Clustering Skills for Industrial Learning

Rajiv Srivastava, Swapnil Hingmire and Girish Keshav Palshikar

TCS Research
54B Tata Research Design & Development Center
Hadapsar, Pune 411013, India
swapnil.hingmire, rajiv.srivastava, gk.palshikar@tcs.com

Abstract

The services organizations such as Information Technology (IT) need re-skilling of their employees for reducing continually emerging skill-gaps. For addressing this critical issue in the Learning Management Systems (LMS), automated mechanisms for skill discovery and their clustering are required. We propose novel use of the non-parametric, hierarchical Dirichlet process to address challenging issue of discovering number of clusters among existing skills. The proposed solution also allows incremental addition of emerging skills to an existing or a new cluster. We show the utility of the incremental discovery of clusters among skills for the IT domain.

Introduction

The growing service industries such as Information Technology (IT) hiring, utilization, and transformation of skill-sets is critical (Takey and de Carvalho 2015)(Skorková 2016)(Nugent, Dean, and Ayers 2010). We define a skill as cognitive and technical abilities acquired through learning and practice to complete a job. The large, global IT services organizations work with thousands of technical and domain skills in industries such as Banking, Manufacturing, Healthcare. In IT industry, the technical skills map to the programming languages, databases, frameworks, middle-ware, and so on.

The discovery of new skills and their organization helps in maintaining the skill's life-cycle related infrastructure and human-resource management in an organization. The Learning and Development (L&D) function of an organization, especially in technology services organization is responsible for discovering new skills, cataloguing and organizing training material and deliver it to the appropriate set of employees.

Following are the specific challenges:

- An employee's selection for up-skilling through a training or project assignment depends on the skill clusters she belongs to. The employees are moved to emerging skills within the same clusters. Any clustering algorithm which reorganizes skill clusters and does not grow clusters incrementally is not suitable for use in providing training recommendation or project allocation.
- A new skill name is discovered by comparing an extracted skill name from a resume or a blog with the organizational repository of skills. But automatically adding the new skill to appropriate cluster of skills is challenging as it involves identifying whether the incoming skill belongs to an existing cluster or requires seeding a new cluster.
- The manual methods of skill clustering suffer from inadequacy of expertise over large set of skills, subjectivity, increase in the rate of emerging skills, and increase in industries using IT skills.

The problem we address is to identify clusters of skills such that skills in one cluster are about similar technology or domain. Please note that the number of clusters is not specified a-priori. Also, the new skills should be added to existing clusters or if evaluated to be different should seed a new cluster. We pose this problem of clustering skills as that of clustering documents where the documents contain a rich description of skill and its features.

Prior Art: A prior art for clustering of the student skill profiles, (Nugent, Dean, and Ayers 2010) proposes a hierarchical agglomerative clustering or k-means clustering, requiring, for K skills, the specification of 2^K clusters. The number of skill set profiles/clusters can quickly become computationally intractable. Moreover, not all profiles may be present in the population. The authors present a flexible version of k-means clustering allowing empty clusters. We need incremental growth of existing clusters which is not supported.

For document or text clustering, there are several classes of methods available. The most common methods are tf-idf based, latent semantic indexing, agglomerative and hierarchical clustering algorithms, k-means, clustering with frequent phrases, probabilistic document Clustering, topic models and online clustering with Text streams (Aggarwal and Zhai 2012). The online clustering method, Online Spherical k-Means Algorithm (OSKM), divides the incoming stream into small segments, each of which can be processed effectively (Zhong 2005). A set of k-means iterations are applied to each such data segment in order to cluster them. Similarly, when newly arriving data points do not naturally fit in any particular cluster, these need to be initially classified as outliers. A similar approach uses bursty features as markers of new topic occurrences in the data stream (He et al. 2007). These algorithms involve incoming streaming documents as in case of new skills emerging in the market but has limitation that the initial number of clusters or parameter k need be defined. Also, it has an issue that the old clusters if inactive may get replaced by new clusters. All algorithms share these limiting behaviours for our use-case of skill clustering in an organizational context. The word embeddings or skill embeddings are used for computing skill similarity (Wong et al. 2017). But these can not be used for clustering based on cosine distance as the appropriate number of cluster is not known. Generally, if we increase the number of clusters, the sum of distances from respective cluster centres reduces. Hence, arriving at the number of clusters is difficult using embeddings.

Proposed Method: To overcome the challenges we propose novel use of hierarchical Dirichlet process (HDP) based topic modelling to cluster skill definition documents (Teh et al. 2006). HDP is a Bayesian non-parametric technique that does not require number of topics to be specified a-priori. We use HDP as the topic modelling algorithm for inferring a set of topics (themes) that are discussed in the skill definition dataset. Each topic or theme is described using its most important words. A skill definition can be described using its most important topics or themes. Each skill is clustered into its most likely theme. A cluster is represented by its most likely words. We use these words to generate the cluster label. We used HDP to infer topics on the skills dataset discussed in next section.

