Kauttonen J, Khan UA, Aunimo L, Nyqvist A and Klemetti A (2024) Topic mining for theses and job ads in ICT sector: can higher education institutes respond to job market demands? *Front. Educ.* *9*:1322774. doi: 10.3389/feduc.2024.1322774

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 12 March 2024
DOI 10.3389/feduc.2024.1322774

OPEN ACCESS

EDITED BY
Marta Moskal,
University of Glasgow, United Kingdom

REVIEWED BY
Perman Gochyyev,
University of California, Berkeley, United States
Kirsty Kitto,
University of Technology Sydney, Australia

*CORRESPONDENCE
Janne Kauttonen
janne.kauttonen@haaga-helia.fi

†These authors have contributed equally to this work

RECEIVED 16 October 2023
ACCEPTED 29 February 2024
PUBLISHED 12 March 2024

CITATION
Kauttonen J, Khan UA, Aunimo L, Nyqvist A and Klemetti A (2024) Topic mining for theses and job ads in ICT sector: can higher education institutes respond to job market demands?
Front. Educ. 9:1322774.
doi: 10.3389/feduc.2024.1322774

Topic mining for theses and job ads in ICT sector: can higher education institutes respond to job market demands?

Janne Kauttonen[1*†], Umair Ali Khan[1†], Lili Aunimo[1], Antti Nyqvist[2] and Aarne Klemetti[3]

[1]RDI and Competences, Haaga-Helia University of Applied Sciences, Helsinki, Finland, [2]Library and Information Services, Haaga-Helia University of Applied Sciences, Helsinki, Finland, [3]School of ICT, Metropolia University of Applied Sciences, Helsinki, Finland

Introduction: This study aims to tackle the challenge of ensuring higher education students are equipped with high-demand skills for today's job market. The focus is on aligning the knowledge acquired during their studies, as represented by final-year thesis projects, with the skills and topics specified in actual job advertisements.

Methods: We developed a computational framework that uses automated subject indexing to extract representative skills and topics from two major datasets: thesis abstracts from Information and Communication Technology (ICT) programmes of Finnish Universities of Applied Sciences, and ICT-related job ads from a top Finnish job portal. Our dataset spans 12 years, comprising 18,254 theses and 107,335 ads. The framework includes a subject indexing model for keyword extraction, dimension reduction techniques for data simplification, clustering algorithms to group similar items, and correlation analysis to compare similarities and differences between the two datasets.

Results: The analysis uncovered both similarities and differences between thesis topics and trends in job ads. It highlighted areas where education aligns with industry demands but also pointed out existing gaps.

Discussion: Our framework not only helps to align the education provided with industry demands but also ensures that higher education institutes can stay up-to-date with the latest skills and knowledge in the field, thereby better equipping students for success in their careers. While the framework was applied to the ICT sector in this instance, its design allows expansion into other fields offering a data-informed approach for continuous development of teaching curricula and methodologies.

KEYWORDS

thesis topics, job skills, topic clustering, correlation analysis, curriculum design and improvement, natural language processing

1 Introduction

One of the major objectives of a higher education institute is to prepare the students for employment (Oraison et al., 2019). In this connection, the curriculum of a higher education institute reflects the pursuit of changing trends and its emphasis on preparing students to become better professionals by equipping them with skills that match the ones required in the job market.

For this purpose, higher education institutes regularly revise their curriculum to remove subjects that become obsolete and do not meet the job market requirements and/or existing research trends. In the same way, the subjects pertaining to emerging technologies with future prospects are included in the curriculum.

Compared to other fields, Information and Communications Technology (ICT) programs in higher education institutes are especially subject to continuous change due to the rampant growth of this sector. ICT curricula require more frequent and more careful revisions due to ever-growing and rapidly changing job market demands in the ICT sector. In addition, ICT curriculum revision requires a far-sighted and visionary approach because the subjects covered in a curriculum that is in high demand today may become obsolete at the time of a student's graduation. Hence, higher education institutes must constantly strive to maintain the practical utility of their courses and explore the degree of alignment between their curriculum and workplace skills.

Requirements for the current workforce are continuously changing, especially in areas with rapid and radical changes such as the IT sector (Brasse et al., 2023). To cope with the requirements of job market in the ICT sector, it is indispensable to design a more practical and employment-focused curriculum to prepare students for professional careers and employment (Martin et al., 2000; Moore and Morton, 2017). However, this requires an in-depth and continuous analysis of the job market and feedback from various stakeholders. The traditional method of curriculum designing and revision involves including the industrial and market representatives in the process along with academicians; however, the feedback from industrial representatives is often limited and does not give full insights into the rapidly changing job market trends.

In Finland, completing a thesis is a requisite for all undergraduate and postgraduate students enrolled in Information and Communication Technology (ICT) programs as part of their degree completion process. The thesis showcases the students' skills and knowledge gained during their degree and reflects the alignment of program courses with current trends. The thesis allows students to apply the tools and methods learned and the skills acquired in the ICT program, demonstrating their readiness for the job market and applied research. Students' theses are often done in collaboration with companies, indirectly reflecting job market needs. Therefore, analyzing thesis topics is useful for identifying the skills higher education institutes aim to produce in their students and assessing how well they keep up with the job market and research trends in the ICT sector. The insights gained from thesis topic analysis can help higher education institutes revise their curriculum to align with current research and job market trends. At the same time, job advertisements reflect the needs of job markets. This data is well available via online platforms and has been often used in studying needs of employers, also within ICT sector (see, e.g., Gurcan and Cagiltay, 2019; Pejic-Bach et al., 2020; Khaouja et al., 2021; Brasse et al., 2023).

This study aims to address the following four research questions:

> RQ1: What are the main topics of the professional skills of students in the ICT field?
>
> RQ2: What are the main topics in job ads in the ICT field?
>
> RQ3: How do trends of these topics correlate temporally with each other over multiple years?
>
> RQ4: How can we computationally measure and quantify these aspects using open data sources and computational methods?

To the best of our knowledge, this is the first body of work that addresses these research questions in ICT using a standard vocabulary and data covering a long time period (over 12 years). This study builds on and goes beyond the previous work on comparing the needs extracted from job ads with the curriculum (see, e.g., Woolridge and Parks, 2016; Ketamo et al., 2019). These previous studies compare job ads with curriculum text, while we extract the skills of the students from their thesis abstracts. While curricula are defined on the level of universities, our thesis data more closely corresponds to individual student skills and topics obtained during their education from a practical standpoint. We also propose a different methodological approach that goes beyond those applied earlier.

The research presented in this paper is motivated by the idea of finding trends of topics in ICT programs in higher education institutes in Finland and correlating them with the job skills in demand. For this purpose, we first extracted the key topics of theses completed between 2010 to 2022 in all the Finnish universities of applied sciences. In the same way, we extracted topics from job ads published during the same time period in one of the major job-seeking platforms in Finland. Subsequently, we performed a clustering of the thesis and job ad topics using data analysis and machine learning techniques and found the optimal number of groups in both. This allowed us to identify the labeling of the major domains. Finally, we computed the degree of temporal cross-correlations between the frequencies of the thesis and the job topics, which allowed us to quantify alignment between the two data sets.

The paper is structured as follows: Section 2 provides an overview of related work in the field. Section 3 outlines the data acquisition, cleaning, and extraction of topics and skills, as well as trend analysis. Section 4 presents the results of topic modeling, trend analysis, and clustering of thesis topics and job ads, including a comparison of trend time series spanning 12 years. In Section 5, we analyze the results qualitatively. Finally, Section 6 concludes the paper.

2 Related literature

To the best of our knowledge, there has been a lack of research attention given to the crucial issue of using quantitative computational methods in identifying the correlation between thesis topics and job skills to determine whether curricula align with the real job market. While a few studies have touched upon this issue, they have only done so partially, focusing primarily on soft skills without conducting in-depth quantitative analyses, or they have used only one, potentially small, data source from jobs markets or the education sector (not both).

Stanton et al. (2011) explored the skills and knowledge required for eScience professionals. Using focus groups, interviews and on-site internships by students, the authors organized the evidence into a job analysis to be used for curriculum development at schools of information and library science. Similarly, Zimmer and Keiper (2021) assessed the impact of using action research to determine the knowledge, skills, abilities, and values industry leaders require from the graduates of the sports management program. The research follows the establishment of eight program learning outcomes to guide the next steps of curriculum development. These studies have limited scope and are based on subjective evaluation. Another line of research

studies the alignment between skills in job ads and resumes of job seekers (see, e.g., Gugnani and Misra, 2020; Smith et al., 2021). However, this research concentrates on methods for mining skills from text and not on the implications of the results on curriculum development methodology.

In another study which is limited to only job ads analysis, Pitukhin et al. (2016) performed a statistical analysis of required competencies in IT job ads by partitioning the competency descriptions into three parts, i.e., type of competency, level of mastering, and the subject of competency to represent each job ad as a point in feature space. The authors opine that this information model of Information Technology (IT) competency opens up the possibility of constructing an ontology for employer-required IT competencies which can be compared with academic competencies to align a curriculum with job needs. However, the study falls short of further investigating the efficacy of the proposed information model for this purpose.

Stanton (2017) learned the priorities of recruiters by conducting surveys. Subsequently, they gathered data from three job boards and compared it with academic program requirements across the United States. The work only focuses on the subjective comparison of soft skills. Rios et al. (2020) performed a descriptive analysis of job ads to empirically rank-order skill demand. While this work only uses keywords and simple heuristics, it is also focused only on soft skills.

Hilliger et al. (2022) developed a 2-cycle building-testing curriculum analytics tool to support continuous curriculum development in higher education settings. The first cycle consisted of designing the first version of the tool and evaluating its use throughout a case study spanning a 3-year continuous improvement process. The second cycle consisted of redesigning the tool according to the lessons learned during the first cycle. The tool does not take into account the job skill trends for continuous curriculum development.

Chen (2022) targeted AI curriculum in business schools to provide recommendations for future AI curriculum development. In this study, only curriculum texts were analyzed using a simple key-term search approach. The scope was also very narrow, targeting only AI-related studies instead of the full ICT-field.

Matsuda and et al. (2018) applied a simplified, supervised LDA method to analyze curricula from the top 47 universities in engineering and technology. They mapped course syllabies into 18 Knowledge Areas defined in CS2013 guidelines by ACM and IEEE Computer Society. They further utilized hierarchical cluster analysis, principal component analysis, and non-negative matrix factorization to compare curricula between universities to help design an appropriate curriculum and combat "locality bias".

