Enterprise Expertise Characterization

Master's Thesis

Nidhi Singh
Enterprise Expertise Characterization

THESIS

submitted in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Nidhi Singh
born in Delhi, India

Web Information Systems Group
Department of Software Technology
Faculty EEMCS, TU Delft
Delft, the Netherlands
www.ewi.tudelft.nl

Center of Advanced Studies, IBM
Johan Huizingalaan 765
1066 VH Amsterdam
Netherlands
www.ibm.com/nl
Enterprise Expertise Characterization

Author: Nidhi Singh
Student id: 4242246
Email: nidhi.singh.igit@gmail.com

Abstract

All enterprises have expertise in the form of their employees, and their expertise management is crucial as it helps in locating the right person for the right job, problem solving and question answering among others. This task can easily become complex for a large organization, like IBM, which is the subject of our study. Enterprises face the challenge of sharing their employee’s expert knowledge and employ tools to register, communicate, and search employees as a knowledge resource. The majority of these tools fail to take into account the following information about the person: demographics, credibility, behaviour, reputation, and accessibility. Whereas, external social sources are rich in these user dimensions, e.g. LinkedIn (a business oriented social network). Using this as motivation, we performed a case study at IBM in which we combine external social media source and enterprise social sources to better understand their employee expertise. The study involved 211 IBM employees, whose relevant social data was collected. We use this data to analyze their expertise related fields in both the external and enterprise sources. This analysis let us conclude that addition of an external social media source adds more value to the internal expertise data of the employee. We also validate this claim with the help of an information retrieval system setup.

Thesis Committee:

Chair: Prof. Dr. Ir. G.H.J. Houben, Faculty EEMCS, TU Delft
University supervisor: Dr. Ir. Alessandro Bozzon, Faculty EEMCS, TU Delft
Company supervisor: Drs. Robert-Jan Sips, Center of Advanced Studies, IBM Netherlands
Committee Member: Dr. Martha Larson, Faculty EEMCS, TU Delft
“The possession of facts is knowledge, the use of them is wisdom.”

– Thomas Jefferson
The journey to my graduation has been a roller coaster ride with ups and downs, and would not have been possible without the support of a number of individuals.

I would like to thank my direct supervisors, Alessandro Bozzon and Robert-Jan Sips from IBM for being patient with me and help me complete this journey. Alessandro advised me with the scientific questions and Robert-Jan provided all the support needed from IBM to carry out this study.

I would also like to extend my thanks to the graduation committee members, specially Martha Larson, for taking out time to read and critique my work.

I wouldn’t have been able to achieve this without the blessings and support of my parents, who are also my greatest critiques and advisers even from far away.

Nidhi Singh
Delft, the Netherlands
November 5, 2015
CONTENTS

3.4.1 Expert Finding at IBM ........................................... 26
3.4.2 Expert Finding System: Implementation ........................ 27
3.5 Discussion ................................................................. 28

4 Expertise Characterization inside and outside IBM .......... 30
4.1 First prototype: analysis ............................................... 31
4.2 User Demographics ..................................................... 32
4.3 People Tagging Evaluation ............................................ 34
4.3.1 Vocabulary Evaluation ............................................. 35
4.3.2 User Tag Semantics ................................................ 36
4.4 People Network Evaluation ............................................ 41
4.5 Discussion ................................................................. 43

5 Improving Expert Finding with External Social Sources ..... 46
5.1 Approach ................................................................. 46
5.2 Experiment Setup ....................................................... 47
5.2.1 Document collection .............................................. 48
5.2.2 Query topics ........................................................ 48
5.2.3 Relevance judgments ............................................. 48
5.3 Performance metrics ................................................... 49
5.4 Results ................................................................. 50
5.5 Discussion ................................................................. 50

6 Conclusions and Future Work ........................................ 52
6.1 Contributions .......................................................... 52
6.2 Conclusions ............................................................ 53
6.3 Future work .............................................................. 54

