Collaborative analytics and recommendations in social learning networks

Sgagia Sofia
SID: 3305170016

SCHOOL OF SCIENCE & TECHNOLOGY
A thesis submitted for the degree of

Master of Science (MSc) in E-Business and Digital Marketing

DECEMBER 2018
THESSALONIKI – GREECE
Collaborative analytics and recommendations in social learning networks

Sgagia Sofia
SID: 3305170016

Supervisor: Ioannis Magnisalis
Supervising Committee Members:
  Assoc. Prof. Name Surname
  Assist. Prof. Name Surname

SCHOOL OF SCIENCE & TECHNOLOGY
A thesis submitted for the degree of
Master of Science (MSc) in E-Business and Digital Marketing
Abstract

This dissertation was written as a part of the MSc in E-Business and Digital Marketing at the International Hellenic University. Collaborative analytics were thoroughly examined in the context of an unofficial learning social network and insight into how actual users coexist and cooperate online was gained. The aim of this thesis was also to highlight the semantic technologies and propose a schema that encapsulates semantically enriched data along with data that derive from learners’ online interactions in order to leverage users experience by exploiting the whole spectrum of learning activities and adding a new sense to practices that already take place.

In this journey of learning and practicing new things I want to my thank my supervisor Mr Ioanni Magnisali for being the most helpful companion, address and support my many questions, my family and the beloved classmates of another academic year that made clear to me the importance of collaborating, sharing and learning together. Lastly, I want to pay a special tribute to all of those people that have contributed to the collection, maintenance and ensured open access to the enormous amount of data, part of which was used to shape this research.

Sofía Sgagia
7.12.2018
Contents

ABSTRACT ................................................................................................................................. III

CONTENTS .............................................................................................................................. V

1 INTRODUCTION ................................................................................................................... 1

2 LITERATURE REVIEW ........................................... ERROR! BOOKMARK NOT DEFINED.
   2.1 SOCIAL LEARNING NETWORKS ...................... ERROR! BOOKMARK NOT DEFINED.
   2.2 COLLABORATION ANALYTICS ..................... ERROR! BOOKMARK NOT DEFINED.
   2.3 RECOMMENDATION TYPES ......................... ERROR! BOOKMARK NOT DEFINED.

3 EXPLORATORY DATA ANALYSIS ......................................................................................... 13
   3.1 UNDERSTANDING THE NETWORK ..................... ERROR! BOOKMARK NOT DEFINED.
   3.2 THE DATASET .............................................. ERROR! BOOKMARK NOT DEFINED.
   3.3 PREPROCESSING .................................................... 7
   3.4 IMPLEMENTATION ..................................................... ERROR! BOOKMARK NOT DEFINED.
      3.4.1 Correlation Matrix ........................................ Error! Bookmark not defined.
      3.4.2 Weka results .............................................. Error! Bookmark not defined.
      3.4.3 Power BI visualizations ................................. Error! Bookmark not defined.

4 DATA MULTICONNECTION ............................... ERROR! BOOKMARK NOT DEFINED.
   4.1 DATASET AND PRE-PROCESSING ....................... ERROR! BOOKMARK NOT DEFINED.
   4.2 IMPLEMENTATION .................................................. ERROR! BOOKMARK NOT DEFINED.

5 CONCLUSIONS ............................................. ERROR! BOOKMARK NOT DEFINED.

BIBLIOGRAPHY ....................................................................................................................... 35

APPENDIX ............................................................................................................................... 39
1 Introduction

Collaborative analytics occur when internet users involve in any kind of participatory activities. According to Web Analytics Association, web analytics refers to all the methods, tools and practices that aim to capture, measure, analyze and report data that are generated online and aim to enhance and understand user’s online behavior [1]. Collaborative analytics it is an emerging trend for over a decade in the enterprise world where business needs procured new processes where collaboration between and within departments, human force, and business units is essential, while technological advances in the form of online coworking spaces, platforms to share content and new work-monitoring techniques enabled gathering and processing a large amount of data regarding performance, productivity and efficiency of humans and operations. At the present, collaborative analytics support internal business intelligence and DevOps departments and enable data-based decision-making regarding human resources and quality management. Yet, collaborative technology and practices are thought to be a catalyst to the solutions of current business challenges which encompass disciplined teamwork and collaborative problem-solving. However, the business world is not the only one that highly assesses the value of collaborative analytics.

In the field of education, coworking and co-creation have long been existing along with data gathering and learner's assessment. Nowadays as education shifts towards computerized environments, gathering, interpretation and visualization of data become important [2]. The emergence of platforms that provide massive online courses, available to individuals, teams or co-workers, forced analytical approaches in the field of online learning since the need for insights, personalized services and innovation is attached to the upgrowth of these platforms. The term learning analytics has been used to describe metrics and indexes that reveal when to what and how online learners interact with the context all well as with each other when plenty of assignments require collaborative work. According to Unesco “learning analytics uses intelligent online data and promises to transform educational research into a data-driven science, and educational institutes into institutions into organizations that make evidence-based decisions” [3].
Thus, as information is apparent everywhere across the internet users are not confined to any specific site or online institution to shape their learning path and seek knowledge out of official spaces and ordinary ways. Discussion forums, wikis, Q&A platforms, project-based learning sites form a versatile field that is yet to be fully explored and analyzed. Many of those sites have a network structure, users have an identifiable profile and are connecting with each other through various actions. When such a network has its purpose in the exchange of knowledge and participatory knowledge forming, it can be called a social learning network, despite the fact that the major components that identify a social learning network as it has been defined: learners, instructors, and learning modules, may be indiscernible and constantly changing, yet present [4].