Various previous works have analyzed job ads to extract trending and relevant technical skills. Dawson et al. (2021) analyzed 8 million Australian job ads from 2012 to 2020 and developed a skill set distance mapping tool to assist in job transitions. A proprietary, non-open NLP system was applied to extract 11,000 unique skills from ads. Gurcan and Cagiltay (2019) used 2,638 Turkish job ads and a semi-automatic LDA method to create a systematic competency map comprising the essential knowledge domains, skills, and tools for big data software engineering. Based on the popularity of topics, they extracted the most in-demand skills, such as those in programming and data warehousing, that should be emphasized in engineering education. Pejic-Bach et al. (2020) analyzed 1,460 job ads related to Industry 4.0 organizations globally and applied TF-IDF and clustering methods to analyze frequent phrases reflecting in-demand knowledge. The work highlighted the demand for soft skills in addition to hard ones, and the usefulness of text mining in tracking changes in rapidly developing industries. More recently, Brasse et al. (2023) analyzed 1.16 million manufacturing industry job ads between 2018 and 2020 from Germany. They applied skill dictionary matching to convert text into skills and Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018), hierarchical clustering, and manual curation to obtain 57 skill topics. These were further reduced by domain experts to 33 "future skill" clusters, which can help governments, companies, educational institutions, and individuals to adapt to the future workforce. For other related studies targeted at ICT skill identification from job ads, we refer to the review by Khaouja et al. (2021).

The literature review indicates that previous studies have not explored the correlation between thesis topics and job skills in curricula, although the thesis reflects the course content and skills acquired during a degree program. Moreover, existing literature has only briefly touched upon this issue, with a primary focus on soft skills. While job ads data have been widely utilized, there is lack of research using education-related data, such as curricula or theses. Many existing research relies on subjective expert evaluation, which can have benefits if done correctly, but is necessary laborious and costly. There is a compelling need to further investigate this topic to ensure that students are equipped with the necessary skills to succeed in the job market.

3 Data and methods

In this section, we describe the data set and methods applied in the analysis. The overview diagram of our data processing and computational framework is shown in Figure 1. The framework starts with text samples (theses and job ads) which are processed via Annif and then added to the pipelines of "Sample clustering & visualization" and "Trend analysis & comparison". The first pipeline produces a robust categorization of documents performed individually for each dataset and provides answers to RQ1 and RQ2. The second pipeline compares two datasets to find common categories and temporal similarities for term occurrences, thus answering RQ3. Taking these together, we can then evaluate RQ4. Individual steps are discussed in detail in the following sections.

3.1 Thesis data set

Information on theses was extracted from Theseus website,[1] which contains theses published by Finnish universities of applied sciences (28 organizations with 24 currently active). Theseus website is an open-access repository funded by the Rectors' Conference of Finnish Universities of Applied Sciences,[2] maintained by the National Library of Finland[3] and the consortium of University of Applied Science libraries.[4] We scraped the site collecting all

1 https://www.theseus.fi
2 https://www.arene.fi
3 https://www.kansalliskirjasto.fi/en/services/repository-services
4 http://www.amkit.fi/en

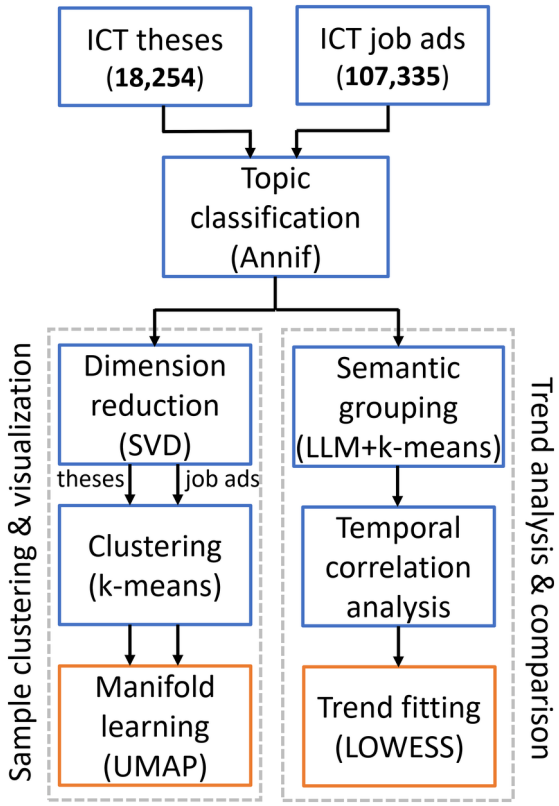

FIGURE 1
The overview diagram of our data processing and computational framework starting from data and ending with visualizations.

meta-information of published theses. This structured meta-information included author, release date, language, title, type of thesis (bachelor or master), degree program, discipline, and abstract. For many theses, full Portable Document Format (PDF) versions were also available but were not analyzed as their format was highly inconsistent, varying between authors, organizations, disciplines, and time periods. This often resulted in failed or noisy text extraction. Therefore, instead of analyzing the whole thesis, we opted to analyze only its abstract, which contains dense descriptions of the whole thesis and was deemed sufficient in extracting representative topics. Furthermore, topic extraction from the thesis abstract was computationally efficient and produced compact sets of key topics.

We first extracted abstracts of all the theses stored in Theseus from January 2010 to July 2022 containing a total of 213,737 theses. This represents 70 percent of all 305,838 theses done in Finnish Universities of Applied Sciences during that time period. The remaining 30 percent are mostly those not merged into Theseus from earlier legacy data systems and those deemed non-public due to confidentiality agreements involving private businesses and the use of sensitive data. Instead of using all available theses, we chose two fields for further analysis: Information and Communications Technology and Business Information Technology. We analyzed both bachelor's and master's theses in these two categories, both belonging to the ICT sector. This data set consisted of 18,254 theses. Abstract data was only lightly curated including the removal of any special tokens (e.g., control characters) and language detection using fastText-based language identification model (Joulin et al., 2017),[5] which was applied in the verification of the existing, sometimes erroneous language tags in metadata. Otherwise, the processing of the abstracts relied on the Annif tool (Suominen, 2019)[6] and its built-in preprocessing.

3.2 Job advertisement data set

This data set included 4,146,628 job advertisements between 2000 and 2022, published on websites run by the Ministry of Economic Affairs and Employment of Finland.[7] Each advertisement included field title, job type, occupation title, publishing date, area, language, and job description. Similar to the theses data set, we chose only ICT -related occupations for the analysis. The selection was done using the occupation titles, from which we manually chose, by leveraging hierarchy of occupational titles, a total of 74 ICT-related occupations out of all 1803 options. These occupations are listed in the Supplementary Table S1. The occupations were based on the official occupations ontology published by Statistics Finland.[8] This resulted in a total of 107,335 job ads that were included in the analysis.

Job ads contained free text description fields written by employers, often containing descriptions of the employing organizations, required soft and hard skills, qualifications, and instructions on how to apply for the position. Pre-processing of data was similar to that of theses, i.e., only including language detection and special token removal for job descriptions. For this data, we also replaced emails and URLs with special tokens (< EMAIL > and < URL >) as these were common and might interfere with topic extraction. Otherwise, processing of texts relied on the Annif tool (Suominen, 2019) and its built-in preprocessing. For each job ad, we processed the text description with the ad title.

3.3 Computation of topics

We addressed the problem of retrieving topics from abstract and job ad texts through a process known as topic classification. Unlike topic modeling, which is typically an unsupervised machine learning method such as Latent Dirichlet Allocation (LDA; Blei et al., 2003) which does not use predefined labels, topic classification is a supervised approach. Supervised techniques use labeled data and can be more effective in accurately identifying specific topics relevant to each document (Bai et al., 2009). This technique connects the content of a document text with predefined topic labels based on a large amount of training data. Therefore, the problem of extracting topics from texts was addressed as a multi-label classification problem where multiple, non-exclusive labels (topics) can be assigned to thesis abstracts and job ads. This task is also known as topic or subject indexing (Suominen, 2019).

We used Annif (Suominen, 2019), an open-source toolkit for subject indexing developed by the National Library of Finland. This

5 https://fasttext.cc/docs/en/language-identification.html
6 see https://annif.org
7 https://tyomarkkinatori.fi
8 https://www2.tilastokeskus.fi/en/luokitukset/ammatti

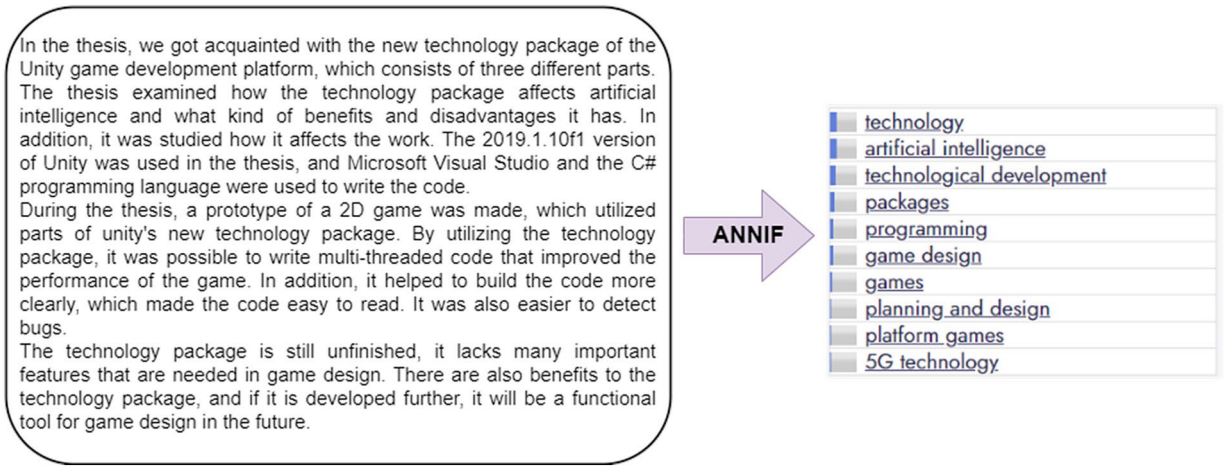

FIGURE 2
Example of extracting topics from a thesis abstract through Annif with a selected ensemble.

tool was chosen mainly for the following reasons: (1) it's proof-tested for quality and robustness by previous works (see, e.g., Suominen et al., 2022; Ahmed et al., 2023; Inkinen et al., 2023), (2) it's used by various public libraries in Finland, Germany and Sweden, (3) its open-source and trainable with own data, (4) it comes pretrained for Finnish, English and Swedish, and (5) it uses generic YSO ontology suitable for all types of texts. Annif's framework comprises a lexical subject indexing algorithm for finding correlations between subjects in the vocabulary and words in documents, a text classification algorithm, and a general-purpose machine learning algorithm. For improving performance, the algorithms are generally combined into an ensemble and the final prediction of subjects is made by using a decision function applied to the predictions of individual algorithms. In this paper, we used ensemble models, which was the most accurate pretrained model available (Suominen, 2019).

