Using data mining methods for research in co-operative education

SHIVANGI CHOPRA
LUKASZ GOLAB[1]
T. JUDENE PRETTI[2]
ANDREW TOULIS
University of Waterloo, Waterloo, Canada

This paper describes two classes of advanced data mining methods that can obtain actionable insight from co-operative education data: text mining of job descriptions and graph mining of job interview data. While these methods are not new in general, they have not been widely used in co-operative education research. A technical overview of each method is provided, followed by a case study using real data from a large North American university. The case study illustrates that the proposed methods can enable students, employers and institutions to make better data-driven decisions. For example, text mining of job descriptions can reveal sought-after skills while graph mining of interview relationships can characterize the extent of competition for jobs.

Keywords: Data mining, data science, co-operative education, text mining, graph mining

Work-integrated learning (WIL) takes many forms across the globe. In 2018, Co-operative and Work-Integrated Learning (CEWIL) Canada defined WIL as "a model and process of curricular experiential education which formally and intentionally integrates a student's academic studies within a workplace or practice setting" (CEWIL, 2018). Co-operative education (co-op) is one of the nine types of WIL described by CEWIL Canada, of which the defining characteristics include alternating periods of academic study and relevant paid work experience. Co-op and other forms of WIL have become popular as they can provide an enhanced learning experience for students and a talent pipeline for employers (Eames & Coll, 2010; Thiel & Hartley, 1997).

In general, research on co-operative education and other forms of work-integrated learning has considered three perspectives: of the student, of the employer and of the educational institution (Haddara & Skanes, 2007). From the student's perspective, much of the focus has been on the impact of co-op on skill and career growth, and on characterizing the attributes that make co-op students successful based on survey data and workplace supervisor evaluations; (see for example Coll, Zegwaard, & Hodges, 2002b; Ferns & Moore, 2012; Gault, Redington, & Schlager, 2000; Rainsbury, Hodges, Burchell, & Lay, 2002; Stengel, Young, Chaffe-Stengel, & Harper, 2010; Zegwaard & Hodges, 2003a; Drewery, Pretti, & Barclay, 2016). From the employer's perspective, there has been work on studying employer expectations; (see for example Coll, Zegwaard, & Hodges, 2002a; Hodges & Burchell, 2003; Moletsane, 2011; Nevison, Cormier, Pretti, & Drewery, 2018). In the broader context of employment, not necessarily co-op employment, there has been research on understanding what makes job advertisements attractive to prospective employees; (see for example Barber & Roehling, 1993; Feldman, Bearden, & Hardesty, 2006; Reeve & Schultz, 2004). From the institution's point of view, there are studies on assessing the effectiveness of and improving co-operative academic programs; (see for example Hays & Clements, 2011; Ralph, Walker, & Wimmer, 2009; Zegwaard & Hodges, 2003b).

[1] Corresponding author: Lukasz Golab, lgolab@uwaterloo.ca

[2] Author is Associate Editor of IJWIL. To maintain anonymity of the reviewers and preserve the integrity of the double-blind review process, the review was managed by the Editor-In-Chief outside the IJWIL administration process.

Much of the prior research on co-operative education uses data obtained by surveying or interviewing students and employers. Surveys tend to suffer from low response rates, and, as a result, datasets used in prior work contain on the order of 100 datapoints or fewer. Likewise, with research based on data collected through interviews, the datasets are also typically quite small. In contrast, the observation leveraged in this paper is that a co-op process at a large university generates a variety of interesting data that can be collected and analyzed: textual data such as job descriptions, relational data denoting which student applied to/interviewed with which employer, and numeric data such as workplace evaluations. While large-scale data mining methods have not yet been widely applied in co-op education research, there is prior work in the workforce literature that uses data mining to, for example, predict retention from performance evaluations (Chien & Chen, 2008), extract skills from job postings (Aken, Litecky, Ahmad, & Nelson, 2010) and recommending jobs to social network users by matching job descriptions with user profiles (Diaby, Viennet, & Launay, 2013; Malherbe, Diaby, Cataldi, Viennet, & Aufaure, 2014). The goal of this paper is twofold: 1) to provide a technical overview and a discussion of the pros and cons of advanced data analysis methods that can be applied to co-op data, and 2) using real data from a large North American university, to demonstrate the kinds of actionable insight that can be obtained through the proposed methods.

This paper focuses on two advanced data analysis methods: text mining and graph mining. In general, the purpose of data mining methods is to automatically extract potentially useful information and interesting patterns from ``raw'' data: for example numbers, text, and database tables, (Witten, Frank, Hall, & Pal, 2016). Specifically, text mining methods aim to discover semantics or context behind a document or a piece of text, such as the topic being described or the sentiment being expressed. Additionally, text clustering and categorization methods can be used to group similar documents together (Gupta & Lehal, 2009). On the other hand, graph mining methods are widely used in social network analysis and can discover groups of closely connected users; for example, groups of Facebook friends or groups of Twitter users who follow or re-tweet each other (Tang & Liu, 2010).

As will be shown in this paper, these methods can be useful for research in co-operative education. For example, job descriptions are natural candidates for text mining methods that can identify frequently occurring words (corresponding to, e.g., sought-after skills), cluster the available jobs into job categories or types, and show how the required skills and available jobs have changed over time. This information can enable universities to track employers' talent needs, inform students about the types of co-op jobs available to them, and help the institution to align its curriculum with job market needs. Notably, analyzing historical data captured through years of submitted job advertisements can be significantly more effective than surveying or interviewing a sample of employers, where researchers would have to rely on employers' recollection of, for example, how the job requirements have changed.