In this shifting environment, analytics plays a huge role in revealing user’s behavior, incentives, and needs but also in identifying trends and patterns that go beyond user’s compartmentalization and many times can provide domain-specific information. Taking into consideration that the majority of social learning networks are becoming more popular, are open, free to use and their data are publicly shared, their analytics are of great interest. The interest does not only lie with what data analysis can reveal but also lies with the analytics’ outcome and how they can serve social learning networks users. Whether that analytics are structured and can be connected to other sources of data or return to users as feedback regarding their actions. Feedback under the form of recommendation, which is another emerging trend in the educational field, that has already gained ground among online learners, is a possible outcome as well.

In the present thesis, two case studies try to address those two matters, exploration of collaboration in an unofficial Q&A social learning network and a proposal regarding the semantical connection of data sources towards a recommendation schema. The chosen network is StackOverflow, a network that attracts computer programmers and has its base in a Q&A mechanism. The connected sources are ESCO the European jobs classification system and EDX, the popular free e-learning platform.

In the second chapter, the literature review will cover how researchers have address problems like data gathering, classification, consolidation, and interpretation of learning analytics. An important note here is that in the literature the term collaborative analytics is
not so far covering forms of analytics outside the business context. Instead, term learning analytics is widely used to describe all forms of data that reflect cooperation aimed at learning. Moreover, forms of recommendations in both official and unofficial spaces and means of learning are also examined.

In the third chapter, an exploratory data analysis is performed with data from StackOverflow with the purpose of understanding and mining information that is meaningful and can be used to classify users. In the fourth chapter a case study is presented where data from three different sources are combined and linked to visualize a recommendation mechanism. In the final chapter conclusions and future challenges are listed and discussed.
2 Literature Review

This chapter is focused around the area of collaborative social learning networks and recommendation within or out of these networks. The concept of collaboration is depicted through various open-source platforms on the internet where users coexist, cooperate and/or co-create content for learning purposes. Moreover, the aim is to examine how the knowledge gained through collaborative analytics is be returned to the users and under what kind of form. The review is divided into three categories: social learning networks, collaborative analytics, and recommendation types.

2.1 Social learning networks

Collaborative technologies featured in the current day social web offer a snapshot vision of the next generation of learning opportunities. Environments such as MOOCs, Edx, GitHub offer a wide range of formal and informal learning opportunities to individuals and groups worldwide. These social web technologies hold the potential to increase opportunities for individuals to learn by cooperating and advance their intellectual and social development. According to [5] those platforms can be considered as social learning networks as they consist of members and modules that communicate with each other and can be depicted through a graph diagram.

A social learning network has its goal in providing the technological features that make online education in general, and collaborative learning in particular, effective and appealing. The rise of those networks is synced with the existence of new educational trends e.g. lifelong learning that emerged for accommodating complex and multivariate learners needs.

Social learning is thought to be a catalyst in this situation as it promotes knowledge exchange and collective thinking as well. Alterations are referring mainly to the technological disruption of both learning environments (academic or not) and learning as a
perception. Manifestation’s of this are distance learning practices, massive open online courses platforms, learning forums, project-based learning platforms. That has brought also a change in the way learning is experienced online, including practices as collaboration, gamification, online mentoring, lifelong learning [6]. Web 2.0 technology set the framework for social learning networks to flourish by enabling blogs and forums where content is generated and shared, providing cross service platforms where content can be multi-edited [7]. In the last decade, collaborative learning has attracted much attention from research teams as it is considered an ameliorative practice in online learning environments. MOOCs platform has launched a project, financed by the European Commission, called colMoocs in which "conversational agents" are introduced as an extra software feature in an attempt "to trigger learners' discussions", to serve student's intra-communication, as well as the ability for further knowledge management and progress self-monitoring. According to authors agents' impact would ideally be equivalent to another leaner opinion on the learning subject or material. It is highlighted that the project's aspirations are the greater development and usage of the tool as an innovation driver for another computer-aided learning system. Moreover, the additional scope of the project is to cater for the continuing educational needs of European citizens and facilitate timely skills acquisition that can improve professional future advances [8]. Authors in [9] describe a practice within e-learning systems that use forums spaces or emails for intercommunication between participants of online courses. SCHOm system proposes the integration of an instant exchange messaging application in e-learning systems, taking into consideration the unprecedented level of technological advances that affect our way of living. The outcome is expected to positively affect collaborative learning activities. The rise of mobile connectivity is identified as a major player in the process of shaping more innovative and collaborative ways of learning [6]. Worth noted platforms that already provide collaboration space for their members are ELLI - Effective Lifelong Learning Inventory and EnquiryBlogger | Learning Emergence [6].

2.2 Collaborative Analytics

Authors argue that social learning analytics have a significant role to play in shifting the environment of computer-aided learning by providing insights deriving from non-conventional forms of learning interactions such as forum discussions, messaging, tagging,
searching and rating across collaboration platforms [6]. Under different appellation, authors also state that collaboration analytics is a determinant factor for any attempts of improvements in computers supported collaborative learning. Either with the form of conversation dialog analysis or analysis of various forms of user traces in the system [10]. Content analysis is a subcategory of social learning analytic which can play an important role in social learning as the content that users create online within a learning context can be available for commendation, assessment or editing.