Experimental set-up and results

Data Preparation

We have created the skill definition dataset for total of 2521 skills which includes the technical as well as domain skills. We crawl the web including Wikipedia, Tutorials Point[1] and other sites for corpus creation. We initialize a list of root IT skill categories from Wikipedia such as 'Computing platforms' and search through sub-categories for pages related to individual skills. We pre-process text by breaking into meaningful sentences, removing duplicates and html tags. For the skills not present in the Wikipedia, we compose a definition document using the top $k=10$ searched documents using a web search engine. We use the similarity among web-search responses to identify the similar surface forms of a skill. We generate all possible unigrams, bigrams, trigrams using unique as well as distinct words present in all surface forms of a skill. An example list of possible unigrams, bigrams, and trigrams using words 'tomcat' (unique), 'apache', and 'server' is as follows: (i) tomcat, (ii) apache tomcat, (iii) tomcat apache, (iv) tomcat server, (v) server tomcat, (vi) apache tomcat server, and so on. To normalize a skill name, we search all the combinations and replace by the standard name, which is the corresponding skill or competency name in use within the organization (Zhao et al. 2015). For the given the example the standard name of skill is 'Apache Tomcat Server 6 / 7 / 8'. In the current version we ignore the version numbers associated with skill names.

Cluster label	**Aboutness of cluster**
insurance_health_risk_financial_ policy_tax_life_public_care_united	Insurance
java_apache_php_javascript_platform_ server_framework_org_language_xml	Java Technologies
sap_supply_chain_high_hana_logistics_ low_erp_planning_inventory	SAP Technologies

Table 1: Discovered labels for sample skill clusters

Topic	**Member skills**
Java related	Java SE 5 / 6, Java Solutions Java Unit Testing Frameworks (Junit: *Unit: *Mock: Cactus: etc) Foundation: Core Java (JDK 1.3 / 1.4) Java EE Solutions
SAP related	SAP Financial Supply Chain Management, SAP Financial Accounting (FI), SAP HANA Solutions (Implementation), SAP Success Factor - LMSs, SAP Quality Management (QM)
Insurance related	Life Insurance & Pensions, Life & Annuity, Property & Casualty Insurance, Health Insurance, Data Validation

Table 2: Sample discovered skill clusters for topics

In this experiment 37 clusters were discovered and their most likely 10 words are used to generate labels (Table 1). The top 5 member skills of a few topic based clusters, discovered using HDP, are given in Table 2.

User Validation

We note that the number of skill clusters discovered by the HDP based method is 37 over 2521 skills. This is much lower as compared to the manually identified 297 hierarchical clusters by Learning and Development group of Human Resource Management function over-time. We obtained the quality score for the clusters and labels on a 1 to 5 scale, 5 being the highest. The average of quality score from 15 experts for 12 skill clusters is 3.75. For the cluster labels, the average quality score is 3.94. For users the intra-cluster skill similarity is satisfactory, with a few exceptions due to their bias for skills from technology vendors such as SAP.

Conclusion

We have shown that the HDP can be used for skill definition document clustering in an incremental manner with satisfactory results. The reported work for recommending clusters for emerging skills is incorporated in a system for managing the skill catalogue and provisioning of learning infrastructure to ensure availability of skills based on the projected requirements. We are improving the generation process of skill definition document for coherent clustering.

[1] https://www.tutorialspoint.com/

References

Aggarwal, C. C., and Zhai, C. 2012. *Mining text data*. Springer Science & Business Media.

He, Q.; Chang, K.; Lim, E.-P.; and Zhang, J. 2007. Bursty feature representation for clustering text streams. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, 491–496. SIAM.

Nugent, R.; Dean, N.; and Ayers, E. 2010. Skill set profile clustering: the empty k-means algorithm with automatic specification of starting cluster centers.

Skorková, Z. 2016. Competency models in public sector. *Procedia-Social and Behavioral Sciences* 230:226–234.

Takey, S. M., and de Carvalho, M. M. 2015. Competency mapping in project management: An action research study in an engineering company. *International Journal of Project Management* 33(4):784–796.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association* 101(476):1566–1581.

Wong, T.-L.; Xie, H.; Wang, F. L.; Poon, C. K.; and Zou, D. 2017. An automatic approach for discovering skill relationship from learning data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 608–609. ACM.

Zhao, M.; Javed, F.; Jacob, F.; and McNair, M. 2015. Skill: A system for skill identification and normalization. In *Twenty-Seventh IAAI Conference*.

Zhong, S. 2005. Efficient streaming text clustering. *Neural Networks* 18(5-6):790–798.