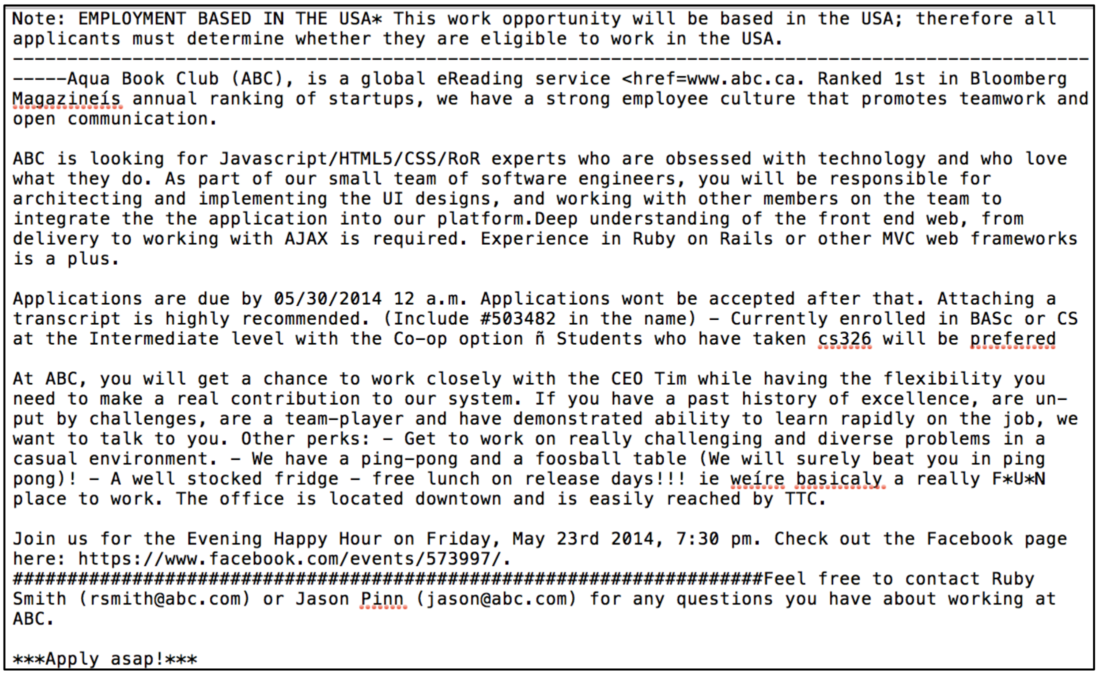
Furthermore, graph mining methods can be applied to co-op interview data. The idea is to represent the interview process as a 'social network', with students connected to other students if they have interviewed for at least one job in common, and jobs connected with other jobs if they interview the same students. Graph mining methods can then be applied to identify densely-connected communities, which are groups of students or employers who compete with each other. Characterizing the extent of competition in a co-op market is an important problem. For example, employers may not have a good understanding of the available talent pool and may not be allocating their recruiting resources effectively. Likewise, students may not be aware of the extent of competition for various types of jobs, and therefore they may not know which jobs are realistically within their reach.

For completeness, it is worth noting that numeric workterm evaluation data can also be analyzed to determine if students and employers are satisfied with each other. However, workterm evaluations are not discussed in this paper as they can be analyzed using standard statistical methods (see Jiang, Lee, & Golab, 2015 for recent work on this topic). For similar reasons, recent work on data analysis to understand the connection between co-operative education and entrepreneurship is also omitted (Andrade, Chopra, Nurlybayev, & Golab, 2018).

The next two sections describe text mining and graph mining methods, respectively, in more detail. Each section includes an example-driven description of the corresponding method, followed by real-life examples of actionable insights that can be obtained and a discussion of caveats and limitations. Finally, the paper concludes with directions for future work.

TEXT MINING METHODS

This section discusses text analysis of co-op job descriptions, which are a rich source of information about the desired skills, company culture and working environments. The goal is to extract and cluster informative terms from job descriptions: technical skills, soft skills, perks (e.g., free food or proximity to public transport and other terms indicating the nature of the job. This information can then be used to understand employers' talent needs and to inform students about the types of available co-op jobs.



```
Note: EMPLOYMENT BASED IN THE USA* This work opportunity will be based in the USA; therefore all
applicants must determine whether they are eligible to work in the USA.
----------------------------------------------------------------------------------------------------
-----Aqua Book Club (ABC), is a global eReading service <href=www.abc.ca. Ranked 1st in Bloomberg
Magazineís annual ranking of startups, we have a strong employee culture that promotes teamwork and
open communication.

ABC is looking for Javascript/HTML5/CSS/RoR experts who are obsessed with technology and who love
what they do. As part of our small team of software engineers, you will be responsible for
architecting and implementing the UI designs, and working with other members on the team to
integrate the the application into our platform.Deep understanding of the front end web, from
delivery to working with AJAX is required. Experience in Ruby on Rails or other MVC web frameworks
is a plus.

Applications are due by 05/30/2014 12 a.m. Applications wont be accepted after that. Attaching a
transcript is highly recommended. (Include #503482 in the name) - Currently enrolled in BASc or CS
at the Intermediate level with the Co-op option ñ Students who have taken cs326 will be prefered

At ABC, you will get a chance to work closely with the CEO Tim while having the flexibility you
need to make a real contribution to our system. If you have a past history of excellence, are un-
put by challenges, are a team-player and have demonstrated ability to learn rapidly on the job, we
want to talk to you. Other perks: - Get to work on really challenging and diverse problems in a
casual environment. - We have a ping-pong and a foosball table (We will surely beat you in ping
pong)! - A well stocked fridge - free lunch on release days!!! ie weíre basicaly a really F*U*N
place to work. The office is located downtown and is easily reached by TTC.

Join us for the Evening Happy Hour on Friday, May 23rd 2014, 7:30 pm. Check out the Facebook page
here: https://www.facebook.com/events/573997/.
#########################################################################Feel free to contact Ruby
Smith (rsmith@abc.com) or Jason Pinn (jason@abc.com) for any questions you have about working at
ABC.

***Apply asap!***
```

FIGURE 1: A sample job description.