Authors in [6] identify five major factors that have affected the immersion of social learning analytics. Those are technological advances, open source, and free content initiatives, socioeconomic changes, demand for innovation and alterations in the essence of learning itself. Firstly, the expansion of internet connectivity, network speed, and access and the rise of social networks has left us with virtual common spaces where communication and collaboration between users can flourish and take many various forms ranging from status updates to projects creation. An important aspect has to do with the evolution of knowledge accreditation progress. Online courses certificates and web-published portfolios or projects in popular domain-specific communities seem to gain ground on institutional titles, previously seen as a mandatory asset for asserting knowledge efficiency. The third factor is referring to the increased demand for free content within the internet and also commenting in initiatives such as Open Education Resource (OER) which is becoming active in publishing high-quality learning content online for free. Another trend mentioned is the range of services that are nowadays available for online learners (analytics reports, personalized advice, benefits) and how the concept of freely shared content apply to them.

Authors in [11] outline the prospects, for different e-learning stakeholders, of the deployment, aggregation, and combination of both learning and social analytics. It is stated that this practice can lead to more adaptive, interactive and functional social learning environments where datasets of information will be well connected and take the form of collective knowledge in order to serve participants needs. Studies in this direction are already present.

Researchers in [12] make use of an existing social network (Facebook) as a base to introduce a twofold learning analytics system that will in one part perform discourse
analysis regarding users conversations upon relevant topics and analysis of learning activities. The aim of this system is to facilitate collaborative learning activities through the provision of insights in a user-friendly format.

As mentioned, apart from monitoring interactions of users in social networks, analytics can be produced through other types of user interaction within an existing system. Regarding the form that collaborative analytics can take, studies found focus around two main pillars: visualized reports and recommendations, although the distinction is not always clear.

Efforts of acquiring and visualize analytics from platforms where users coexist have already taken place. An example of leveraging collaborative analytics is Uatu. Researchers designed and implemented in a real environment a visualization system that monitors online editing of a google doc and generates visual reports about individual’s contribution, overall and personal progress. Those reports can be segmented by user status. eg. editor, supervisor. The system was tested in real time environment by eight computer science students collaborating in software documentation writing. Findings highlight it's collaborative activities context of collaborative writing and editing [13]. Tagging is also examined as a collaborative learning practice. Authors claim that it can be a semantic extension to the analytics part that can facilitate both learners communication and knowledge inquiry as well [14]. Worth mentioned practice is also collaborative visualization, which takes place when “visual representations of data are shared with more than one person with the Common goal of contribution to joint information processing activities in learning environments [15].

Despite the fact that participation in online learning conversations can improve overall learners’ performance, challenges that have to do with data privacy and analytics measurements manipulation are still present. Moreover, challenges lie with social learning analytics, focusing mainly on their quality, coherence, distinctiveness, and means of measurement. Additionally, analysis expands to the unique character or social learning environments and how this affects the shaping of analytics [6]. While big data technologies can be also utilized to handle the massive amount of data that derives from learners interactions and extract useful information [16], only in the recent years' technologies like AI and Big Data has been introduced to harness existing data. Hence many of those
systems are built with technologies that provide no extension or upgrade options making the initial collection of information harder [10]. The overall conclusion is that the abundance of this type of analytics can lead to smarter, more adjustable and accurate recommendations notwithstanding the privacy concerns that accompany any handling of personal data.

2.3 Recommendation Types

A challenging area in the field of e-learning is recommendations. In a broader sense, in the literature the topic of recommendation is delineated by two different concepts: Intelligent Tutor Systems (ITS) and Recommendation Systems (RS). Both concepts have emerged during the last two decades and were initially deployed for commercial use or for business needs [17]. However, due to the rise of computer-supported collaborative learning, they have found a place in the emergent learning communities where the need for more adaptive and smart systems is necessary.

As an Intelligent Tutoring System we can describe a computer software that can assist the acquisition, by providing timely advice and feedback according to individual’s needs. Such a system encapsulates some form of intelligence along with automated processes and in its best practice, substitutes a human tutor [18] Within ITS, feedback can take different forms and adapt to learners' interactions while working on a group. Moreover, it can offer immediate advice to learners by keeping track of their interplay as this is depicted through various online activities and procure intelligent and flexible solutions for more personalized recommendations. Improvement of those systems and their integration to computer-supported learning communities can lead to high-quality learning experiences both for groups and individuals [10]. Researchers who benchmarked three different ways of feedback inside a Collaborative Intelligent Tutoring System and specifically to a pair of students working remotely on a programming language course found that when domain feedback was accompanied by collaboration feedback learner's performance was significantly improved [19].

At the e-learning context, RS can be defined as powerful computing software that filters the vast amount of fragmented information and has the ability to decide whether any
combination of this information is recommendable for a specific user. This outcome can be achieved through various technics like collaborative filtering, content-based filtering or a hybrid version. Each one of them utilizes many different data resources like user’s characteristics, learning progress, generated content etc. [20]. Nonetheless, that depends on the available sources of information and the dynamics that each system wished to render to its users.