Annif's topic suggestions for a text are based on a preselected ensemble and a vocabulary based on general Finnish ontology[9] which is a trilingual (Finnish, Swedish, English) ontology consisting of more than 30,000 general and field-specific concepts. YSO encompasses the content description needs and concepts within the Finnish cultural sphere. It can be efficiently applied to interdisciplinary indexed material having diverse themes. Other potentially suitable ontologies also exist, most importantly European Skills, Competences, qualifications and Occupations[10] ontology, which includes professional occupations, skills, and qualifications relevant for the EU labor market and education and training. However, there are currently no pretrained, open source and validated models available to automatically extract ESCO terms for Finnish texts. Also, while ESCO is well suited for job ads that typically contain very specific and detailed skills, it's less suited for theses abstracts, which describe more general concepts. For these reasons, we considered YSO and Annif well-suited for the current work.

Figure 2 shows the YSO tags extracted by Annif by an ensemble consisting of a text classification algorithm (here fastText model; Joulin et al., 2017). Each topic term is predicted independently and has a score between 0 and 1. Stricter analysis can be performed by selecting topics based on their number or a probability threshold, which is a free parameter. Topics with a higher score represent higher prediction confidence. Additional demonstrations can be found in Supplementary Tables S2, S3. In this work, we set the threshold at 0.05, which resulted in an average of 12 terms for each document (i.e., a thesis abstract or a job ad). With a smaller threshold and additional terms, also probability of non-related topics become noticeable. After testing different thresholds, we concluded that the main results and conclusions of this study did not depend on the specific choice of this parameter.

3.4 Clustering analysis

We applied clustering analysis to find out the co-occurrence tendencies of samples and Annif terms. These analyses were applied to the rows and columns of the term occurrence matrix from Annif. The sample clustering contains data points as vectors where each vector represents the set of individual thesis terms. On the other hand, the data points in term clustering are individual terms. Term occurrence matrix is binary, corresponding to whether a specific topic was present in a given text or not according to Annif.

3.4.1 Sample clustering

After extracting the thesis and job ad topic using Annif, we clustered the topics using unsupervised machine learning in order to find the number of relevant clusters or groups each topic belongs to. The steps involved in the topic clustering are described as follows.

3.4.1.1 Vectorization of samples

Before clustering, it is important to convert the topic corpus into a numerical format through the process of vectorization. We used the Term Frequency - Inverse Document Frequency (TF-IDF) (Christian

9 Yleinen Suomalainen Ontologia aka YSO; https://finto.fi/yso/en
10 ESCO; https://ec.europa.eu/esco

et al., 2016) vectorization. In a typical scenario, TF-IDF is applied to whole texts, rather than keywords, for unsupervised learning problems for grouping documents. Nevertheless, here we applied it to our pre-extracted topics which were treated as individual words.

3.4.1.2 Dimensionality reduction of term occurrence matrix

The vectorization process generates sparse data having a large number of dimensions in the feature space spanned by terms (here generated Annif). This high-dimensional data can lead to the "curse of dimensionality" which can dramatically impact the clustering performance (Bellman and Kalaba, 1961). Hence, it is desirable to reduce the number of dimensions in the feature space while preserving the inter-variables relationships. For this, we applied Singular Value Decomposition (SVD) (Baker, 2005), a common linear projection technique for dimensionality reduction in sparse data, which is concerned with decomposing a sparse data matrix into its constituent parts. SVD of a high-dimensional data matrix can be calculated using iterative numerical methods with only limited computational resources (W. Zhang et al., 2012).

Selecting the right number of input features (dimensions) in the truncated data is data-dependent. After computing an SVD transform, we evaluated the same model for different numbers of dimensions and selected the number of dimensions that yielded the smallest cross-validation error. Cross-validation was performed by splitting data into training (90%) and testing (10%) parts, where transformation was fitted on training data and then applied to test data to measure the error caused by dimension reduction. This procedure is known as matrix completion (Halko et al., 2011). Due to the stochastic nature of the cross-validation procedure, we run the computations 10 times for each dimension to calculate the mean cross-validation error. Supplementary Figure S1 shows the mean cross-validation error for various dimensions of the thesis abstract data set. The minimum cross-validation error on this near-convex curve marks the optimal number of dimensions.

3.4.1.3 Clustering with K-means

After dimensionality reduction with optimal dimension count, we used the truncated data to train a K-means clustering model. K-means (MacQueen, 1967) is an unsupervised clustering algorithm to group data items into a specified number of groups (clusters). K-means performs well for clustering data with a spherical shape. Since K-means clustering only groups the data items in a specified number of clusters and cannot find the optimal number of clusters before grouping, we generated multiple clustering models and used a combination of metrics to compute clustering quality. These metrics included silhouette score (Rousseeuw, 1987) and Davies-Bouldin score (Davies and Bouldin, 1979). A high silhouette and Davies-Bouldin scores indicate that the clusters are well apart and are distinguishable. For each cluster, we computed the sum of the normalized (between 0 and 1) silhouette and Davies-Bouldin scores to find a balanced average. Supplementary Figure S2 depicts the evaluation of the quality of different numbers of clusters. The point where the curve of the sum of silhouette and Davies-Bouldin scores attains a maximum value (up to 2) represents the optimal number of clusters. Since we still had multiple dimensions after SVD, we used UMAP technique which allows visualization of high-dimensional data using only two dimensions.

3.4.2 Term clustering

Term clustering is concerned with finding groups of terms that have a high inter-cluster semantic similarity. While the sample clustering was performed before for theses and job ads data sets separately, the term clustering in this study was done by combining the terms of both data sets and grouping them using K-means clustering. In order to obtain *semantic relationships* among the terms, we used a pretrained miniLM Large Language Model (LLM) (Wang et al., 2020) available from the Huggingface repository.[11] This model can create dense, 384-dimensional vectors of terms which can be then further clustered. The K-means clustering of the term embeddings was in the same way as sample clustering. The final clustering was further revised manually to verify validity and to give descriptive names to clusters.

3.5 Trend analysis

For analyzing temporal trends of theses and job ad topics, we pooled data quarterly, resulting in a total of 50 quarters between 1/2010 and 6/2022. While exact dates were available in the data, those were considered inaccurate, particularly for theses, where the date indicates when information was added online, not necessarily the actual completion of the thesis. We also wanted to reduce the high variability present in daily data.

Figure 3 shows the quarterly count of theses and job ads with trends. It is evident that while the number of theses published in the public Theseus database in Finnish universities continues to increase, the job market exhibited notable seasonal trends. A significant decline in the number of job advertisements can be seen during the period of 2013–14 and 2019–20. The first decline can be attributed to the highest unemployment rate in Finland in 2013–14 (Statistics Finland, 2013). The other decline represents the smaller number of jobs advertised during the Covid-19 peak period. The thesis trend remained unaffected even during the Covid-19 peak period. This could result from a high degree of remote studying with lesser economic impact compared to job markets.

After obtaining quarterly time series for theses and job ads, we picked those terms that appeared in both data sets at least 20% of time points (i.e., 10 out of 50). Terms that appeared less, were considered noisy and too rare to be analyzed temporally. Then we computed cross-correlations between the two data sets for every term individually. We allowed up to a 4-year lag (i.e., 16 quarters) and located the maximum point of correlation within this time period. We temporally averaged the correlations with the two nearest neighbors. The rationale for this was to reduce spurious, unrealistically high correlation peaks. We assumed that changes in a genuine topic increase or decrease take longer than 3 months and considered individual high peaks unreliable. The statistical significance of correlations was evaluated by comparing them against those of permuted data. We used phase-mixing permutations implemented in

11 https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

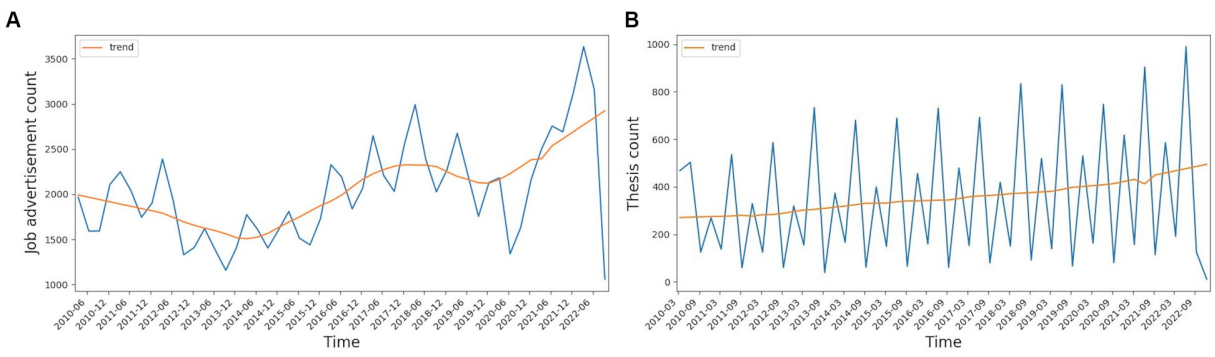

FIGURE 3
Counts of **(A)** job ads and **(B)** theses in the period 2010–2022 on a quarterly basis. Figures also include trends that demonstrate long-duration progress for counts.

BrainIAK toolbox (Kumar et al., 2021). Permutations were computed individually for each term and two-tailed *p*-values were estimated from distributions containing 20,000 iterations.

Finally, to make it visually easier to follow the development of trends over the years, we applied the Locally Weighted Scatterplot Smoothing (LOWESS) method to estimate trends (Cleveland, 1981). LOWESS is a non-parametric smoothing that eliminates noisy data values, sparse data points, or weak inter-relationships to foresee trends. To find the optimal curve for data distribution, LOWESS takes a subset of the entire data and performs a weighted linear least squares regression throughout that data. The subset of the selected data is called a fraction which determines the quality of the smoothing. For determining the optimal fraction, we first performed LOWESS for different values of fraction between 0 and 1 and selected the value that yielded the least prediction error for a single test quarter not used during training. The same fraction was then used for all terms.

4 Results

In this section, we summarize our main findings of the analyses. Definitions and mutual relationships of all terms extracted by Annif and listed in our results are available on the official YSO website.

4.1 Thesis topic clustering

After the vectorization and dimensionality reduction, the optimal number of feature dimensions found by evaluating the SVD model on different values on the thesis data set was 38. We then used the truncated data set to train a K-means clustering model on different values of clusters. Evaluating the clustering performance for different values of clusters, as described in Section 3.4.1.3, the optimal number of clusters was 28.

After performing the K-means clustering, we extracted top-10 terms nearest to the cluster centroids. After analyzing cluster terms using the full TF-IDF matrix, we manually assigned a label to each cluster. Table 1 shows the representative and the label assigned to each cluster. Figure 4 depicts the UMAP visualization of all samples and our 28 clusters. Full tables with more terms are shown in Supplementary Table S4.