Bibliography ................................................................. 55
List of Figures

3.1 Process Pipeline of the Research Framework ................................. 16
3.2 Overview of data gathering process ........................................... 19
3.3 Representation of LinkedIn Schema ........................................... 21
3.4 Representation of IBM connections schema ................................. 22
3.5 Overview of tag data preparation .............................................. 23
3.6 System Overview of IBM Expertise ............................................ 28
4.1 Tag count distribution ............................................................ 37
4.2 Distribution: difference in tag count ......................................... 38
4.3 Difference in tag count across BUs ........................................... 39
4.4 Cosine Similarity for LinkedIn and IBM Connections tags ............. 40
4.5 Cosine Similarity for LinkedIn and IBM Connections tags-categories . 41
4.6 Distribution of Friends Count in both Networks ............................ 42
4.7 Difference in Friends count distribution and across BUs ................ 44
4.8 Difference in Friends Count across dimensions ............................ 44
Chapter 1

Introduction

Enterprises ranging from small to medium to large all have expertise in the form of their employees. Their expertise management is crucial since it helps in locating the right person for the right job, problem solving and question answering among others. This task can easily become complex for a large organization, like IBM, for which this study is conducted.

Enterprises use all sort of systems to manage their expertise and majority of these systems are designed to be used by the employees to do their day-to-day expertise finding tasks but the effectiveness of these systems has always been questioned. There are several studies in which users were interviewed to understand what they ask their network and their motivation for asking those questions [1].

One of the major findings of one of these recently conducted study at IBM [2] revealed that only 1% of their participants "searched an internal tool" to solve their expert-finding needs. Most of these participants reported that they turned to their colleagues for these tasks. The study described an expert-finding system available for use within the enterprise as an internal tool. The majority of the participants also highlighted that traits of a candidate like experience (summary of education and work), group affiliations and current role, which aid in selection of a candidate expert were missing from the tools. Thus, with our work we propose that the inclusion of external social network sources e.g. LinkedIn\(^1\) which includes these lacking features, can assist in better understanding of the candidates.

1.1 Problem Statement

The broad research question investigated in this thesis is:

\(^1\)https://www.linkedin.com/ - business oriented social network
1.2 Methodology

How can the combination of social networks and enterprise social sources result in better understanding of employees’ expertise?

This question can be answered by examining following sub-research questions:

**RQ1:** What is expertise? What are the different models to define expertise? How is expertise determined in enterprise setting different from expertise derived in public Social Networks?

**RQ2:** How to assess expertise? What are the state-of-the-art methods to do this?

**RQ3:** Do the external social media sources duplicate the data already contained in enterprise sources?

**RQ4:** Which attributes of external social media sources can help in improving expertise assessment?

**RQ5:** To which extent can the combination of the enterprise and external data sources improve the expertise assessment in expert finding system?

In order to answer these questions, enterprise and public social media sources relevant to the questions, were required to be identified. Furthermore, a framework to facilitate the collection of data from these sources needed to be implemented.

1.2 Methodology

This research study provides a thorough examination of prior work done in this field, discussed in Chapter 2. The main goal here is to understand the field of expert finding systems and how expertise is determined in enterprise environment and public social media. This knowledge is means to answering **RQ1** and **RQ2**. In further text, we will refer to both private and public media sources in enterprise as Enterprise Social Media (ESM) and external social networks as Public Social Media (PSM).

This study is extensively dependent on user data, therefore, a framework is developed for data collection. This task involved studying enterprise social media and public social media (Linkedin and Facebook), and building crawlers to extract user data from these sources. The technical details of the social media sources and the developed framework are discussed in Chapter 3. Data collected through crawlers is analyzed for understanding the user demographics and their profiles on separate social media. The goal is to identify the relevance of these sources in locating expertise of an employee. Hence, aids in answering **RQ3** and **RQ4**.