A recent examination of the specific characteristics that best fits in a collaborative learning platform, led to the conclusion that high personalization on feedback and the provision of great user experience in the interaction between user and system play an important role in learners’ satisfaction. Consequently, a system must take advance of the learners online social activity and illustrate their progress compared to the groups they belong to [21]. In [22] authors propose a cloud-based recommendation system that will semantically encapsulate data from social networks regarding knowledge or other assessment-related activity and provide recommendations within an e-learning system under various forms. The model is based on the concept of ontology and artificial immune systems. It is also mentioned that such a system can be used for a job-profiling recommendation as well. Additionally, the creators of another recommendation engine, which can be integrated into social learning environments, propose the utilization of various social analytics metrics that range from friendship index to topic segmented user ratings. Authors consider this approach to be vital for the specificity and uniqueness of the recommendations [23]. Rating practices are listed as a decisive factor in the design process of a recommendation system in an e-learning environment. User rating can be perceived as the “good learner rating” but also take other forms. A comparative study suggests that this type or RS can improve learner’s performance [24].

Pertaining to the facilitation of online collaboration and having as a base computer-supported collaborative learning and mobile supported collaborative learning theories, authors propose a mobile application that will be able to provide suggestions about the most suitable learning partners within a platform with the purpose of creating "communities of practice", namely online groups whose members share similar level of knowledge upon a specific topic and the willingness to cooperate or discuss upon relevant subjects. This application feature can be integrated into existing online social networks also. Data from a social network, collaborative filtering, and social knowledge mining techniques are to be used to funnel the recommendation algorithm. Their purpose is to leverage existing data from multiple available sources including user's personal library preferences, when
exist, in order to suggest more coherent thus diverse groups of learners. Lastly, evaluation metrics regarding the quality of service and users satisfaction showed a positive impact on the cognitive level of learners as well as in the socialization changes of them [7]

Based on the belief that clear explanation of how recommendation arises, can enhance recommendation engines impact researchers propose a tool that aims to explicit present how recommendations about forthcoming learning actions are formed and provided by the tutor, issuing a system which implements data from learners’ collaborative activities and present outputs in a visual way. Specifically, the system presents a top-down approach decision tree to break down the path that leads to the specific recommendable outcome. By providing simplicity and accuracy authors state that feedback becomes more compelling for the learner and thus more likely for them to accept it. Additionally, provides structured insights into the collaboration process that learners can use variously [25].

To conclude, there is evidence that online learners are more prone to excel in group learning as soon as they were getting feedback about the advantages, they could acquire [26]. However, an important aspect that can affect the usefulness and effectiveness of recommendation engines is choosing the most suitable type of information. The lavishness of having an extreme volume of user data available can lead system's developers to choose unnecessary features and produce more complex and unfunctional systems.

Based on the upon literature review research the questions this thesis will try to address are the following two research questions.

R. question 1
What kind of knowledge can we mine from collaboration in social learning networks?

R. question 2
Can this knowledge be utilized to further expand and improve the learning process or be of any other use for the end user?
3 Case Study design

Nowadays students and professionals take advantage of new forms of online knowledge to leverage their skills and build their career. In the field of technology, most of the online courses include practical exercises, real-time deployment of features, code production and a lot of personal involvement in projects in order for learners to ground the taught subjects. In a parallel manner, online learners, professionals, and students seek information regarding their projects and share their work or their experience to help each other learn better and accomplish their goals. This process has led to the creation of huge collective knowledge repositories. As users interact and get involved in forums, wikis, sites, and platforms produce analytics that can provide insights regarding what kind of value do those interactions disclose, how the offered help is perceived and assessed from the users and if their engagement helps the communities grow and meliorate the created content. We believe that this commonly created content can serve as a significant complementary learning companion and contribute to the efforts of people trying to grasp the knowledge both with scientific and empirical manner. In the following pages an exploratory analysis of the StackOverflow Q&A network will be presented. The specific network stands out for its popularity across millions of users that engage with technology in many different levels.

3.1 Understanding the network

Created in 2008 by Jeff Atwood and Joel Spolsky, Stack Overflow is part of the Stack Exchange Network and is one the largest and most trusted online communities where developers and other technology-related specialisms participate to seek information, learn, share their knowledge and find new job opportunities. The site is using a Q&A model that facilitates and promotes collaborative editing for providing the best answer to each question. This means that many users can answer the same question by adding new solutions, highlight any mistakes in other answers or edit previous answers. This form of collaboration that leads to higher content quality and adds value to the community work. Also found that developers even use the platform to spot and document existing bugs and introduce new features on their work.
According to site’s insight report for 2018 [27], more than 50 million professional and aspiring programmers visit Stack Overflow each month to help solve coding problems and develop new skills, with 21 million of these estimated to be professional developers and university-level students. The level of experience presents a wide range among developers, and a full third of professional developers on Stack Overflow learned to code within the past five years. Additionally, merely 25% of the respondents were enrolled in a formal college or university program full-time or part-time. However, 90% of the users answered that they have taught themselves a new language, framework, or tool outside of their formal education. Among professional developers, almost half said that they had taken an online course like a MOOC both in 2018 and 2017 research. The platform has been established in the mind of the developers as a strong learning companion. Over 80% of respondents rely on Stack Overflow Q&A when learning something new, which makes it along with the official documentation, the most common way developers level up their skills.