Method Description

Figure 1 shows an anonymized example of a job description targeting undergraduate computer science students. It includes the following information:

- Required technical skills: Javascript, Ruby on Rails;
- Required soft skills: team player, ability to learn;
- Job duties: architecting and implementing UI (user interface) designs;
- Desired mindset and attitude: obsessed with technology, love what they do;
- Perks: ping-pong table, free lunch; and
- Company culture: casual environment.

In practice, job descriptions are written directly by employers, and therefore they are not standardized or well-structured. They may include administrative and formatting elements such as links to company websites, contact emails, timestamps, and of course common English words. Thus, the technical challenge is to extract useful information from job descriptions (and then to cluster or group the job descriptions to determine what types of jobs are available).

To address this challenge, the proposed text mining method starts with a parser that extracts relevant and frequently occurring words from unstructured job descriptions. Afterwards, existing text mining tools such as Latent Semantic Analysis (LSA) and text clustering are used to characterize the types of available jobs.

The parser, implemented in Python, works as follows. To remove unnecessary words, a vocabulary (i.e., a list of words) is created, call it list A, consisting of publicly available lists of stopwords (Bird, Klein, & Loper, 2015), common English words (Pearson Education, 2007), misspellings ("Lists of common misspellings/For machines," n.d.) and abbreviations, company names, locations and persons' names. However, one must be careful to not remove informative terms. For example, Ajax is a town in Canada and would be included in the vocabulary of terms that can be removed. However, Ajax is also a web development toolkit. To address this problem, another vocabulary is created, call it list B, of words that should not be removed. This vocabulary consists of terms listed as skills on a resume help web site (The Balance Careers, 2013) and terms listed as job duties in the Canadian National Occupation Classification (Government of Canada, 2013). Note that list B only contains a subset of relevant words; for example, it is missing many specific technical skills, perks and company culture descriptors.

The parser handles various elements of linguistic morphology (Bauer, 2003), including inflections (different forms of the same word to reflect tense or voice; e.g., work, worked and working are all converted to 'work' and all contribute towards the frequency count of 'work'), derivations (e.g., un-professional, un-cut), contractions (e.g., you're, we're), inconsistencies in writing (e.g., java script, java-script, javascript) using common techniques from the field of information retrieval (Croft, Metzler & Strohman, 2015; Porter, 2001). Additional details of the parser can be found in (Chopra, 2017).

To summarize the parsing step, each job advertisement is parsed, words are standardized, and words occurring in list A but not list B are removed.

The second part of the proposed method is designed to analyze the extracted job attributes. This is done in two ways:

1. To highlight popular skills, attitudes, working environments and perks, words that occur at least once in a large percentage of job descriptions are reported and visualized.
2. Next, clustering is used to identify the different types of available co-op jobs within an academic discipline. Following previous work on text clustering (Ding & He, 2004; Song & Park, 2007; Sorour, Mine, Goda, & Hirokawa, 2014), Latent Semantic Analysis (LSA) (Scikit-learn developers, 2011) is first applied to job descriptions, with each job description represented as a job vector. The ith coordinate of a job vector is equal to the inverse document frequency (IDF) (Croft, et al., 2015) of the ith word in the set of all possible words, provided that this word is mentioned in the given job description at least once (and zero otherwise). This way, less common words are given higher weights. For example, if a word occurs in 1000 job descriptions at least once, the corresponding co-ordinate of these 1000 job description vectors is set to $\frac{1}{1000}$, whereas if it occurs only in ten job descriptions, the corresponding vector entries of those ten job descriptions are $\frac{1}{10}$. The purpose of LSA is to reduce the dimensionality of job vectors from the number of distinct words down to one hundred. Each reduced dimension corresponds to a latent concept in the data. Finally, a clustering algorithm (Jain, 2010) is applied to the transformed job vectors, and a few top terms (again, ranked by IDF) from each cluster are extracted and reported as representatives. The output of a clustering algorithm consists of a disjoint partitioning of the data such that datapoints (i.e., job vectors) within the same part are similar (i.e., include similar latent concepts) and datapoints in different parts are dissimilar. The clustering algorithm chosen in this paper is k-means, which is simple and widely used, but other clustering algorithms could also be applied (Jain, 2010).

Insights

This section gives examples of insights that can be obtained by analyzing job descriptions using the proposed method (see Chopra, 2017; Chopra & Golab, 2018 for further details). The dataset used in this analysis was collected by a large North American undergraduate institution and consists of all 17,057 job postings that were advertised and filled in 2014.

The first example invovles a comparison of jobs obtained by information technology (IT) students (those studying computer science or software engineering) with those obtained by finance students (those studying accounting, actuarial sciences or finance). After identifying frequently occurring words in the job descriptions, they are visualized as word clouds in Figure 2, with very frequent words written in large font for emphasis. Soft skills are highlighted in green. Note that soft skills such as communication, teamwork and learning are frequent in both IT and finance jobs; this emphasizes the importance of soft skills in post-secondary curricula. However, as expected, the technical skills are different: IT jobs mention programming languages such as C++ and Java whereas finance jobs are more likely to mention Microsoft Excel and accounting. Upon closer inspection, the top five sought-after programming languages in IT jobs were found to be Java (mentioned in 33 percent of job postings), C++ (33 percent), JavaScript (31 percent), C (24 percent) and Python (22 percent). There are also interesting differences in the descriptions of mindsets and work environments: IT jobs are more likely to mention passion, creativity and love (of technology) whereas finance jobs mention client relationships and interpersonal skills.