Furthermore, we develop an expert finding system augmented with data extracted from public social media sources, to perform qualitative analysis. These
qualitative measures aid in system evaluation and acts as means to answer our **RQ5**. The system and experiment details are described in Chapter 5.

### 1.3 Contribution

**Framework** The first action point and the major contribution of this work is the development of a framework to carry out this work. The framework consists of three main components:

- **Data Gathering Component**: This component outlines the process to pick which data sources will be required to facilitate this study, development of data schema and crawler scripts to timely collect and store this data whenever available to the system.

- **Game and participants invitation**: The access to user data, related permissions were obtained through a sister project. A game [3] was developed as part of this project, to measure user engagement in enterprise environment. This tool also allowed users to invite their peers. Further, in this text, we will refer to this project as the **game tool**. To support this study, we developed an authentication system which allowed users to connect to their respective public and enterprise social accounts and enabled them to use this connection in other aspects of the game tool.

- **Analysis**: Lastly, analysis component was developed to explore and study this collected data set across different hypothesis. Moreover, to facilitate this analysis a data preparation module was developed, which cleansed the data for analysis tasks.

Chapter 3 describes the above introduced components in detail.

**Data set: Participants’ Enterprise and Social Media Data** Usable information to conduct a data intensive user study is a scarce resource. With this work, we curate a data set from social media sources within the enterprise and outside. The tool developed to gather this data, asked for appropriate access to user data. A complete description of data set schema is provided in Chapter 3. This data set contains rich information about users and their social networks. This information is utilized in our work to analyze and validate an expert finding system, but can also be applied further for characterizing other dimensions of user’s profile, like identifying his social media behaviour patterns etc.
**Study of Participants’ Expertise on Public and Enterprise Social Media**  Third contribution of this work is the analysis carried out on the gathered data. The aim of this analysis is to emphasize the added value of an external social media source to the enterprise social media and expert finding system. Through this analysis we were able to prove that an external social source data enriches the expert finding related tasks on a certain number of dimensions, these dimensions are introduced in Chapter 2 and discussed with respect to data set in Chapter 4.

**Validation of Social Media Enriched Expert Finding System through an Experiment**  The last contribution of this work is the validation of the new data enriched expert finding system with expert finding system using sources used in current implementation at IBM.

### 1.4 Thesis Outline

In Chapter 2 we will review related work on expertise and expert finding systems, we will discuss in detail the research design and framework our work in Chapter 3. More information about the data set and its properties will be discussed in Chapter 4, along with Social-Tagging evaluation of both enterprise and public social media sources. Chapter 5 will present evaluation of expert finding system encompassing both enterprise and public social media, with experiments, its setup and results. Finally, the conclusions and future recommendations will be discussed in Chapter 6.
Chapter 2

Background and Related Work

Expert finding is a well explored research field in computer science. This chapter introduces this field and discusses prior work done in research and development of Expert finding systems.

The need to find answers to our queries forces us to search for resources in the form of people or documents with that knowledge. In earlier times, the best way to do this was find a person or a book, but the challenge was to find the right resource in the least time. This problem was partly solved with the advent of computers and documents turning digital, search engines took the task of indexing every digitized content, and the problem reduced to “ask a program” or “ask a person”. Nonetheless the problem of finding and contacting that right person who can answer your query remains a challenge. This defines the need for expert finding systems, the topic of our interest.

Expertise location is “ask a person” task and expertise finding systems are a plenty both in enterprise and academic domain. Thus, to understand the design of these systems, and the parameters that influence their produced results, we classify these systems into categories discussed in subsequent sections.

2.1 Expert Finding

Early approaches to expert finding included manual construction of database, housing skills and knowledge of individuals in the organization. This took considerable effort to setup and maintain. Next, came in the automated approaches, automatically mining the enterprise repositories to build profile of an expert. Most of these automated approaches handled expert finding in specific domains, extracting information from known document types. For example, mining source code documents to see who last modified the code and where, this was an indication of a candidate’s involvement in a project and size and frequency of changes hinting towards
his expertise. Further advancement came in with the need to index heterogeneous
documents to make these systems more widely applicable. Most of this work was
performed in industry and the details of the published solutions remained high-level.