4.2 Theses trend analysis

To demonstrate the time courses of clusters and individual terms, we selected the three largest clusters (depicted in Figures 5G–I) and the top 2 terms closest to the centroid in each cluster (depicted in Figures 5A–F). Each subplot shows the frequency of terms or multiple terms (cluster), the LOWESS trend, and the predicted short-term trend for the next quarter as a guide to the eye. These visualizations are useful in providing insights into a topic's and cluster's popularity and future adoption. The cluster sizes are listed in Table 2 later. Topic clusters and trends allow us to answer RQ1 related to the professional skills of students in the ICT field.

4.3 Job skills clustering

We performed vectorization, dimensionality reduction, and K-means clustering of the job ad data set in the same way as described in Section 4.1 for the theses data set. The optimal number of feature dimensions and the K-means clusters for the job ad data set were 56 and 38, respectively. These were higher than those for the theses data set (38 versus 28), which was not surprising considering the notable higher number of samples (588% more). Similar to Tables 1 and 3, Figure 4 lists representative terms and the label assigned to each cluster, and Figure 6 depicts the UMAP visualization of the 38 clusters. Full tables with more terms are shown in the Supplementary Table S5.

With two different sets of clusters, one for theses and another for job ads, we could compare them against each other by computing overlaps of terms in clusters. For this, we computed the number of the same terms between all cluster centroids. With the top-10 terms per centroid, this resulted in an overlapping ratio of 0, 10, …, 100 percent. In Table 2 we list all clusters with their corresponding number of samples. For 16 clusters, there was at least a 30% overlap between the two data sets. Fifteen of these are shown

TABLE 1 28 topic clusters and their manually assigned labels found in the theses abstract data set with 18,254 theses from ICT-field.

No.	Example terms	Cluster label
1.	Software development, computer programs, applications (computer programs)	Software engineering
2.	Data security, safety and security, data protection	Data security
3.	Automation, data systems, processes	Product development and design
4.	Web pages, websites, content management	Content creation and management
5.	Technology, information technology, technological development	Technological development
6.	Projects, project work, project management	Project management
7.	Machine learning, artificial intelligence, algorithms	Artificial intelligence
8.	Games, computer games, digital games	Game design, programming and playing
9.	Working life, work, employees	Work life
10.	Mobile devices, cell phones, applications (computer programs)	Mobile devices, apps and OS
11.	Programming, programming languages, computer programs	Programming
12.	Measurement, measuring instruments (devices), measuring methods	Measurements and testing
13.	Marketing, electronic commerce, business	E-commerce and digital marketing
14.	Modeling (creation related to information), Three-dimensional imaging, planning and design	Computer aided design
15.	Servers, Linux, operating systems	Servers and OS
16.	Online services, services, cloud services	Internet and Cloud services
17.	Windows XP, Windows 7, Windows 95	Windows OS
18.	User-centeredness, usability, users	UI designs
19.	Wireless data transmission, telecommunications technology, mobile communication networks	Wireless communication
20.	Customers, customer orientation, customer service	Customer services
21.	HTML, web pages, applications (computer programs)	Web programming
22.	Universities of applied sciences, students, learning	Education and Learning
23.	Organizations (systems), leadership (activity), development (active)	Organizations and leaderships
24.	Video, editing, image processing	Image and video processing
25.	Data communications networks, information networks, protocols	Data communication and networks
26.	Social media, marketing, digital marketing	Online advertising
27.	Databases, information management, data systems	Information management
28.	Research, questionnaire survey, research methods	Research and surveys

in Figure 6. In addition, a 30% overlap was found for the 'data security' -related cluster, which was labeled as 'Safety and Security' for the job ads data set.

4.4 Job trend analysis

Similar to Section 4.2, we visualized term occurrence frequencies and LOWESS trends for the three largest clusters and their top-2 terms in Figure 7. Topic clusters and trends allow us to answer RQ2 related to the main topics in job ads in the ICT field.

4.5 Comparison between theses and job ads

In the final part of the analyses, we compared the two data sets by computing the correlations between term frequency time series covering timespan 1/2010–6/2022. To compare the two data sets, we created a common terms list by choosing those terms that occurred in each data set at least 20% of the time (i.e., at least 10 quarters out of 50), resulting in 1072 terms. From this set, we further removed 17 terms of locations (towns and countries), which were considered non-interesting for the comparison analysis, hence resulting in the final 1,055 terms in analysis. In the job-ads data set, there was a total of 1,319,788 term occurrences with the top-10 being 'enterprises', 'development (active)', 'know-how', 'information technology', 'employees', 'experiences (knowledge)', 'working life', 'services', 'staff', 'leadership (activity)'. For the theses data set, there were 224,095 term occurrences with the top-10 terms being 'final projects (education)', 'enterprises', 'data systems', 'planning and design', 'computer programs', 'development (active)', 'programming', 'Internet', 'services' and 'applications (computer programs)'. Comparing term occurrence histograms for selected 1,055 terms, Pearson correlation was 0.56 with a p-value of 3.5e-88, indicating a statistically significant term frequency similarity between the two data sets.

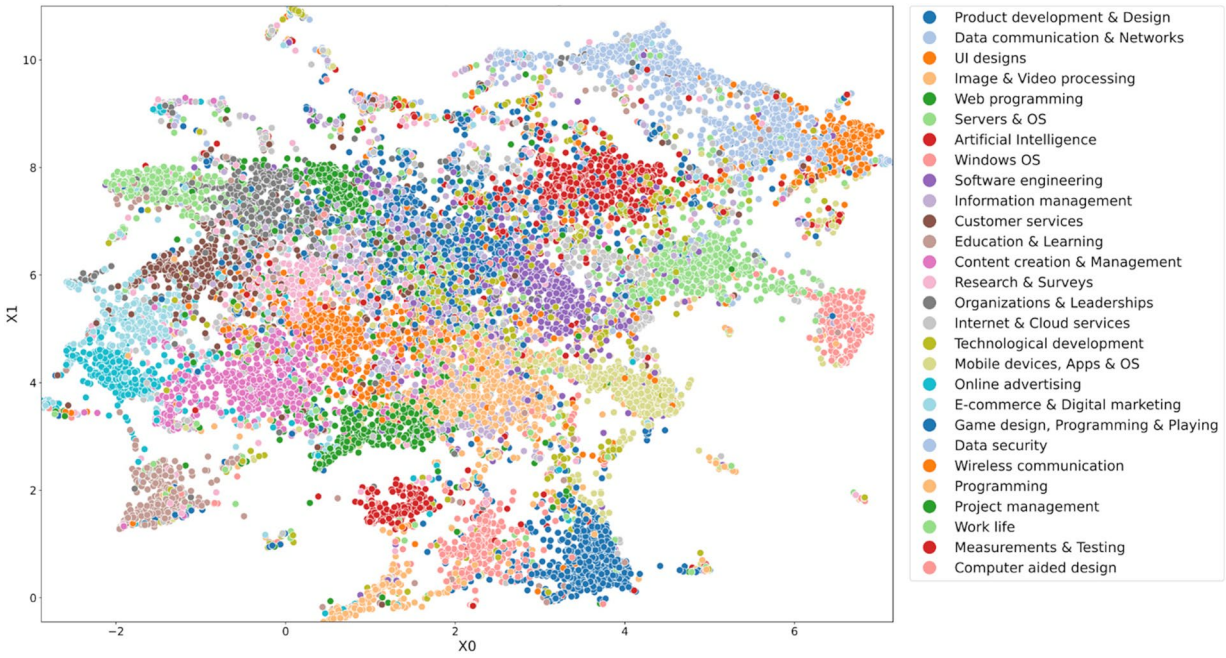

FIGURE 4
UMAP visualization of TF-IDF matrix of the thesis abstracts data set. Colors indicate 28 topic clusters and their manually assigned labels.

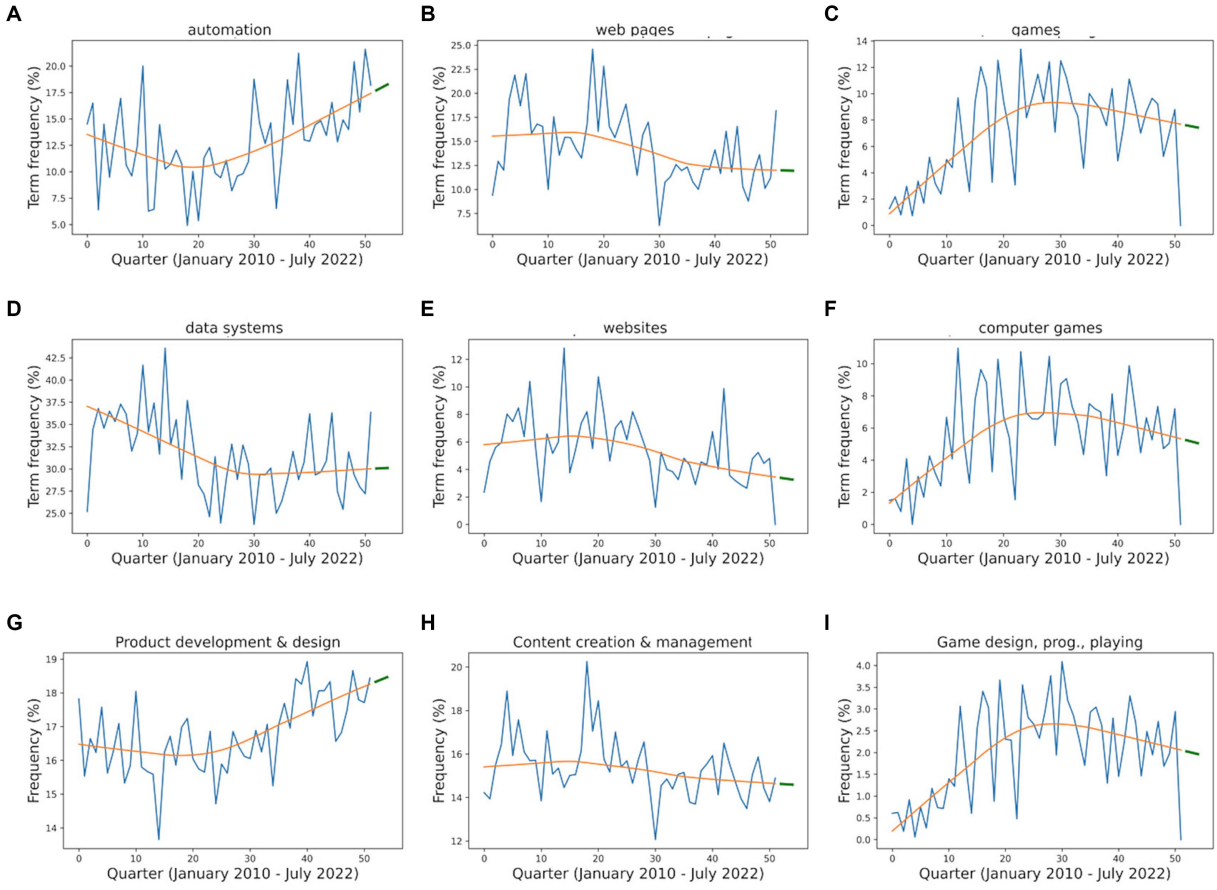

FIGURE 5
Term frequencies and trends of selected thesis topics **(A–F)** and thesis clusters **(G–I)**.