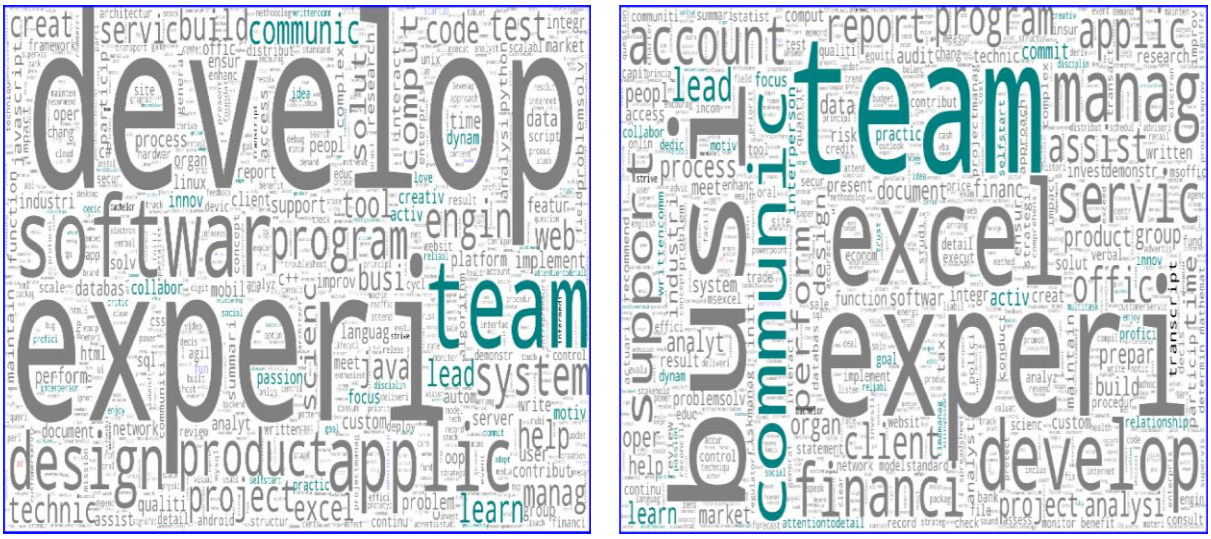


FIGURE 2: Word clouds of terms from information technology (left) and finance (right) job descriptions.

Next, two Venn diagrams are shown in Figure 3, which characterize the overlap between junior jobs (obtained by lower-year students in years 1 and 2) and senior jobs (obtained by upper-year students in years 3 and 4); while word clouds can effectively visualize frequent words, venn diagrams are useful for visualizing two intersecting sets of words. Again, IT is on the left and finance is on the right. It appears that all IT and finance jobs require soft skills such as communication and collaboration. However, junior IT jobs require scripting and HTML, whereas senior IT jobs mention advanced technologies such as distributed and scalable systems and security. Furthermore, common words in junior finance job descriptions include file, arrange, update and Microsoft Office, which suggests clerical and data entry positions. On the other hand, senior finance jobs are more likely to mention risk managing, statistics, modeling and investing. These results can help manage the expectations of junior students: it may take until senior years to obtain a co-op position that leverages advanced skills and technologies.

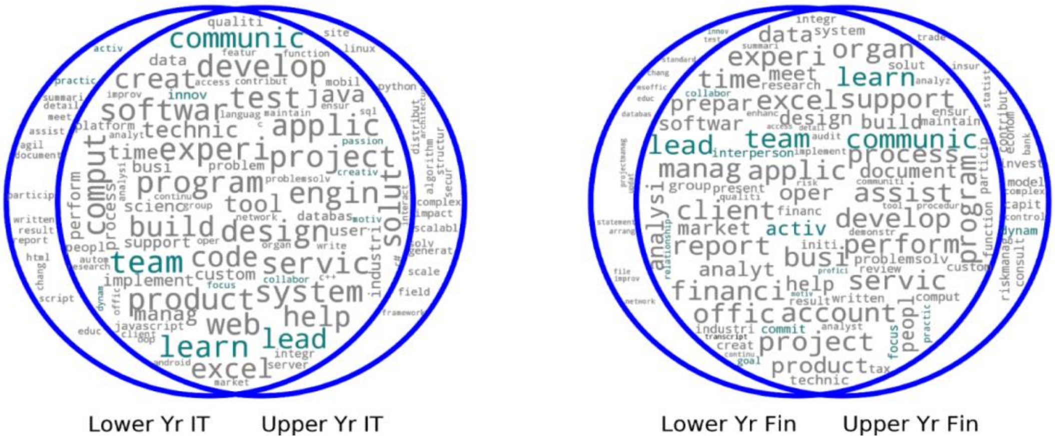


FIGURE 3: Overlap between the top 100 most frequent attributes in junior and senior information technology (left) and finance (right) job descriptions.

After investigating frequently occurring words, the next example shows a clustering of IT job descriptions to understand the types of available jobs. Different numbers of clusters were tested between 2 and 30 (the k-means clustering algorithm requires the number of clusters as input). Results using ten clusters are shown in Table 1; using fewer clusters led to different types of jobs being assigned to the same cluster, whereas using more clusters led to similar types of jobs belonging to multiple clusters.

Table 1 shows the six largest clusters in 2014 sorted by size; the remaining four clusters had under 2% of the total number of jobs each. The table include representative words from each cluster (which appear frequently within the cluster but not in other clusters), a manually-assigned label summarizing the job type, and three percentages: the percentage of all jobs assigned to this cluster, and the percentages of junior and senior jobs within this cluster. The higher of the last two percentages is highlighted in bold font to indicate whether a cluster consists of more junior or senior jobs.

Based on the clustering results, the IT co-op market can be categorized as follows. The five largest clusters cover 87% of IT jobs, spanning web development (22%), programming (21%), software start-ups (18%), business analysis (16%) and mobile app development (10%). Furthermore, troubleshooting jobs are mostly filled by junior students, whereas jobs mentioning company culture, many of which are startups, tend to be filled by senior students.

TABLE 1: Six largest clusters of 2014 information technology jobs, used in the research.