Forms of participation in the platform also differ. Based on the insight developers can come to StackOverflow only to find answers to their questions, while others participate in the community by asking, answering, voting for, or commenting on questions. Specifically, a 40% of survey respondents participated in Stack Overflow a few times per month or more often. One of the most important findings is that the sense of community is very strong among site users. Characteristically the biggest part of the respondent’s answers that they “consider themselves part of the community”.

High participation of both students and professionals in the platform, the level of estimation that it has among its users, the uniqueness of the idea of sharing and dynamically creating collective knowledge, the depiction of the technological advances and the linkage with the labor market changes make StackOverflow good candidate for this case study. However the most important characteristic regarding collaboration analytics, lies in the Q&A mechanism of the site, where a huge amount of interactions take place, generating data that this use case will try to leverage. Analytically user can post a question that can be specific or not at a certain degree and use up to five tags to describe it and increase its visibility into the specific threads. Following, other users can upvote or downvote you question based on its relativity and the compliance with site's terms of use. Users that are willing to share their expertise regarding the referring subject can comment and
offer their answers. The user that posted the question can accept or reject the answer, and upvote or downvote the answer. Every user can vote for the answers or even edit them. Every time an upvote action takes place the user gets points according to the following matrix:

<table>
<thead>
<tr>
<th>Action</th>
<th>Score Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>User’s answer accepted</td>
<td>+15</td>
</tr>
<tr>
<td>User’s question voted up</td>
<td>+5</td>
</tr>
<tr>
<td>User’s answer voted up</td>
<td>+10</td>
</tr>
<tr>
<td>User’s editing voted up</td>
<td>+2</td>
</tr>
</tbody>
</table>

Table 1

There is also negative scoring for the equivalent downvote actions

<table>
<thead>
<tr>
<th>Action</th>
<th>Score Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>User’s answer voted down</td>
<td>-2</td>
</tr>
<tr>
<td>User votes down answer</td>
<td>-1</td>
</tr>
</tbody>
</table>

Table 2

As we see the downvoting of the answers has a negative impact on the voter too. This is considered an act of protection for the community as it mobilizes users towards more active practices such as editing the answer or flag it as irrelevant. Additionally, there are some extra features preventing users from gaining or losing an unworthy reputation. Those are "Community user" and the "Protect Question". Community user according to site documentation is an automated process that owns edits by anonymous users, all community posts and the downvotes of spam threads. A question is protected when no one can add post an answer. Users that have gained the privilege to flag a question as protected do so to protect mainly new users from adding answers that can negatively affect their reputation or result to a spamming.

The assuming notion is that user’s reputation, that is formed by the accumulation of points of each post or answer, is a well calculated and trusted procedure that combine quantitative as well as qualitative metrics to measure participation and collaboration. Higher
reputation score is rewarded with more available actions in the platform and higher awareness inside the community. Worth mentioned is that apart from high reputation users can earn “badges” that differ between answers, question and overall participation and are listed in their profile. For example, someone who answers a question with score of 1 or more can be listed as "teacher" and someone who answers its own questions with a score of 3 or more is considered a “self-learner”.

3.2 The dataset

Stack Exchange which is the parent site of StackOverflow platform publishes almost every three months a year the updated site’s data bump in the archive.org platform. The dataset is published under Attribution-Share Alike 3.0 Creative Commons License and it can also be found in the Google’s Big Query public datasets. Author's will was to utilize the whole dataset for a complete analysis. However, limited computational power, lack of free tools and a large amount of data that unfolds in millions of rows led to the conclusion that Big Data related methods of analysis would fit better for that approach.

The alternative was to access StackOverflow data through the Stack Exchange Data Explorer (SEDE) where data are stored in a relational database schema, therefore, can be queried and downloaded in CSV format. Though there exist a limitation of 50,000 instances. Worth mentioned is that except SEDE queries, there is also an API provided with predefined queries in a quite user-friendly interface that does not require knowledge of a query language. Nevertheless, result limit is only 30 instances per day and is provided in JSON format, making it useful only for specific information retrieval.

The datasets used in the following analysis were retrieved through SEDE with SQL queries. Before presenting and detailing the datasets used it is important to depict how site information is stored in the StackOverflow database. As mentioned above the whole data are stored in a relational database schema which uses tables that are connected with each other over columns that store unique values for every table row. Usually, this refers to an ID column and it is called a primary key. A table X is connected with a table Y when X's primary key is featured in Y. For Y this column serves as a link to the table X and it is called foreign key.
The current database schema comprises of four main tables: Posts, Users, Comments, Votes and twenty-three tables that hold complementary information of the main four. Posts contain all the questions that users do while Comments holds all the answers users give. Votes hold the votes of all the posts questions and Users store data like id, name, reputation, date of induction and total upvotes for all the users in the platform. For the further understanding of the database structure a visual representation of the tables and the columns that were used is provided below. Other tables were excluded due to space limitations.

As we wanted to analyze users’ behavior, user's level of participation in knowledge forming and user's knowledge validation User's id, name, reputation, and upvotes were collected. Then all the comments that users did where retrieved from Comments with id, score, and username. In order to drill more into the dataset, the existence of tags was also important so all the tags of the questions of the answers the users did where fetched from post table. As the limitation of the API is 50,000 instances, we descended the result by comment' score since it is the base measure for the analysis and we wanted to avoid low or even negative values. The code for this query can be found in the appendix section.

