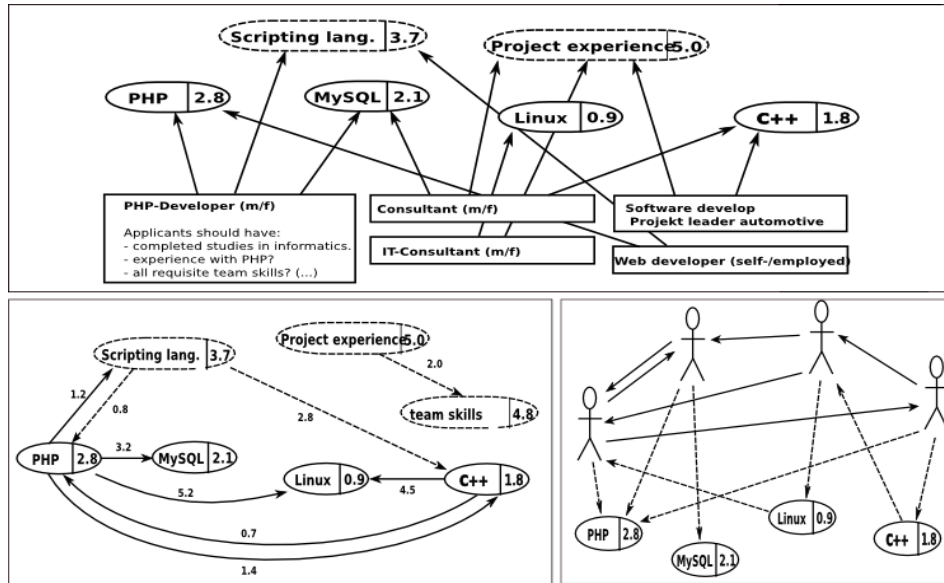


Figure 3: Three-level model for individual professional development.

From our analysis of job offers in the IT domain, we can tell that most offers target surface level competences that can immediately be applied to present tasks and the currently used tools. Unfortunately these competences do not address the deeper understanding of a domain which may allow the individual to keep track with future changes. Thus surface level competences tend to become obsolete quickly, whereas deep level competences may be of long term value, because they enable the individuals to transfer their knowledge to new problems and tasks. Furthermore we assume that deep level competences allow the individual to learn a corresponding bundle of surface level competences, because of the knowledge about the underlying concepts.

Having modeled relevant job profiles in the ontology (as shown in 2.2) enables the support system to assist the professionals in their career planning. Together with modeled relationships between the competences, a path from the professional’s current profile to the targeted position can be inferred and suggested (see Figure 4). Since there will usually be more than one path from the current set of competences to the targeted set of competences, the other two levels of the comprehensive model are considered while recommending a certain path. The market level increases or decreases the importance of particular surface skills. This is important to stay employable. The social network level increases or decreases the importance of the competences that are held or expected by the network buddies (either peers or organisations) of the individual IT worker. In the end the support system will present a set of to-be-acquired competences based on an overall ranking value derived from the learning effort needed, the gain or loss of social capital and the market demands.

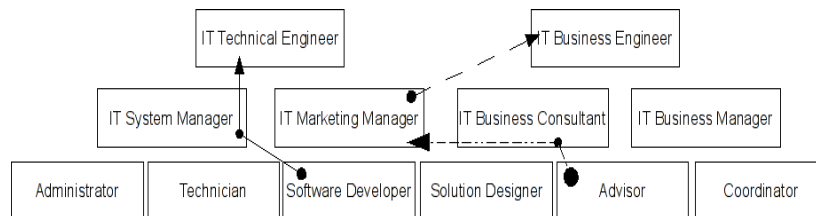


Figure 4: Alternative trajectories in individual professional development.

As the model is based on data that is frequently updated like the market demand, the evolving ontology and the changes in the social network it is capable of capturing the dynamics of the specific branch of trade. A small scale study showed that the proposed model is applicable to real cases [25]. If it is used and maintained over a longer period of time it is also possible to capture the dynamics of the personal profile. This may lead to an even better decision support, because the zone of proximal development [26] can be determined more accurately.

4. Summary

Techniques from data mining and information retrieval appear to be a promising approach for automatic knowledge extraction about competences and their relationships from job offers. We could detect clusters of similar job profiles as well as weighted relationships between competences. Furthermore, we proved the importance of specific professional competences in comparison with general and social competences. Negative rules on competences which cannot be explained easily could be a stimulus for human resources creating these job offers.

Concerning the automation of the extraction of competence ontologies, there are still some drawbacks: Although the job offers were broken down into tf/idf weighted terms automatically and filtered using stopword lists and an online dictionary⁴, the list of considered competences had still to be filtered manually to exclude terms that could not be considered as competences. Furthermore not all relations detected automatically were reasonable. So, to create a reliable ontology, domain experts are still needed for validation purposes.

On the basis of this knowledge extraction approach, a model for supporting individual professional development considering the dynamics of the market and the individual career intentions has been formulated and will be further applied and tested in an ongoing academic-industry cooperation project.