Label	Representative words	%All	%Jr.	%Sr.
Web Development	javascript, html, web, css, sql, c#, server, java, net, jquery	22%	**64%**	36%
Programming	c++, c, languag, linux, python, oop, scienc, algorithm, perl, script	21%	46%	**54%**
Start-up Culture	startup, python, javascript, featur, code, web, love, stack, fun, passion	18%	39%	**61%**
Business Analyst	sql, analyst, test, solut, c#, script, execut, financi, document, busi	16%	**69%**	31%
Mobile Development	io, android, mobil, app, platform, java, agil, iphon, devic, c	10%	**61%**	39%
System Administrator	hardwar, troubleshoot, configur, instal, network, desktop, server, user, xp, resolut	6%	**87%**	13%

Discussion

As shown, job description mining can reveal actionable insight for students, employers and the institution. The institution can provide students with a better understanding of co-op opportunities in various disciplines and therefore help them select the right academic program and career. Additionally, the institution may use frequently appearing words and the clustering of jobs in various disciplines to produce more effective promotional material for its co-op programs and to help attract strong students. Furthermore, students can find out what types of jobs are available to them and what soft and technical skills are required. In particular, clustering can be used to segment the job descriptions to make it easier for students to find jobs they are interested in and institutions can align their curricula with job market needs. For example, as many disciplines seem to emphasize teamwork, institutions can incorporate more team exercises in their curricula. New tools and methods may be introduced in courses when the corresponding words begin to appear in job descriptions.

In particular, text clustering is a useful tool for automatically partitioning and summarizing a large collection of loosely structured documents (such as job descriptions). For example, the results in Table 1 summarize the types of available IT jobs (and, when combined with the academic levels of students who obtained the jobs, they reveal which jobs tend to be available only to senior students). This method is especially useful for job descriptions which are usually not pre-categorized, i.e., employers are not required to identify a job category or type when submitting a description.

Job description mining results should be interpreted carefully due to the following factors.

- Diversity in size and age of companies, for example, the IT discipline has many modern companies that emphasize a fun work culture, while other disciplines such as finance have more traditional companies which might emphasize client relationships.
- Incorrect job descriptions which may not reflect the true nature of the job; for example, employers may write or modify the job descriptions to suit the company's public image.

A limitation of many clustering algorithms such as k-means is that the desired number of clusters must be specified upfront. This is usually difficult to do; for example, it is not clear how many types of co-op jobs exist in a given job description dataset. As a result, analysts often need to run clustering algorithms multiple times with different numbers of clusters, as was the case in this paper.

Finally, a general limitation of data analysis methods is that they largely focus on the question of *what* rather than *why*. In other words, they can identify interesting patterns and correlations in the data but they usually cannot explain why they occur without additional statistical analysis or without collecting additional information. For example, the job description used earlier the mining results revealed that senior IT students tend to obtain jobs that mention a fun and engaging work environments. However, it cannot immediately be concluded that all employers should include such comments in their job descriptions to increase their chances of attracting top talent. Perhaps top students are mainly attracted to large established companies and a fun working environment is an added bonus. Thus, further analysis and perhaps interviews with students and employers would be required to confirm any cause-and-effect relationships.

GRAPH MINING METHODS

The previous section discussed job description mining to understand what skills employers are looking for and what types of jobs they offer to co-op students. After advertising jobs, the next step in the co-op process is for employers to select candidates for interviews. This section shows how to use graph mining methods on interview data to determine which groups of students and employers compete with each other.

Method Description

The first step is to transform interview data into a graph, i.e., a collect of nodes (or vertices) and edges that connect the nodes. It is assumed that interview data consist of (student id, job id) pairs. Two graphs are then constructed: a student graph, in which two students are connected if they interview for at least one job in common, and a job graph, in which two jobs are connected if they interview at least one student in common. Next, a community detection algorithm is executed on both graphs. The Louvain Method is used in this paper (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008; Lambiotte, Delvenne, & Barahona 2008), which is implemented in the Networkx Python package (Hagberg, Swart, & Schult, 2008). The goal of community detection is to segment or cluster the nodes in a graph such that nodes belonging to the same cluster or community are strongly connected while nodes in different communities are sparsely connected.

To illustrate the method, Table 2 shows a simple example, similar to that shown in (Toulis & Golab, 2017), which describes interviews of nine students (labelled 1-9) for eight jobs (labelled A-H). Figure 4 shows the corresponding student and job graphs. The job graph contains two communities. Note that jobs within the same community are connected to each other, but not to jobs in other communities. Communities in the student graph can then be correlated to the job communities in which the students had the most interviews. For example, students 1-5, who make up student community 1, interviewed for jobs A-D, which make up job community 1.

TABLE 2: Example table of interviews describing nine students (labelled 1-9) interviewing for eight jobs (labelled A-H)

Job ID	Student IDs
A	1, 2
B	1, 2
C	2
D	1, 3, 4, 5
E	5, 6
F	6, 7, 8, 9
G	7, 8, 9
H	7

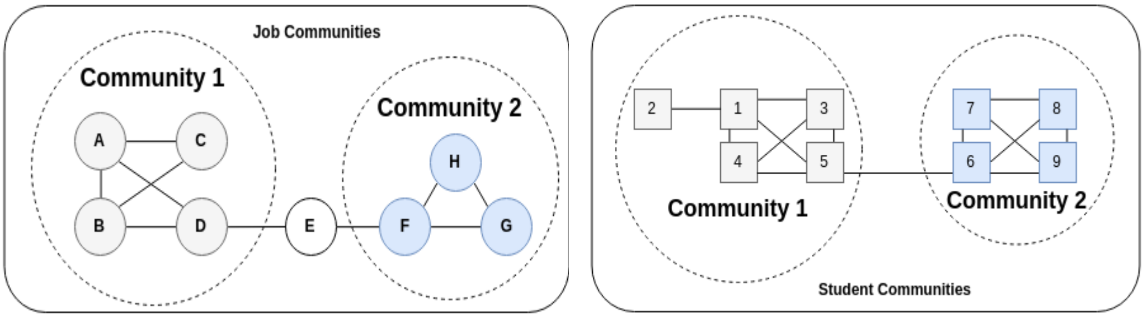


FIGURE 4: On the left, a job graph is shown based on the data from Table 2. On the right, a student graph is shown based on the data from Table 2.