The outcome extracted in csv had the following format
3.3 Preprocessing

Data preprocessing was required for the reduction of the dataset in order to understand better the dataset and reduce the number of instances. Excel application was used for that. The first step was to identify how many unique users were there. With Remove Duplicate function 27,785 users were identified. Next, in order to understand how many answers each user contributed, we pivoted with Power Pivot function Answers Id and Users Id columns and then aggregated them to get answers per user column. However, there had to be a measure for the quality of the aggregated answers. For that reason, a column with the average score value for all the grouped answers was created through Grouping By function within Excel’s Power Query. Tags column was removed in this phase of the analysis.

Table 4

<table>
<thead>
<tr>
<th>User Id</th>
<th>Count of Answers</th>
<th>Average score of answers</th>
<th>Users Reputation</th>
<th>Users Upvote</th>
</tr>
</thead>
</table>

3.4 Implementation

For the analysis and the presentation of the findings, three tools were used. Weka, MS Excel, and Power BI. Weka is an open source data mining tool created by the University of Waikato of Australia that runs in java environment. Excel is a Microsoft 2016 version is a Microsoft proprietary software for data analysis. Power BI is an open source software created by Microsoft for data visualization. November 2018 version was used.
3.4.1 Correlation Matrix

After the preprocessing steps, a correlation analysis was performed to reveal any correlation between the created measures (Count of answers, Average score of Answers) and the raw data (Users Reputation, Users UpVotes) and help us understand further the attributes. As shown in Picture 2 there exists a strong positive linear relationship between Count of Answers and Users Reputation and a moderate positive linear relationship between User's UpVotes, a way of evaluating users’ participation in the platform, and Count of Answers. Moreover, a moderate positive linear relation exists between Users UpVotes and Users Reputation. Regarding the interpretation of the results, an increase of Users Reputation coexists with an increase in the Number of Answers and Users UpVotes for each user and an increase in the Users Upvotes coexists with an increase of the Answers as well. Both results were expected, taking into consideration that as users engage more in the platform, reading questions, asses them and give answers they have good chances to raise their Reputation. We can also see that there is a very low negative correlation between Average score of answers, User’s Reputation, and User's UpVotes. This is due to the average value selection which normalized the results so that the range of the score of the answer for each user is not well represented. To our best knowledge, the mean of answers score could have given better results.

![Correlation matrix](image)

The above correlation matrix was created with Excels Data Analysis tool pack.

3.4.2 Weka results

In Weka, cluster method was chosen for its ability to reveal any representative patterns in the data, which could characterize the users, in our case. Cluster analysis was performed with four attributes: User’s Reputation, Count of answers, Average score of Answers and Users UpVotes. The reason all of the attributes were selected for clustering is that they describe user's performance both in a quantitative and qualitative way. Answers
Count and Average score give the quantity and the quality of the given answers, while Users Reputation and Users UpVotes provide insight regarding users’ behavior. 27784 instances were clustered using the EM algorithm and percentage split (split on 66% of the dataset and 13 iterations performed). This method gave three clusters with the following characteristics: the first cluster contains 889 instances or 9% of the dataset with mean of Users Reputation 77836, mean of Average Answer’s score 38.3, mean of Count of Answer 6.3 and Mean of User’s UpVotes 3825. The second cluster contains 3547 instances with Mean of Reputation 11063, mean of Average Answer's score 55.3, mean of Count of Answer 1.7 and mean of User’s UpVotes 1014. Third cluster and the biggest one as it contains the 53% of instances has mean of Users Reputation 2359, mean of Average Answer’s score 31.8, mean of Count of Answer 1 and mean of User’s UpVotes 31.8.

As we see the mean of Average score of Answers does not have any significant variation between clusters. However, we see that user with higher Reputation that have contributed more answers and have upvoted more posts to form a separate cluster. Instances that belong in this cluster represent 9% of the dataset.

Apart from EM algorithm Simple K-means and FarthestFirst algorithms where also used though results were not useful for analysis because they both clustered the majority of instances, 93% and 98% corresponded in one cluster.

### 3.4.3 Power Bi Visualizations

While Weka provides a variety of clustering algorithms and preprocessing filters. However, visualizations are poor, so another software was used. Power BI provides Clustering Visuals using R programming language using K-Means algorithm. However, these visuals are more interpretable when two attributes are used for clustering. Attributes where also scaled for the same reason. In Picture 3 instances where clustered based on Users Reputation and Count of Answers. Three clusters were formed and as we can notice below the existence of some outliers were detected. Outliers in our case indicate that in our dataset exist a user that has to gather a relatively bigger reputation and has given more answers.
Next, Answers Score and Number of answers were used for clustering. As shown in Picture 3 clusters are formed across the axes and the detected outliers are either users with a high number of answers or users that have a high average score of answers.
Additionally, instances were clustered with Answers Average Score and Users Reputation. Values were scaled.

Picture 5

Picture 6
The analysis of how Users reputations index is formed in the platform, the strong correlation that exists among reputation and number of answers for each user and the clustering results gives ground to propose that those two indexes could be used to categorize users and serve as a primordial analytics measure.