TABLE 2 Job ad and thesis clusters arranged according to size of the cluster (descending).

Rank	Job ads cluster	Ad count	Theses cluster	Thesis count
1.	**Internet services and information networks**	3,073	**Product development and design**	1,055
2.	Software development	2,950	Content creation and management	885
3.	**Customer services**	2,804	Game design, programming and playing	841
4.	Employment	2,802	**E-commerce and digital marketing**	806
5.	**Programming**	2,757	**Mobile devices, apps and OS**	780
6.	Business and enterprises	2,749	**Servers and OS**	774
7.	Business and project cooperation	2,668	**Internet and cloud services**	765
8.	Employees and staff	2,594	Online advertising	745
9.	Legislation	2,540	**Data security**	738
10.	Financial administration	2,634	Measurement and testing	727
11.	Professional skills	2,452	**Information management**	721
12.	Career development	2,404	**Software engineering**	720
13.	Work life experience and Interaction	2,331	Technological development	718
14.	Enterprises and leadership	2,321	Organization and leadership	690
15.	Teamwork	2,312	**Data communication and networks**	684
16.	**Web programming**	2,290	UI design	682
17.	Information administration	2,286	Research and surveys	646
18.	**Working life**	2,284	**Programming**	638
19.	**Servers and OS**	2,282	**Working life**	623
20.	Business, marketing and corporate services	2,261	**Project management**	610
21.	Automation	2,165	**Education and learning**	578
22.	Customership	2,154	**Customer services**	565
23.	Recruitments	2,145	**Web programming**	530
24.	**Project management**	2,114	Computer-aided design	504
25.	**Education and learning**	2050	Artificial intelligence	421
26.	Access to employment	2043	Wireless communication	397
27.	**Information management**	2011	**Windows OS**	389
28.	**Product development and design**	1967	Video and image processing	317
29.	**Data communication and networks**	1867	–	–
30.	**E-commerce and digital marketing**	1849	–	–
31.	Java programming	1,585	–	–
32.	**Safety and security**	1,547	–	–
33.	Software design	1,534	–	–
34.	Public health	1,517	–	–
35.	**Windows OS**	1,339	–	–
36.	**Mobile devices, apps and OS**	1,316	–	–
37.	**Software engineering**	1,311	–	–
38.	Consultancy	1,203	–	–

The clusters in colored fonts are common between thesis and job ad data sets.

4.5.1 Term clustering analysis

The optimal number of semantic term clusters in k-means clustering was 20 and each cluster contained from 19 up to 98 terms. Clusters were manually annotated and curated by moving 11 terms to more related clusters. All 20 clusters with their total number of common term occurrences for each data set are depicted in Figure 8. Common terms were distributed very differently in the two data sets. For example, terms in 'Work life' were most common in job ads (15%), but very rare in thesis data (2%).

4.5.2 Temporal correlation analysis

We computed temporal correlations for all common terms individually and for 20 clusters. We obtained the peak absolute correlation up to a 4-year lag (positive or negative). Total 72 terms (out of 1,055) and 5 (out of 20) clusters reached absolute correlation high enough to reach statistical significance at uncorrected $p < 0.01$. Out of those, 17 terms and 4 clusters also surpassed $p < 0.05$ with FDR adjustment. In addition, the mean overall correlation was 0.123 with 95 percentile limits of −0.268 and 0.514. The results are summarized in Table 4 which shows semantic clusters, the number of terms in each cluster, example top terms in clusters, and correlation coefficients of the cluster time series.

Next, to study topics where these were either temporally ahead or lagging job ads, we investigated terms whose peak correlation occurred at least 4 quarters (1 year) away from zero. A total of 34 terms are listed in Table 5, each with a statistical significance of at least $p < 0.01$ (uncorrected). Bolded ones also surpassed $p < 0.05$ FDR adjusted. For 22 terms (65%) job ads data set was ahead compared to theses.

Finally, we considered two example themes of interest: The gaming industry and data-analysis/AI. Out of all 1,055 terms, we picked 11 and 5 that corresponded to these particular themes. For the game industry, we chose the following terms: 'Game industry', 'game programming', 'game design', 'game sector and computer', 'video-', 'console-', 'mobile-', 'platform-', 'online-' and 'digital games'. For data analysis, we picked the following terms: 'Statistics (data)', 'deep learning', 'artificial intelligence', 'machine learning' and 'data mining'. We summed the occurrence frequencies of these together and computed trends using the LOWESS method. Results are depicted in Figure 9. The correlation for data analysis and AI was high (0.75), while the similarity was low for the gaming industry (correlation −0.11), both obtained with zero lag. These results allow us to answer RQ3 related to the temporal correlation of topics.

5 Discussion

In this paper, we have analyzed a large number of theses of students in higher education institutes and job advertisements in Finland in the ICT sector. We studied topics and trends in education and job markets and compared these two. We wanted to utilize big data sets that are open to the public, regularly updated, and with historical data available. Our overall aim was to develop a data-analysis framework that could be utilized in the development of operations of applied universities. By using our framework, we wanted to gain insight into how well the topics and themes in student theses correspond to demands in job markets. In the development of the framework, we applied computational methods from fields of natural language processing, supervised and unsupervised machine learning, and statistical analysis. Although we concentrated on the ICT sector to narrow the scope, the methods we applied are agnostic to this selection and can be applied to any field.

In our clustering analysis of thesis abstracts, we found 28 different topics (see Table 2). The three largest clusters were 'Product Development and Design', 'Content Creation & Management', and 'Game Design, Programming and Playing'. For the job ads data set we found 38 clusters with the three largest being 'Internet services and Information networks', 'Software development' and 'Customer service'. These large clusters shed light on content found in theses and job ads, as well as the main focus of demands of the skills requested in the job market for ICT professionals. We identified a total of 16 topic clusters shared between theses and job ads data sets. We also found that despite the differences, the relative occurrences of topics were still highly similar (see Figure 8). These results cover RQ1 and RQ2.

Our data spanned over 12 years, hence this allowed us to perform longitudinal trend analysis. We pinpointed topics and themes whose popularity was similar or dissimilar, and which one was following the other. We identified 5 clusters (out of 20) and 72 terms (out of 1,055) that were significantly correlated. Interestingly, all except one of these correlations were positive with the mean correlation being 0.124. This strongly positive temporal correlation indicates that there was clear correspondence in the popularity of topics in theses and job ads in general. By inspecting and visualizing time series, one can look for specific topics of interest. For example, even though the demand for software development skills has been large in the job market during the time period under study, it has a declining trend, as can be observed in Figure 7H. Also, a strong recent interest in data analysis and AI was evident for both data sets (see Figure 9B). This covers our RQ3.

University students have a relatively free choice of thesis topic as long as it belongs to the study area of the study program. Thus, thesis topics also reflect the personal areas of interest of the students and not only the learning outcomes of the curriculum or the skill set of the students. Final year thesis projects are a critical component of a student's education as they not only assess a student's mastery of the knowledge, concepts, and skills gained throughout their program but also demonstrate the program's alignment with contemporary trends and priorities. A large proportion of theses in the ICT sector are commissioned by companies. These projects are especially valuable from an industry standpoint as they prepare students to tackle real-world problems and challenges, giving employers insight into a student's skills and ability to apply their academic knowledge in practical settings.

5.1 Theoretical contributions

The theoretical contributions of this work are both conceptual and methodological. The broad context of this research is to address the difficult societal problem of how to provide a skillful workforce for the constantly changing job market. Both governmental bodies and management of higher education institutes need quantitative information to plan education to match societal needs. On the other hand, students also need timely information to decide what topics to study and write a thesis. Our approach to these aspects was to leverage data mining and analysis of open data continuously produced by the job market and education system. The previous work on the subject either uses one data source (job ads or curricula) and/or concentrates on developing new topic extraction methods. Furthermore, none of the previous works has performed a longitudinal analysis of the data and looked at temporal correlations between data sources.

Our methodological contribution includes both suggesting the usage of public job ads and theses data sets and applying a combination of computational methods to this data. Leveraging public, frequently updated data sources in particular is appealing due to being cost-effective and allowing longitudinal analysis spanning several years.

TABLE 3 38 topic clusters and their manually assigned labels found in the job ads data set containing 107,335 ads from ICT-field.

No.	Example terms	Cluster label
1.	Employment (legal relationship), working hours, employees	Employees and staff
2.	Programming, computer programs, software developers	Programming
3.	Information administration, municipalities, data systems	Information administration
4.	Work communities, work, working life, work comfort	Working life
5.	Safety and security, security systems, data security	Safety and security
6.	Online services, internet, information networks	Internet services and information networks
7.	Information management, databases, data systems	Information management
8.	Professional skills, know-how, education and training	Professional skills
9.	Financial administration, electronic financial management, enterprises	Financial administration
10.	Information technology sector, enterprises, services	Information technology
11.	Access to employment, recruitment of employees, employees	Employment
12.	Telecommunications technology, data communications, data communications networks	Data communication and networks
13.	Business operations, business, enterprises	Business, marketing and corporate services
14.	Marketing, digital marketing, Internet	E-commerce and digital marketing
15.	Infrastructures, consultancy agencies, management (control)	Consultancy
16.	Communication, customer experience, business life	Customer experience
17.	Personnel selection, recruitment of employees, access to employment	Access to employment
18.	Business, development (active), projects	Business development
19.	Project management, project work, project leadership	Project management
20.	Automation, installation, upkeep (servicing)	Automation
21.	Customer service, services, customers	Customer services
22.	Product development, innovations, technology	Product development and design
23.	Cooperation (general), development (active), projects	Business and project cooperation
24.	Web pages, HTML, JavaScript	Web programming
25.	Education and training, universities of applied sciences, universities	Education and learning
26.	Windows XP, Windows 95, Windows 7	Windows OS
27.	Testing, testing methods, quality control	Software engineering
28.	Public health service, health services, health sector	Public health
29.	Managers and executives, leadership (activity), enterprises	Enterprises and leadership
30.	Tasks, presentations (introductions), central government	Legislation
31.	Success, organizations (systems), leadership (activity)	Career development
32.	Applications (computer programs), design (artistic creation), planning and design	Software design
33.	Applications (documents), access to employment, workplaces	Recruitment
34.	Java ME, Java, Java EE	Java programming
35.	Maintenance, servers, information networks	Servers and OS
36.	Mobile devices, cell phones, operating systems	Mobile devices, apps and OS
37.	Business, enterprises, experiences (knowledge)	Business and enterprises
38.	Customer orientation, customer service, customers	Customership

Methodologically, we suggested the usage of multi-label topic classification (here Annif tool), clustering within documents and topic semantics, and temporal cross-correlation between data sets with permutation statistics. While the most common approaches in topic modeling are based on unsupervised machine learning methods (Blei et al., 2003; Dumais, 2004; Hofmann, 2013), they do not give valuable insights into a document's thematic concepts and fall short of retrieving all relevant subjects in a document (Suominen, 2019). Although they do not require any prior training, they fall short of producing compact, context-oriented, neatly packaged topics.