In addition to community detection, it is useful to identify nodes with high *closeness centrality*, which are nodes with the smallest average shortest path length to other nodes, where shortest path length indicates the fewest edges that would have to be traversed to reach one node from another. For example, in Figure 4, the shortest path length between nodes E and G is two. These nodes (jobs) are noteworthy as they are likely to be multi-disciplinary positions that interview a diverse set of students and compete with a diverse set of other jobs for these students. For example, node E in figure 4 has high closeness centrality as it is close to (and competes for students with) jobs in job community 1 and job community 2.

Insights

This section illustrates the benefits of graph mining using a dataset from the same institution as before, corresponding to all interviews that took place in the summer of 2016. This dataset consist of 4100 undergraduate students and nearly 2000 jobs offered by 700 distinct employers. On average, each student had 3.5 interviews, and each job interviewed 7 students. The dataset also shows which student was ultimately hired for each position, though some positions remained unfilled.

The results presented below characterize the competition for IT jobs among computer science and software engineering students; see Toulis & Golab (2017) for full details and see Jiang & Golab (2016), for a graph-mining study on the competition for co-op jobs among academic programs.

In general, the Louvain Method found eight clusters in the job graph, three of which contained mostly IT jobs. Upon further inspection, it was possible to establish a ranking of these three communities.

- The first community contained sought-after IT jobs at large companies such as Facebook and Google. Most of the students who interviewed for these jobs were seniors (in their third or fourth years of undergraduate study).
- The second community contained small IT companies and start-ups which mostly interviewed and hired junior students (in their second year of study).
- The third community had mostly quality assurance and software testing jobs, which are perceived by students as less desirable work. Most students competing for these jobs were in their first year of study and had little prior work experience.

When analyzing competition, it was found that some small IT companies and start-ups from the second community interviewed the same students as the established companies from the first community.

However, a majority of these top students accepted positions from top IT companies, and the smaller companies ended up hiring more junior students.

Furthermore, it was found that some jobs in the first IT community did not hire any students. Manual inspection of these unfilled jobs showed that they were outliers: their descriptions were interesting enough to attract students, but their actual quality was lower. For example, some of these jobs exaggerated the role a student would be hired for, for example, "Big Data Hacker". The students these jobs interviewed were hired by other top jobs in the first community.

These results suggest that the smaller companies which are able to attract significant student attention are underestimating their competition and have difficulties competing for top co-op talent.

Interestingly, centrality analysis revealed that the most central job in the top-tier IT community was a data scientist position, suggesting that data science roles are more multi-disciplinary than traditional IT positions.

Discussion

Analyzing co-op interview data through graph mining can reveal the nature of competition in the job market, including the top employee and employer clusters. Student community analysis revealed the types of jobs that different types of students can obtain, whereas job community analysis showed which employers compete for the same talent. As a result, both students and employers can be better informed about the job market. In particular, the findings can be used to manage the expectations of small start-up companies participating in a large IT co-op job system: some of these companies may be unaware of the level of competition they face and they may want to consider targeting more junior students.

The main advantage of graph mining methods is that they consider relationships within the data. For example, text mining methods can cluster job descriptions and identify the types of available jobs, whereas graph mining methods identify competing jobs and students. In Table 2, it is not clear which students and jobs compete with each other by looking at the data, but representing interview relationships as graphs in Figure 4 clarifies the extent of competition.

However, before using graph mining methods, it must be ensured that the relationships used to create the graph are meaningful; otherwise, finding densely connected communities or central nodes does not make sense. In social media, relationships are clear: Facebook friends, LinkedIn connections, Twitter followers or similar. In co-op datasets related to competition, there are at least two choices: two students can be connected if they applied to, or interviewed for, at least one job in common; similarly, two jobs can be connected if at least one student applied to or interviewed for both of them. Interview relationships were analyzed in this paper because they are arguably stronger than application relationships: any student can in principle apply to any job, but in order to obtain an interview, a student must be perceived by the employer as qualified for the job. In other words, just because the same students applied to the same jobs does not immediately mean that these students compete with each other.

CONCLUSIONS AND FUTURE WORK

This paper presented two data-intensive methods for mining co-operative education data: text mining of job description and graph mining of student-employer interview relationships. Using real data, it was shown that these methods can lead to new data-driven insight that provides a deeper understanding of the co-op job market. Students can use the results to better target and prepare for co-op opportunities. Employers can better understand the competition they face and avoid missed hiring opportunities. The institution can obtain a better understanding of employers' needs and the types of co-op jobs available to students in different academic disciplines. These examples have shown that data mining techniques are not just powerful for numeric data, but also in deriving insights from text-based and graph-based data. In general, it can be argued that co-operative education is an important new application area that showcases the power of data analytics, machine learning and data-driven decision making.

From a practical perspective, there are several challenges in undertaking these type of data analytics techniques for research. There needs to be a high level of confidence in the integrity of the data, and it may require significant time in cleaning datasets prior to analysis. Researchers and data custodians need to work closely together to ensure a thorough understanding of the mapping between business problems and research questions and thus the selection of the appropriate data analysis technique. Following the application of the technique, they must also work closely to ensure proper understanding of data and interpretation of the results.

Naturally, there is more data-driven work that can be done. An immediate extension of the proposed text mining methods is to apply them to students' resumes in order to understand what skills students have to offer and what types of students employers can select from. Furthermore, combining job description mining with resume mining can determine whether there exists a gap between employers' needs and students' talents. Similarly, text mining of course descriptions can help determine whether there is a gap between what employers are looking for and what is being taught in classrooms.

Additionally, an interesting extension of the proposed graph mining methods is to build a recommender system for students and employers. The idea is to recommend new jobs to apply for or new students to interview based on nodes that are close to the given student or job in the graph.