Based on that, in the following Picture, the whole dataset was induced into Power BI under the following form, sorted by Users Reputation.

![Picture 7]

![Picture 8]
A directed graph visual was used to depict how many users in our dataset answer questions with specific tags, what is their reputation and how many answers gave for those specific tags. Tags were used as sources and users Id where used targets. Links were weighted by Count of Answers and Users were also filtered by their Reputation Score. Specifically, only users with a reputation equal or higher than 10,000 are depicted. The threshold is indicative. In Picture 9 tags where filtered by the word Python and JavaScript as shown below.

Picture 9
As we see in Picture 11 and 12, under the hashtag Python user with ID 100297 has given the most answers among users whose reputation is higher or equal than 10.000.

Under the hashtag JavaScript, three users have given the majority of answers among users with a reputation higher or equal to 10.000.
4 Data Multiconnection

In this chapter, a possible application of collaborative analytics, extracted from unofficial social learning networks and spaces will be analyzed. A combination of different datasets will be relationally connected, and Power BI software will be used to visualize how data from multiple sources can be mapped together and lead to useful recommendations for the end users. The data from StackOverflow, European Jobs Classification framework (ESCO) and EDX learning platform were chosen as the data sources. This case study highlights the importance of using and combining data that exist online and can be semantically connected in order to provide to the end user information that helps him or her take action towards a direction. In the following case study, the end user is someone who tries to qualify for a job position by acquiring a new skill, as shown below:
4.1 ESCO and EDX

EDX, an initiative started by MIT and Harvard University in 2012, is one of the most known e-learning platform and MOOC provider through the open source project, Open EDX Portal. Online users can have access in more than 1500 courses provided by 340+ universities, institutes and organizations across the world. All courses are free of charge, but users can issue for an official certificate under a certain price. EDX course catalog is available online.

ESCO is a European initiative that adheres to the concept of Semantic Web and specifically towards Labour Market Interoperability. Semantic Web promotes the meaningful linkage of data sources so that the whole amount of data, existing online could be processed by machines instead of humans. The functionality of Semantic Web and the applications built within it require data storage in machine-readable format, open data sources, and interoperable computing systems among others. The purpose of web semantics is the harnessing of existing data and the transformation of world wide web to a globally linked database where the stream of information would be facilitated by machines [28]. ESCO which stands for European Skills Competencies Qualifications and Occupations is a classification system that serves as a "common reference terminology". ESCO can be seen as a directory that includes requirements, necessary skills or good to have skills for various jobs in all European languages. The structure of this information that enables linkage of other data sources is a step towards the integration of European labor market and the gradual interconnection of employers, employees, human resources managers, students, trainees, and other stakeholders. ESCO data are under the form of Open Linked Data and their utilization from developers is highly supported. According to a well-connected labor market, where the connection of multiple data sources will transfuse different utility aspects, could serve as a solution to the disruption that will be caused due to technological advances [28].
4.2 Dataset and Preprocessing

Data used in this case study where retrieved through APIs. ESCO additionally provides all data in CSV format but it has also a well-integrated API and a real-time Local API will soon be available. EDX as well has an API from which one can retrieve the data. Power BI allows users to connect data directly from web sources thus no other tool was used for preprocessing. From ESCO API occupation under the title Web Programming where retrieved in the following format.

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>classId</td>
<td><a href="http://data.europa.eu/esco/model#Skill">http://data.europa.eu/esco/model#Skill</a></td>
</tr>
<tr>
<td>uri</td>
<td><a href="http://data.europa.eu/esco/skill/69bbd53f-fbb0-4476-b4b2-ef7844464e28">http://data.europa.eu/esco/skill/69bbd53f-fbb0-4476-b4b2-ef7844464e28</a></td>
</tr>
<tr>
<td>title</td>
<td>web programming</td>
</tr>
<tr>
<td>Reference language</td>
<td>List</td>
</tr>
<tr>
<td>Preferred label</td>
<td>Record</td>
</tr>
<tr>
<td>Alternative label</td>
<td>Record</td>
</tr>
<tr>
<td>Description</td>
<td>Record</td>
</tr>
<tr>
<td>status</td>
<td>released</td>
</tr>
<tr>
<td>links</td>
<td>Record</td>
</tr>
<tr>
<td>embedded</td>
<td>Record</td>
</tr>
</tbody>
</table>

With Navigation and Concert to tables, function list was expanded and converted to table so that the form after those steps were the following

<table>
<thead>
<tr>
<th>href</th>
<th>URI</th>
<th>title</th>
</tr>
</thead>
</table>
Then title text was trimmed and filtered by the first word of the text range was chosen to create a new column. The new column which was named Text Range was filtered by the word “web” so that our data were in the final form.

Table 7

<table>
<thead>
<tr>
<th>href</th>
<th>URI</th>
<th>title</th>
<th>Text range</th>
</tr>
</thead>
</table>

Filtering gave us all the ESCO occupancies that contained the word “web”.