Our pipeline offers objective, computational measures for mapping and matching educational and job market data, thus

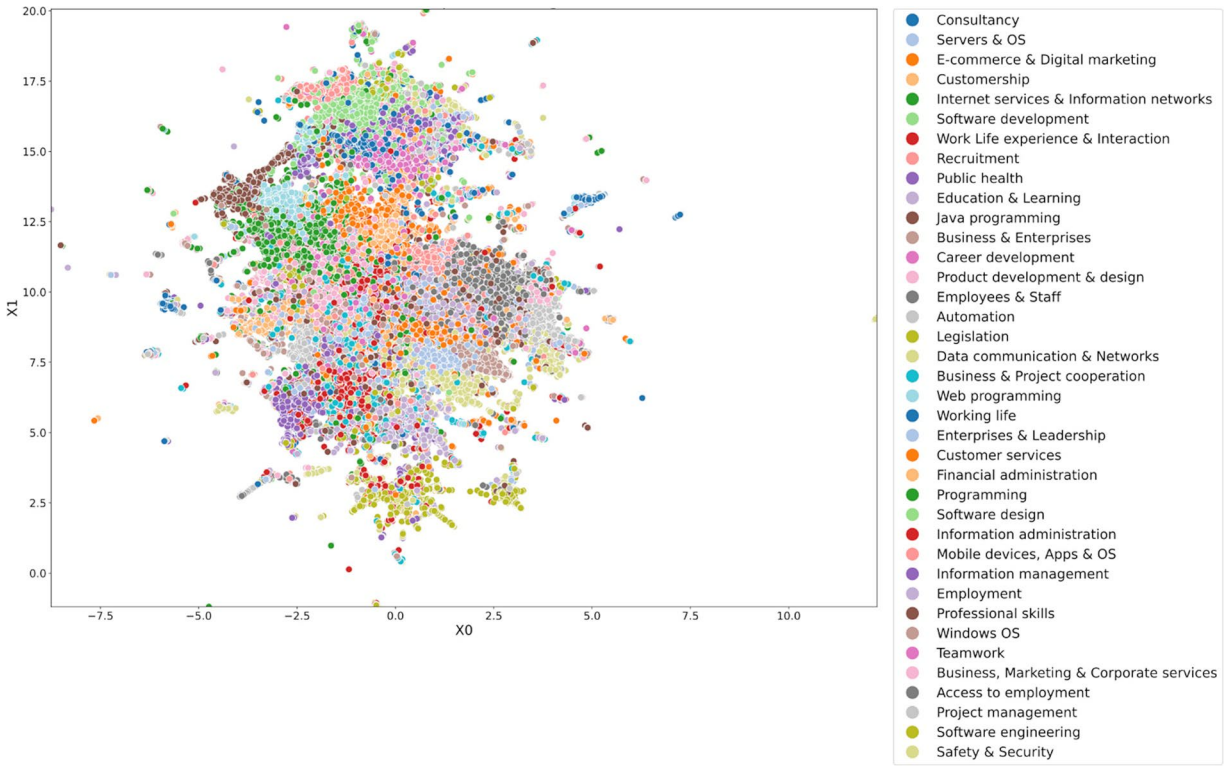

FIGURE 6
UMAP visualization of TF-IDF matrix of the job ads data set. Colors indicate 38 topic clusters and their manually assigned labels.

responding to RQ4. Our suggested analysis pipeline does not depend on any novel or tailored models or algorithms and - if necessary - can be easily configured according to particular data and objectives. For example, one could instead use a different topic extraction or clustering method or similarity metric that is more suitable to the data at hand.

5.2 Contributions to institutions of higher education

The results of the study show similarities and differences in the topics of theses and job ads in the field of ICT during a period of 12 years. Table 2 shows the cluster labels and their sizes in the thesis and job ads data sets. They can be used to observe the topics of theses and job ads. An observation of the names of the clusters shows that the thesis cluster names are more technical and narrower than the job ads cluster names. For example, all but four ('content creation and management', 'measurement & testing', 'organization and leadership' and 'research and surveys') of the non-overlapping cluster names describing the thesis data set are ICT sector-specific. For the cluster names of the job ads data set, we can observe 16 cluster names that are not at all related to the ICT sector. One implication of this finding is that the student should be prepared by the institution of education for the fact that in working life the skills needed may not be as technical and specific as what is required in the thesis work. For a practical curriculum developer, this may imply the insertion of a study module on soft skills that are needed in working life when working in multidisciplinary teams.

Another interesting result of the study is presented in Figure 10 which shows the extent of similarity between the common clusters in the job ads and thesis data sets. We can observe that project management has a high degree of similarity (80% of the terms in the respective clusters overlap), whereas customer service ranks the lowest in terms of similarity (30% overlap in terms). This may imply that the concept of project management is understood in a relatively similar manner in both theses and job ads whereas the concept of customer service has more variation. This may imply that the curriculum related to project management is relatively well aligned with the demand in the job market and that the study contents related to customer service may need some updates. There has been some discussion on the use of different terminology in the institutions of higher education and the job market (see, e.g., Ketamo et al., 2019). However, we observe that the same term may be used to mean a different concept in the context of higher education than in the context of the job market has novelty.

The trend analysis performed for both thesis (Figure 5) and job ads (Figure 7) data sets is potentially very useful for study program directors and curriculum developers as it contains predictions for both. Interesting observations about the predictions for future job ads are those concerning the ascending trend curves for the term 'online services' (Figure 7A), and the cluster 'Internet services and information networks' as well as the descending curves for the term 'software development' (Figure 8B), and the cluster 'Software development' (Figure 7H). We can observe that the demand for traditional software development skills is declining. Rather,

FIGURE 7
Trends of some selected job skills **(A–F)** and job skill clusters **(G–I)**.

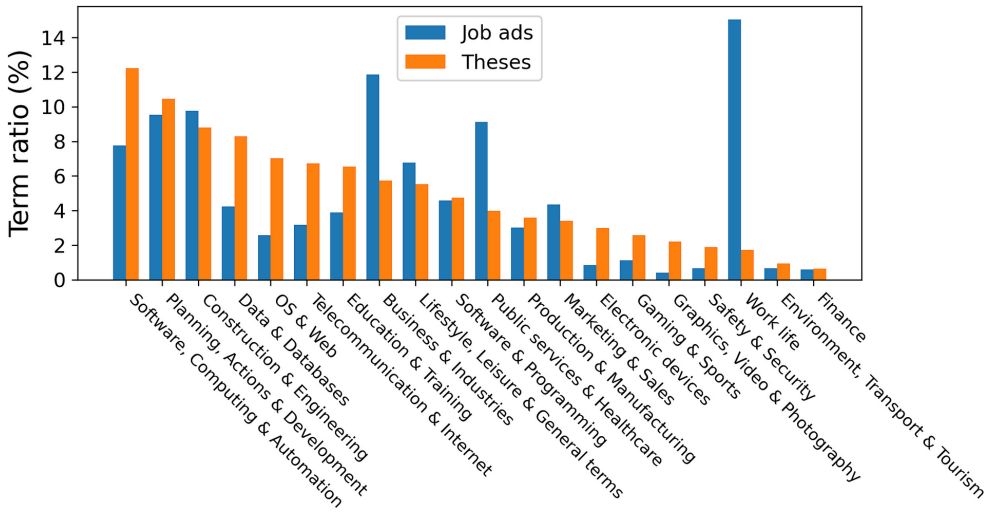

FIGURE 8
Distribution of topics occurring in theses and job ads data sets over 12 years quarters. Data includes 1,055 individual YSO terms that occurred in both data sets, split into 20 semantic clusters. Bars are shown in descending order according to the theses data set.

skills related to the development of online services are increasing. While many skills related to these tasks are overlapping, this is not the case for all. Furthermore, in Figure 5, we can observe a descending trend for both the term 'computer games' and the cluster 'Game design, programming, and planning' in theses. In addition, we found no significant correlation for the gaming sector-related topics (see Table 4; Figure 9A). However,

TABLE 4 Time series correlation results between theses and job ads data.

Semantic cluster	Example terms	Corre-lation
Education and Training (60)	Educational institutions, education and training, study, teaching and instruction, educational technology, schools, higher education	0.17
Business and industries (69)	*Small-scale entrepreneurs*, *housing companies*, *software business*, *software sector*, enterprises, business	−0.23
Lifestyle, leisure and general terms (97)	*Every day*, *names*, *know-how*, media, social media, history	0.17
Work life (39)	*Working life*, *work study*, *job description*, *work*, office work, workplaces	**0.65**
Gaming and sports (28)	*Teams*, *gambling games*, games, computer games, video games, game industry	0.37
Telecommunication and internet (57)	*Wireless communication*, *data communications networks*, *telephone technology*, *base stations*, *TCP/IP*, *wireless technology*, *local area networks*, *servers*, *mobile phone systems*, *telecommunications technology*	**0.63**
Marketing and sales (50)	*Marketing research*, marketing, advertising, digital marketing, commerce, consumers	−0.17
Computing and automation (60)	*Artificial intelligence*, *microcontrollers*, *processors*, *application generators (technical systems)*, *emulation*, computing systems	0.27
Data and databases (63)	*Data*, *data transfer*, *publishing systems*, *machine learning*, data systems, databases	0.46
Safety and security (19)	Data security, safety and security, security systems, security environment, safety and security management, cyber security	0.16
Construction and engineering (67)	*Project learning*, *development (active)*, *key figures*, projects, construction, project work	0.25
OS and web (59)	*iPhone*, *registration*, *Flash (computer programs)*, *Symbian OS*, *accessibility*, *World Wide Web*, *Windows Server*, *Windows 10*	0.68
Finance (23)	*Invoicing*, *payments*, *payment services*, *banking services*, payment systems	0.23
Planning, actions and development (98)	*Opinions*, *aids (implements)*, *tests*, *participation*, *well-being*, *interaction*, *taking advantage*	**0.73**
Software programming (36)	*XML*, *C language*, software developers, programming, programming languages, computer programmers	−0.16
Production and manufacturing (35)	*Production control*, *industrial design*, *management (control)*, production, manufacturing, production planning	−0.25
Graphics, video and photography (54)	*Graphics (visual arts)*, *PowerPoint*, *multimedia (information technology)*, *CGI*, *virtual reality*, video technology	*−0.29*
Electronic devices (47)	*Electrical engineering*, *electromagnetic compatibility*, electronic devices, electronics, electrical devices, devices	**0.59**
Healthcare, well-being and public services (56)	*Restaurants*, *specialists (experts)*, services, public services, health services, customer service	0.30
Environment, transport and tourism (40)	*Resiliency (flexibility)*, *single-family houses*, environment, living environment, environmental protection, transport	0.18

The table contains term clusters with a total number of terms (in parenthesis), example terms, and the peak correlation of the cluster (with lags). Notable correlations are shown in italics ($p < 0.01$, uncorrected) and bold ($p < 0.05$, FDR adjusted), including a total of 72 terms and 5 clusters.

the digital games industry does not show signs of decline in Finland (Kivijärvi and Sintonen, 2021).