Furthermore, prediction methods may be useful in a co-op context. For example, a university admission process could benefit from a model that predicts whether a student will be successful in a co-operative program based on their high school work and extra-curricular background.

ACKNOWLEDGEMENTS

This submission is based on a paper presented at DARLI-AP 2018: the 2nd International workshop on Data Analytics Solutions for Real-Life Applications (Chopra, Jiang, Toulis, & Golab, 2018), and includes a more detailed overview of data-intensive methods, their value and their limitations.

REFERENCES

Aken, A., Litecky, C., Ahmad, A., & Nelson, J. (2010). Mining for computing jobs. *IEEE software*, *27*(1), 78-85. https://doi.org/10.1109/MS.2009.150

Andrade, A., Chopra, S., Nurlybayev, B., & Golab, L. (2018). Quantifying the impact of entrepreneurship on cooperative education job creation. *International Journal of Work-Integrated Learning*, *19*(1), 51-68.

Barber, A. E., & Roehling, M. V. (1993). Job postings and the decision to interview: A verbal protocol analysis. *Journal of Applied Psychology*, *78*(5), 845. https://doi.org/10.1037/0021-9010.78.5.845

Bauer, L. (2003*). Introducing linguistic morphology*. Edinburgh, UK: Edinburgh University Press

Bird S., Klein E., & Loper, E. (2015, July 1). Accessing Text Corpora and Lexical Resources. Retrieved from http://www.nltk.org/book/ch02.html

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), 10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

CEWIL (Co-Operative Education And Work-Integrated Learning Canada), *Co-operative education definition* (2018). Retrieved from https://www.cewilcanada.ca/coop-defined.html

Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert systems with applications*, *34*(1), 280-290. https://doi.org/10.1016/j.eswa.2006.09.003

Chopra, S. (2017). *Job description mining to understand undergraduate co-operative placements* (Master's thesis, University of Waterloo). Retrieved from https://uwspace.uwaterloo.ca/handle/10012/12319

Chopra, S., & Golab, L. (2018). Job description mining to understand work-integrated learning. In K. E. Boyer, & M. Yudelson (Eds.), *Proceedings of the 11th International. Conference on Educational Data Mining (EDM)* (pp. 32-43). Buffalo, NY: Retrieved from http://educationaldatamining.org/files/conferences/EDM2018/EDM2018_Preface_TOC_Proceedings.pdf http://educationaldatamining.org/files/conferences/EDM2018/EDM2018_Preface_TOC_Proceedings.pdf

Chopra, S., Jiang, Y., Toulis, A., & Golab, L. (2018). Data analytics to improve co-operative education. In N. Augsten (Ed.), *Proceedings of the workshops of the EDBT/ICDT 2018 Joint Conference: International Workshop on Data Analytics Solutions for Real-Life Applications* (pp. 16-21). Vienna, Austria: University of Salzburg, Austria. Retrieved from: http://ceur-ws.org/Vol-2083/paper-03.pdf

Coll, R. K., Zegwaard, K., & Hodges, D. (2002a). Science and technology stakeholders' ranking of graduate competencies Part 1: Employer perspective. *Asia-Pacific Journal of Cooperative Education*, *3*(2), 19-28.

Coll, R. K., Zegwaard, K., & Hodges, D. (2002b). Science and technology stakeholders' ranking of graduate competencies Part 2: Students perspective. *Asia-Pacific Journal of Cooperative Education, 3*(2), 35-44.

Croft, W. B., Metzler, D., & Strohman, T. (2015). *Search engines: Information retrieval in practice.* Retrieved from http://ciir.cs.umass.edu/downloads/SEIRiP.pdf

Diaby, M., Viennet, E., & Launay, T. (2013, August). Toward the next generation of recruitment tools: An online social network-based job recommender system. In T. Ozyer, P. Carrington, & E. LIM (Eds.), *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)* (pp. 821-828). Niagara, ON, Canada: IEEE. Retrieved from https://doi.org/10.1145/2492517.2500266

Ding, C., & He, X. (2004, July). K-means clustering via principal component analysis. In C. E. Brodley (Ed.), *Proceedings of the Twenty-First* International *Conference on Machine learning (ICML 2004)* (p. 29). Banff, Alberta, Canada: ACM. Retrieved from https://doi.org/10.1145/1015330.1015408

Drewery, D., Pretti, T. J., & Barclay, S. (2016). Examining the effects of perceived relevance and work-related subjective well-being on individual performance for co-op students. *Asia-Pacific Journal of Cooperative Education, 17*(2), 119-134.

Eames, C., & Coll, R. K. (2010). Cooperative education: Integrating classroom and workplace learning. In *Learning through practice* (pp. 180-196). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-90-481-3939-2_10

Feldman, D. C., Bearden, W. O., & Hardesty, D. M. (2006). Varying the content of job advertisements: The effects of message specificity. *Journal of Advertising*, *35*(1), 123-141. https://doi.org/10.2753/JOA0091-3367350108

Ferns, S., & Moore, K. (2012). Assessing student outcomes in fieldwork placements: An overview of current practice. *Asia-Pacific Journal of Cooperative Education*, *13*(4), 207-224.

Gault, J., Redington, J., & Schlager, T. (2000). Undergraduate business internships and career success: Are they related?.*Journal of Marketing Education*, *22*(1), 45-53. https://doi.org/10.1177/0273475300221006

Government of Canada (2013, Decemeber 13). National Occupational Classification 2016. Retrieved from http://noc.esdc.gc.ca/English/noc/welcome.aspx?ver=16

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications.*Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60-76. https://doi.org/10.4304/jetwi.1.1.60-76

Haddara, M., & Skanes, H. (2007). A reflection on cooperative education: From experience to experiential learning. *Asia-Pacific Journal of Cooperative Education*, *8*(1), 67-76.

Hagberg, A., Swart, P., & S Chult, D. (2008*). Exploring network structure, dynamics, and function using NetworkX* (No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos, NM: Los Alamos National Lab. (LANL).