EDX dataset had the initial form

Table 8

<table>
<thead>
<tr>
<th>channel</th>
<th>Attribute selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>table</td>
<td>2.0</td>
</tr>
</tbody>
</table>

The embed table in column link was expanded to the following form

Table 9

<table>
<thead>
<tr>
<th>Title</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edx.org course feed</td>
<td><a href="https://www.edx.org/api/v2/report/course-feed/rss">https://www.edx.org/api/v2/report/course-feed/rss</a></td>
</tr>
</tbody>
</table>

The embed table in column item was expanded and gave the following form
Table 10

<table>
<thead>
<tr>
<th>guide</th>
<th>title</th>
<th>link</th>
<th>description</th>
<th>pubDate</th>
<th>source URL</th>
</tr>
</thead>
</table>

The table in column source URL was expanded so that more information about the course were available under the form of table 9

Table 11

<table>
<thead>
<tr>
<th>guid</th>
<th>title</th>
<th>link</th>
<th>description</th>
<th>Pub Date</th>
<th>subtitle</th>
<th>subject</th>
<th>prereq-uisites</th>
</tr>
</thead>
</table>

Column Title was filtered by the text “web” since we wanted to collect all the courses which contained “web”. However, because the subject contained more than one fields delimited by a comma, word “web” was extracted so it could serve as the link parameter.
For the Stackoverflow the initial dataset was used. For this dataset the tags related with web development were manually selected as no unique value that could be used for selection was found.

4.3 Implementation

Data where connected and presented in a dynamic dashboard. The graph below shows that when we select the “web developer” occupation form ESCO board, we get an EDX course with the title “Programming for the Web with JavaScript” and the users with the highest reputation that have answered questions that have been labeled with the specific hashtags that were relevant to web development.
EDX data

<table>
<thead>
<tr>
<th>Course Title</th>
<th>Text Between Delimiters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming for the Web</td>
<td>Web with JavaScript</td>
</tr>
</tbody>
</table>

StackOverflow data

- c++: 100.0%
- c#: 100.0%
- java: 100.0%
- javascript: 100.0%
- php: 100.0%
- git: 100.0%
- python: 100.0%
- html: 100.0%

ESCO data

<table>
<thead>
<tr>
<th>ESCO Occupancies</th>
<th>Text Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>web content manager</td>
<td>web</td>
</tr>
<tr>
<td>web developer</td>
<td>web</td>
</tr>
<tr>
<td>webmaster</td>
<td>web</td>
</tr>
</tbody>
</table>

User Info

<table>
<thead>
<tr>
<th>Tags</th>
<th>Sum of Users Reputation</th>
<th>user id</th>
</tr>
</thead>
<tbody>
<tr>
<td>c#</td>
<td>1069036 22656</td>
<td></td>
</tr>
<tr>
<td>c#</td>
<td>779290 17034</td>
<td></td>
</tr>
<tr>
<td>oop</td>
<td>622096 14860</td>
<td></td>
</tr>
<tr>
<td>git</td>
<td>570990 20862</td>
<td></td>
</tr>
<tr>
<td>java</td>
<td>538011 218196</td>
<td></td>
</tr>
<tr>
<td>java</td>
<td>528915 571407</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>69205603</td>
<td></td>
</tr>
</tbody>
</table>

Picture 12
5 Conclusions

5.1 Conclusions

In this thesis, collaborative analytics, social learning networks, and possible recommendation mechanisms were discussed. Social learning networks came across as a new ecosystem for online learners that enclose a great deal of potential for creative learning, knowledge sharing, and collaborative work editing. The conducted study indicated that the analysis, the quantitative and qualitative measurement of users' actions that take place inside these networks uphold valuable information regarding the quality of the generated content and the rating of the editors based on their interactions. There is also evidence that learners participate in many different networks in their effort to grab the knowledge from any available source. The upshot of this is the possibility for these data to be semantically connected and be utilized to produce various recommendations for the end user. Towards this approach, ESCO classification system was used as a reference source because it is an institutional effort to bring together members of the labor market by semantically linking tones of available information for all parties involved.

The limitations of this effort lie the absence of a unified framework that can serve to provide an analytical view of social learning networks. This is due to the fact that each network is very unique, and its analysis requires deep domain knowledge, active participation and user unit analysis. Analytic indexes and KPIs have to be chosen very carefully, though the extraction may be challenging as well. Moreover, since every network is designed for a very specific cause, that serves best its users, the lack of common data references is a drawback for the connection to other networks. The semantical enrichment of data seems to be a solution that will lead in leveraging those data.

Conclusively, extracting insights from networks where new ways of learning emerge, is a challenging task due to networks complexity and variation. However, the added value
coming from people’s actual collaboration is highly evaluated and appreciated so that harnessing this information can lead to high-quality outcomes for the online learners.

## 5.2 Future work

The results of this study support the idea that more research has to be conducted in the regarding the effect of online collaboration. The tremendous amount of data that users generate urge the usage of Big Data technologies along with new approaches and techniques that can shape new tools. Apart from that, domain knowledge is a fundamental issue for future research in the field of social learning networks. Enterprise world and business applications can serve as a paradigm.
Bibliography


Appendix

Code snippet 1: this code was used to extract data from the database

```sql
1 SELECT Comments.Score, Comments.Id, Comments.UserId, Comments.DisplayName, Comments.PostId, Users.Id,
2 Users.Reputation,
3 Users.UpVotes,
4 Posts.Tags
5 FROM Comments
6 RIGHT JOIN Users
7 on Users.Id = Comments.UserId
8 RIGHT JOIN Posts
9 on Comments.PostId = Posts.Id
10 ORDER BY Comments.Score DESC
```