Based on these results, software engineering curriculum developers should study carefully what is the difference and adjust the study offerings accordingly. Thanks to such forecasts, they have time to make adjustments before the changes in demand. Also, the managers of study programs related to digital games should study the reasons behind this forecast and take action before it becomes a reality.

5.3 Limitations of the work

This work is not without limitations. We used public data sources that were not directly designed or collected for this work; hence they were noisy. For example, job advertisements typically include not only relevant job skills but also general company information and recruitment information. Text style and length also vary greatly. As a result, despite limiting ourselves to ICT, Annif also extracted various other, unrelated themes (see Table 3). The same also holds for these data, although to a lesser degree. Identifying only particular segments (e.g., those related to work skills only) from a job ad text is a complex problem itself and goes beyond this work.

Instead of analyzing full theses, we only used thesis abstracts, which are unlikely to represent the full set of methodologies and analytical techniques that the students used in their project or their full theoretical knowledge. Actual knowledge and skills at the graduation of a student would require formal assessment of their skills, which is not (at least currently) done in Finnish universities. Furthermore, content in job ads might not actually correlate with the person who is finally hired as employers often struggle to find people

TABLE 5 All 34 terms whose correlation peak had a lag of at least 1 year (4 quarters, positive or negative) with statistical significance of $p < 0.01$ (uncorrected) with bolded values surpassing $p < 0.05$ (FDR adjusted).

	Term	Quarters	Correlation	Semantic cluster
Job ads ahead	Banking services	16	0.28	Finance
	Registration	16	0.31	OS and Web
	Project learning	12	**0.45**	Construction and Engineering
	Key figures	12	0.31	Construction and Engineering
	Telephone technology	10	0.34	Telecommunication and Internet
	Names	10	0.24	Lifestyle, Leisure and General terms
	Symbian OS	8	0.46	OS and Web
	Graphics (visual arts)	8	0.24	Graphics, Video and Photography
	Software business	8	0.28	Business and Industries
	Wireless communication	7	**0.31**	Telecommunication and Internet
	Payment services	6	**0.33**	Finance
	Payments	6	**0.31**	Finance
	Windows 10	6	0.61	OS and Web
	Participation	6	0.40	Planning, Actions and Development
	Application generators (technical systems)	6	0.36	Software, Computing & Automation
	Opinions	6	**0.23**	Planning, Actions & Development
	iPhone	5	**0.43**	OS & Web
	Virtual reality	5	0.44	Graphics, Video & Photography
	Work study	5	0.29	Work life
	Software sector	5	0.33	Business and Industries
	Artificial intelligence	4	**0.83**	Computing and Automation
	Restaurants	4	0.30	Healthcare, Well-being and Public services
Theses ahead	Emulation	16	0.41	Computing and Automation
	Resiliency (flexibility)	16	0.29	Environment, Transport and Tourism
	Processors	12	0.32	Computing and Automation
	Housing companies	12	**0.41**	Business and Industries
	Microcontrollers	9	0.26	Computing and Automation
	Electromagnetic compatibility	8	0.31	Electronic devices
	Invoicing	8	**0.30**	Finance
	Management (control)	7	0.41	Production and Manufacturing
	Mobile phone systems	6	0.46	Telecommunication and Internet
	PowerPoint	6	0.26	Graphics, Video and Photography
	Data transfer	4	0.44	Data and Databases
	Electrical engineering	4	**0.49**	Electronic devices

For 22 terms job ads were temporally ahead compared to theses.

with the skills in the ad, or deliberately advertise for something that they think will be popular and more likely to result in strong applicants. Both our data sources were *indirect* sources of information from the education system and job markets, hence also our results should be considered as such. Our purpose was to examine how we can leverage public, long-term data sets to map the correspondence of the two and advance curriculum development. Our conclusions from this work should therefore be complemented by other data sources, such as curricula, student skill evaluations and job market surveys.

Annif is a pretrained, generic tool for extracting pre-defined, ontology based YSO topics from textual inputs. While there are other methods developed particularly for extracting hard and soft skills from job ads (see Section Related work), no pretrained and mature multi-lingual models are currently available for Finnish, Swedish and English. Here we focused on the Finnish theses and job market due to data availability and our specific expertise in this area. While our findings are relevant for the Finnish context, we acknowledge the importance of a more global perspective. Further research is needed to extend this approach by

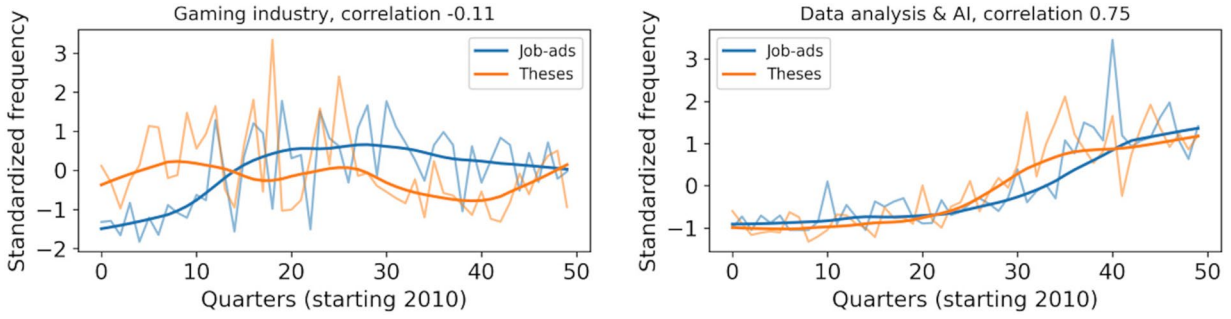

FIGURE 9
Term frequency plots for job ads and theses for **(A)** gaming industry (11 terms) and **(B)** data- analysis and AI (5 terms) themes. Frequency scales have been standardized for easier visual comparison.

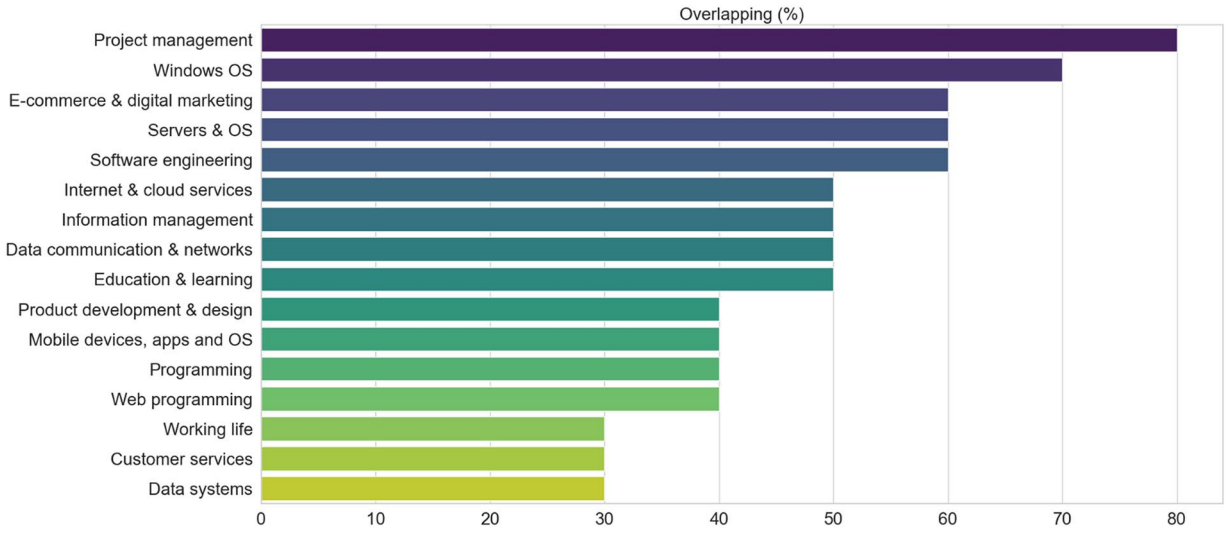

FIGURE 10
The extent of similarity between the common clusters in theses and job ads data sets. The similarity is computed by counting the percentage of the same terms between cluster centroids.

incorporating similar data from other countries, using appropriate national ontologies and topic models suited for analyzing diverse linguistic and regional contexts. This would enhance the generalizability of our findings and provide a more comprehensive understanding of the alignment between education and ICT job market needs globally.

Finally, our methodology treated all data from the 12 years with equal importance, without a specific emphasis on recent trends in ICT. While our approach provides a comprehensive, long-term overview, it may not fully capture the very latest developments in the rapidly evolving ICT sector. This is a relevant aspect when developing curricula with up-to-date developments in the field. Future studies could explore methods to prioritize recent data, such as temporal weighting in topic modeling, or to reflect current industry shifts more accurately. This would offer a more nuanced understanding of how recent changes in the sector impact the alignment between education and job market demands.

6 Conclusion

In this study, we outlined a novel contribution to the field by leveraging topic and semantic clustering to identify the most dominant areas in both thesis topics and the job market. Furthermore, we compared theses against job markets spanning 12 years to identify if popularity and trends are similar. Identification of topics and their temporal development is crucial for both curriculum designers who can adjust course content, as well as for industrial partners who can focus on these areas during student internships.