References

- [1] HR-XML Consortium: Competences (Measurable Characteristics), <http://www.hr-xml.org>, 2007

⁴ We used the web services of the project "Deutscher Wortschatz": <http://wortschatz.uni-leipzig.de/>

- [2] Schmidt, A.; Kunzmann, C.: Towards a Human Resource Development Ontology for Combining Competence Management and Technology-Enhanced Workplace Learning. In: On the Move Conferences 2006, OnToContent Workshop, Montpellier, LNCS, Springer, Heidelberg/Berlin, 2006
- [3] Erpenbeck, J., von Rosenstiel, L.: Handbuch Kompetenzmessung - Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis, Schäffer-Poeschel, 2003
- [4] Gruber, T. R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies 43, 1993, 907-928
- [5] Pinto, H. S.; Martines, J. P.: Ontologies: How can they be Built? Knowledge and Information Systems 6, 2004, 441-464
- [6] Hayashi, Y.; Bourdeau, J.; Mizoguchi, R.: Ontological Support for a Theory-Eclectic Approach to Instructional and Learning Design. In W. Nejdil and K. Tochtermann (Eds.): Innovative Approaches for Learning and Knowledge Sharing (Proc. of EC-TEL2006, Crete, Greece, Oct. 1-4, 2006), pp.155-169, Springer LNCS 4227, Berlin & Heidelberg, 2006.
- [7] Hepp, M.: Possible Ontologies: How Reality Constrains Building Relevant Ontologies, IEEE Internet Computing, 2006
- [8] Zacharias, V.; Braun, S.: SOBOLEO - Social Bookmarking and Lightweight Engineering of Ontologies. In: Proceedings of the WWW 2007 Workshop on Social and Collaborative Construction of Structured Knowledge, Banff, Canada, 2007
- [9] Malzahn, N.; Weinbrenner, S.; Hüskens, P.; Ziegler, J.; Hoppe, H. U.: Collaborative Ontology Development - Distributed Architecture and Visualization. In: Proceedings of the German E-Science Conference, 2007
- [10] Lin, F.; Hsueh, C.: Knowledge Map Creation and Maintenance for Virtual Communities of Practice. In: Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03), 2003
- [11] Miao, Y.; Sloep, P.; Hummel, H.; Koper, R. (2008). An Analysis of Unreliability of Competence Information in Learning Networks and the First Exploration of a Possible Technical Solution. Paper presented at the TENCompetence Workshop: Stimulating Personal Development and Knowledge Sharing. Sofia (Bulgaria) October, 30-31, 2008. (available at <http://dspace.ou.nl/handle/1820/1606>)
- [12] Mierswa, I.; Wurst, M.; Klinkenberg, R.; Scholz, M.; Euler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and data mining, 2006, 935-940
- [13] Colucci, S.; Noia, T. D.; Sciascio, E. D.; Donini, F.M.; Ragone, A.: Measuring Core Competences in a Clustered Network of Knowledge. World Scientific. Knowledge management: Innovation, technology and cultures 6, 2007, 279-291
- [14] Edwards, M.; Tovar, E.; Soto, O.: Embedding a Core Competence Curriculum In Computing Engineering. Frontiers in Education (FIE), New York, 2008.
- [15] Sánchez-Ruiz, L. M.; Edwards, M.; Ballester, E.; Sarrias: Competence learning challenges in Engineering Education in Spain: from theory to practice. International Conference on Engineering Education, San Juan, 2006.
- [16] Quinlan, R.: Induction of decision trees. Machine Learning 1, 1986, 81-106
- [17] Witten, I. H.; Frank, E.: Data Mining – Practical Machine Learning Tools and Techniques. Elsevier Inc, 2005
- [18] Hochbaum, S.: A best possible heuristic for the k-center problem. In Mathematics of Operations Research, 10(2), Seiten 180-184, 1985
- [19] Gionis, A.; Mannila, H.; Tsaparas, P.: Clustering Aggregation. In ACM Transactions on Knowledge Discovery from Data, Vol. 1, Nr. 1, Article 4, 2007
- [20] Chan, K. P.; Cheung, Y. S.: Clustering of Clusters. In *Pattern Recognition*, Vol. 25, No. 2, 1992, 211-217
- [21] Pelleg, D.; Moore, A. W.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: Seventeenth International Conference on Machine Learning, 2000, 727-734
- [22] Flach, P.A.; Lachiche, N: Confirmation-guided discovery of first-order rules with Tertius. In Machine Learning, Vol. 42, Seiten 61-95, 1999
- [23] Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Harshman, R.: Indexing by Latent Semantic Analysis. In Journal of the American Society of Information Science, 1990, 391-407
- [24] Malzahn, N.; Urspruch, T.; Zeini, S.; Hoppe, H. U.: Dynamisierung von Personal- und Kompetenzentwicklung. Jenseits von Virtualität - Arbeiten und Lernen in Projektnetzwerken, 2007, 229-241
- [25] Schröder, S.: Implementierung und Evaluation eines Modells zur Dynamisierung von Personal- und Kompetenzentwicklung in der IT-Branche. Master Thesis, Univ. Duisburg-Essen 2007.
- [26] Vygotsky: Mind and society: The development of higher psychological processes. Cambridge, MA: Harvard University Press, 1978.