Hays, J., & Clements, M. (2011, June). Supervision in work experience for learning programs. In K. Betts (Ed.), *Proceedings of the 17th World Conference on Cooperative and Work-Integrated Education (WACE)*. Philadelphia, USA: Drexel University Retrieved from http://www.waceinc.org/philly2011/conference_proceedings/Refereed%20Papers/Australia/JAYHAY~1.PDF

Hodges, D., & Burchell, N. (2003). Business graduate competencies: Employers' views on importance and performance. *Asia-Pacific Journal of Cooperative Education*, *4*(2), 16-22.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

Jiang, Y. H., & Golab, L. (2016). On competition for undergraduate co-op placements: A graph mining approach. In *EDM* (pp. 394-399).

Jiang, Y. H., Lee, S. W. Y., & Golab, L. (2015). Analyzing student and employer satisfaction with cooperative education through multiple data sources. *Asia-Pacific Journal of Cooperative Education*, *16*(4), 225-240.

Lambiotte, R., Delvenne, J. C., & Barahona, M. (2008). Laplacian dynamics and multiscale modular structure in networks. Retrieved from arXiv:0812.1770v3 [physics.soc-ph].

Lists of common misspellings/For machines (n.d.). In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines

Malherbe, E., Diaby, M., Cataldi, M., Viennet, E., & Aufaure, M. A. (2014, August). Field selection for job categorization and recommendation to social network users. In X. Wu, M. Ester, & G. Xu (Eds.), *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)* (pp. 588-595). Beijing, China: IEEE. Retrieved from https://doi.org/10.1109/ASONAM.2014.6921646

Moletsane, A. (2011). Work integrated learning (WIL) stakeholder expectations in the hospitality industry. In K. Betts (Ed.), *Proceedings of the 17th World Conference on Cooperative and Work-Integrated Education (WACE)*. Philadelphia, USA: Drexel University. Retrieved from http://waceinc.org/philly2011/conference_proceedings/Non-Refereed%20Papers/South%20Africa/Annie%20Moletsane,%20Vaal%20University%20of%20Technology,%20Work%20Integrated%20Learning%20Stakeholder%20Expectations.pdf

Nevison, C., Cormier, L., Pretti, T.J. & Drewery, D. (2018). The influence of values on supervisors' satisfaction with co-op student employees. *International Journal of Work-Integrated Learning, 19(*1), 1-11.

Pearson Education Limited (2007). Longman Communication 3000. Retrieved from http://www.lextutor.ca/freq/lists_download/longman_3000_list.pdf

Porter, M. F. (2001). *Snowball: A language for stemming algorithms*. Retrieved from http://snowball.tartarus.org/texts/introduction.html

Rainsbury, E., Hodges, D. L., Burchell, N., & Lay, M. C. (2002). Ranking workplace competencies: Student and graduate perceptions. *Asia-Pacific Journal of Cooperative Education*, *3*(2), 8-18.

Ralph, E., Walker, K., & Wimmer, R. (2009). Practicum-education experiences: Post-interns' views.*International Journal of Engineering Education* *25(*1):122-130

Reeve, C. L., & Schultz, L. (2004). Job-seeker reactions to selection process information in job ads. *International Journal of Selection and Assessment*, *12*(4), 343-355. https://doi.org/10.1111/j.0965-075X.2004.00289.x

Scikit-learn developers (2011). sklearn.decomposition.TruncatedSVD. Retrieved from http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

Song, W., & Park, S. C. (2007, October). A novel document clustering model based on latent semantic analysis. In L. O'Conner (Ed.), *Third International Conference on Semantics, Knowledge and Grid (SKG 2007)* (pp. 539-542). Xian, Shan Xi, China: IEEE. doi:10.1109/SKG.2007.154

Sorour, S. E., Mine, T., Goda, K., & Hirokawa, S. (2014, April). Efficiency of LSA and K-means in predicting students' Academic performance based on their comments data. In S. Zvacek, M. T. Restivo, J. O. Uhomoibhi, & M. Helfert (Eds.), *CSEDU 2014 - Proceedings of the 6th International Conference on Computer Supported Education.* (pp. 63-74). Barcelona, Spain: SciTePress

Stengel, D., Young, D. R., Chaffe-Stengel, P., & Harper, R. M. (2010). Assessing the academic and workplace skills of undergraduate business interns. *Journal of Cooperative Education & Internships*, *44*(18), 13-22.

Tang, L., & Liu, H. (2010). Graph mining applications to social network analysis. In *Managing and Mining Graph Data* (pp. 487-513). Boston, MA: Springer https://doi.org/10.1007/978-1-4419-6045-0_16

The Balance Careers (2013, October 5). The Best Skills to Include on Your Resume. Retrieved from https://www.thebalance.com/list-of-the-best-skills-for-resumes-2062422

Thiel, G. R., & Hartley, N. T. (1997). Cooperative education: A natural synergy between business and academia.*SAM Advanced Management Journal*,*62*(3), 19.

Toulis, A., & Golab, L. (2017, May). Graph mining to characterize competition for employment. In A. Arora, S. Roy, & A. Bhattacharya (Eds.), *Proceedings of the 2nd International Workshop on Network Data Analytics: Conference SIGMOD/PODS'17 International Conference on Management of Data* (p. 3). Chicago, IL, USA: ACM. Retrieved from https://doi.org/10.1145/3068943.3068946

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016).Data mining: Practical machine learning tools and techniques (4th ed.) San Francisco, CA: Morgan Kaufmann.

Zegwaard, K. E., & Hodges, D. (2003a). Science and technology stakeholders' ranking of graduate competencies part 3: Graduate perspective. *Asia-Pacific Journal of Cooperative Education*, *4*(2), 23-35.

Zegwaard, K. E., & Hodges, D. (2003b). Science and technology stakeholders ranking of graduate competencies part 4: Faculty perspective. *Asia-Pacific Journal of Cooperative Education*, *4*(2):36-48.