Our findings further highlight the overlap between academic and job skills and demonstrate the potential for industry-academia partnerships to be strengthened by increasing industry-oriented course content. Additionally, our trend analysis of individual thesis topics, jobs, and their respective clusters provides valuable insights for curriculum design and improving course content at both breadth and depth levels. Furthermore, the trend analysis enables curriculum designers and education experts to understand the underlying reasons

behind any disparities in trends for a specific term or cluster. This information not only guides study program managers in effectively revising their curriculum, but it also assists industries in reevaluating their preferences and aligning them with future requirements.

There are several ways one could expand and further refine this work. It would be interesting to study the differences in topics between commissioned and non-commissioned thesis topics by expanding work by Buehling and Geissler (2022) regarding the variation of topics between commissioned and non-commissioned PhD theses. While their finding was that there is only a little difference in the topics, it is not clear if this also holds for bachelor's and master's thesis and in the field of ICT. On the scale of individual institutions, the presented framework could be also applied between institutions and disciplines for cross-referencing curricula as suggested by (Matsuda et al., 2018), which could be applied to merge the mutual contents and streamline the course offerings of educational institutes. One can also rank and compare institutions on the basis of their similarity with job markets. Expanding our analysis to cover other countries as well is also important in preparing students for global careers and integrating international perspective into curricula. Finally, from the methodology point of view, there are different supervised approaches to extracting topics of textual data, such as finding hard and soft skills (see, e.g., Gugnani and Misra, 2020; Smith et al., 2021; Zhang et al., 2022). Development of a customized, multilingual model using both job advertisements and theses data types could be more optimal than applying a generic topic model, such as Annif.

Looking forward, our goal is to build an AI-based matchmaking platform that will suggest relevant jobs to students based on their curriculum and thesis projects, while also guiding academicians on course review. Topic clusters and time series analysis allow prediction for rising and declining skills in the future. This could help universities and teachers to develop teaching to respond to developments in job markets continuously. Such a platform has the potential to revolutionize the relationship between industry and academia.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.theseus.fi; https://tyomarkkinatori.fi; https://www.keha-keskus.fi.

Author contributions

JK: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. UK: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. LA: Conceptualization, Investigation, Supervision, Validation, Writing – original draft, Writing – review & editing. AN: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. AK: Conceptualization, Investigation, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the AI Forum project (OKM/116/523/2020) funded by the Ministry of Education.

Acknowledgments

We thank support of 3AMK alliance network (https://www.3amk.fi/en).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2024.1322774/full#supplementary-material

References

Ahmed, M., Mukhopadhyay, M., and Mukhopadhyay, P. (2023). Automated knowledge organisation: AI/ML-based subject indexing system for libraries. *DESIDOC J. Libr. Inf. Technol.* 43, 45–54. doi: 10.14429/djlit.43.01.18619

Bai, B., Weston, J., Grangier, D., Collobert, R., Sadamasa, K., Qi, Y., et al. (2009). Supervised semantic indexing. *Int. Conf. Inf. Knowl. Manag. Proc.* 2009, 187–196. doi: 10.1145/1645953.1645979

Baker, K. (2005). *Singular value decomposition tutorial*. Columbus, OH: The Ohio State University, p. 24.

Bellman, R., and Kalaba, R. (1961). Reduction of dimensionality, dynamic programming, and control processes. *J. Basic Eng.* 83, 82–84. doi: 10.1115/1.3658896

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.

Brasse, J., Förster, M., Hühn, P., Klier, J., Klier, M., and Moestue, L. (2023). Preparing for the future of work: a novel data-driven approach for the identification of future skills. *J. Bus. Econ.* 2023:23. doi: 10.1007/s11573-023-01169-1

Buehling, K., and Geissler, M. (2022). "PhDs with industry partners – assessing collaboration and topic distribution using a text mining methodology" in *University-Industry Knowledge Interactions*. eds. J. M. Azagra-Caro, P. D'Este and D. Barberá-Tomás (Berlin: Springer), 9–24.

Chen, L. (2022). Current and future artificial intelligence (AI) curriculum in business school: a text mining analysis. *J. Inf. Syst. Educ.* 33, 416–426.

Christian, H., Agus, M. P., and Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech* 7, 285–294. doi: 10.21512/comtech.v7i4.3746

Cleveland, W. S. (1981). LOWESS: a program for smoothing scatterplots by robust locally weighted regression. *Am. Stat.* 35:54. doi: 10.2307/2683591

Davies, D. L., and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-1, 224–227. doi: 10.1109/TPAMI.1979.4766909

Dawson, N., Williams, M. A., and Rizoiu, M. A. (2021). Skill-driven recommendations for job transition pathways. *PLoS One* 16:4722. doi: 10.1371/journal.pone.0254722

Dumais, S. T. (2004). Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.* 38, 188–230. doi: 10.1002/aris.1440380105

Gugnani, A., and Misra, H. (2020). Implicit skills extraction using document embedding and its use in job recommendation. *Proc. AAAI Conf. Artif. Intell.* 34, 13286–13293. doi: 10.1609/aaai.v34i08.7038

Gurcan, F., and Cagiltay, N. E. (2019). Big data software engineering: analysis of knowledge domains and skill sets using LDA-based topic modeling. *IEEE Access* 7, 82541–82552. doi: 10.1109/ACCESS.2019.2924075

Halko, N., Martinsson, P. G., and Tropp, J. A. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* 53, 217–288. doi: 10.1137/090771806

Hilliger, I., Aguirre, C., Miranda, C., Celis, S., and Pérez-Sanagustín, M. (2022). Lessons learned from designing a curriculum analytics tool for improving student learning and program quality. *J. Comput. High. Educ.* 34, 633–657. doi: 10.1007/s12528-022-09315-4

Hofmann, T. (2013). *Probabilistic latent semantic analysis*. ArXiv.

Inkinen, J., Lehtinen, M., and Suominen, O. (2023). Annifin ehdotusten osuvuus on parantunut Theseus-julkaisuarkistossa. *Tietolinja* 1:282.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). *Bag of tricks for efficient text classification. Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics: Volume 2, short papers*, pp. 427–431.

Ketamo, H., Moisio, A., Passi-Rauste, A., and Alamäki, A. (2019). *Mapping the future curriculum: adopting artificial intelligence and analytics in forecasting competence needs*. Proceedings of the 10th European Conference on Intangibles and Intellectual Capital ECIIC 2019.

Khaouja, I., Kassou, I., and Ghogho, M. (2021). A survey on skill identification from online job ads. *IEEE Access* 9, 118134–118153. doi: 10.1109/ACCESS.2021.3106120

Kivijärvi, M., and Sintonen, T. (2021). The stigma of feminism: disclosures and silences regarding female disadvantage in the video game industry in US and Finnish media stories. *Fem. Media Stud.* 2021, 1–19.

Kumar, M., Anderson, M. J., Antony, J. W., Baldassano, C., Brooks, P. P., Cai, M. B., et al. (2021). BrainIAK: the brain imaging analysis kit. *Aperture Neuro* 1:411. doi: 10.52294/31bb5b68-2184-411b-8c00-a1dacb61e1da

MacQueen, J. (1967). Classification and analysis of multivariate observations. 5th Berkeley Symp. *Math. Statist. Probability* 1, 281–297.

Martin, A. J., Milne-Home, J., Barrett, J., Spalding, E., and Jones, G. (2000). Graduate satisfaction with university and perceived employment preparation. *J. Educ. Work.* 13, 199–213. doi: 10.1080/713676986

Matsuda, Y., Takayuki, S., and Kazunori, Y. (2018). Curriculum analysis of computer science departments by simplified, supervised LDA. *J. Inf. Process.* 26, 497–508. doi: 10.2197/ipsjjip.26.497

McInnes, L., Healy, J., and Melville, J. (2018). Umap: uniform manifold approximation and projection for dimension reduction. *ArXiv* 2018:3426. doi: 10.48550/arXiv.1802.03426

Moore, T., and Morton, J. (2017). The myth of job readiness? Written communication, employability, and the 'skills gap'in higher education. *Stud. High. Educ.* 42, 591–609. doi: 10.1080/03075079.2015.1067602

Oraison, H., Konjarski, L., and Howe, S. (2019). Does university prepare students for employment?: alignment between graduate attributes, accreditation requirements and industry employability criteria. *J. Teach. Learn. Grad. Employability* 10, 173–194. doi: 10.21153/jtlge2019vol10no1art790

Pejic-Bach, M., Bertoncel, T., Meško, M., and Krstić, Ž. (2020). Text mining of industry 4.0 job advertisements. *Int. J. Inf. Manag.* 50, 416–431. doi: 10.1016/j.ijinfomgt.2019.07.014

Pitukhin, E., Varfolomeyev, A., and Tulaeva, A. (2016). *Job advertisements analysis for curricula management: the competency approach*. In: 9th annual international conference of education, research and innovation proceedings, Seville, pp. 2026–2035.

Rios, J. A., Ling, G., Pugh, R., Becker, D., and Bacall, A. (2020). Identifying critical 21st-century skills for workplace success: a content analysis of job advertisements. *Educ. Res.* 49, 80–89. doi: 10.3102/0013189X19890600

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Smith, E., Weiler, A., and Braschler, M. (2021). *Skill extraction for domain-specific text retrieval in a job-matching platform*. International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 116–128.

Stanton, R. (2017). Do technical/professional writing (TPW) programs offer what students need for their start in the workplace? A comparison of requirements in program curricula and job ads in industry. *Tech. Commun.* 64, 223–236.

Stanton, J. M., Kim, Y., Oakleaf, M., Lankes, R. D., Gandel, P., Cogburn, D., et al. (2011). Education for eScience professionals: job analysis, curriculum guidance, and program considerations. *J. Educ. Libr. Inf. Sci.* 52, 79–94.

Statistics Finland. (2013). *Employment and unemployment in 2013*. Helsinki: Statistics Finland.

Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. *LIBER Q.* 29, 1–25. doi: 10.18352/lq.10285

Suominen, O., Inkinen, J., and Lehtinen, M. (2022). Annif and Finto AI: developing and implementing automated subject indexing. *JLIS.it* 13, 265–282. doi: 10.4403/jlis.it-12740

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv* 2020:10957. doi: 10.48550/arXiv.2002.10957

Woolridge, R. W., and Parks, R. (2016). What's in and what's out: defining an industry-aligned IS curriculum using job advertisements. *J. High. Educ. Theory Pract.* 16:105.

Zhang, W., Arvanitis, A., and Al-Rasheed, A. (2012). *Singular value decomposition and its numerical computations*. Michigan: Michigan Technological University.

Zhang, M., Jensen, K. N., Sonniks, S. D., and Plank, B. (2022). Skillspan: hard and soft skill extraction from english job postings. *ArXiv* 2022:12811

Zimmer, W. K., and Keiper, P. (2021). Redesigning curriculum at the higher education level: challenges and successes within a sport management program. *Educ. Action Res.* 29, 276–291. doi: 10.1080/09650792.2020.1727348