The Text REtrieval Conference (TREC) started an Enterprise Track in 2005 with
the goal to let participants perform experiments on enterprise data and fuel the re-
search in this field. Participants were provided with a crawl of World Wide Web’s
Consortium (W3C) website, a list of candidate experts and a set of query topics.
The task was, \textit{Given a topical query, find a list of candidates who are experts in it.}
This conference continued till 2008 and from most of the submitted solutions two
principle approaches of expert finding emerged, these were formalized by Balog et
al. [4] and further compared in [5]. These approaches were called \textit{document-model}
and \textit{candidate-model}. Each of these model, rank candidates according to probability
of them being an expert given the query, but they differ in how this is achieved. Mc-
Donald and Ackermann [6] distinguish several aspects of expert finding, \textit{Who are the}
\textit{experts on topic X?} and \textit{What does expert Y know?}. The candidate and document
model formalized by Balog et al. [5] focus on the first aspect of this expert finding.

- **Candidate Model:** In candidate model or model 1, a representation of candi-
dates knowledge is build using the documents they are associated with, this
is also referred to as expert candidate profile building. Experts are found by
finding profiles that best match a query.

- **Document Model:** In document model or model 2, documents are retrieved
and ranked based on the query. Experts are then found by identifying can-
didate experts associated with those documents. Here, documents act as a
latent variable between a query and a candidate. Hence, document model
is based on a search engine approach, first look for document, find out who
wrote them and then contact the author.

Both these models assume the existence of \textit{document-candidate} association,
which is stated as, a document $d$ in a collection is associated with a candidate $ca$, if
there is non-zero association $a(d,ca) > 0$. This measure quantifies the degree to
which a document is representative of a candidate’s expertise or it can also quan-
tify the extent to which the candidate is responsible for the document’s content. It
is formulated as, estimation of the probability that a document $d$ is associated with
candidate $ca$. These document-candidate association are build using two key ap-
proaches described below.

- **document-centric**

$$p(d|ca) = \frac{a(d,ca)}{\sum_{d' \in D} a(d',ca)}$$
where D is the document collection. According to this approach, a candidate will be most strongly associated with the top documents resulting from this.

- **candidate-centric**

\[
p(ca|d) = \frac{a(d, ca)}{\sum_{ca' \in C} a(d, ca')}
\]

where C is set of candidate experts. Using this approach we find candidates who made the biggest contribution to document d. Here, we assume that candidate ca is one of the people who made some contribution to d.

### 2.2 Expert Finding Evaluation

Expert finding is an Information Retrieval (IR) problem and its evaluation follows similar performance metrics, that is, how good is the system in retrieving relevant documents. In order to evaluate the effectiveness of IR systems, two separate approaches are used, these are, user-based and system-based. User-based approach relies on user’s feedback on retrieval performance, user-interface and other aspects of the system. It requires observing and analyzing user’s interaction with the system to measure their satisfaction levels. This process is human participation intensive and can prove costly and time-consuming.

On the contrary to user-based approach, a system-based approach relies on experiments to evaluate the performance of retrieval system. Such evaluations require a test collection framework. A test collection comprises of (i) a document corpus, which is a collection of substantial number of documents, (ii) topics that are set of pre-defined queries formulated in a standard format, used by the retrieval system as search-query, and (iii) a set of relevance judgments, made by human assessors. The retrieval effectiveness of the system is measured on the basis of number, fraction and relevance of the returned results to the query topic. System performance is quantified based on a chosen evaluation metrics such as precision, recall, average precision (AP), normalized discounted cumulative gain (NDCG) among others. Per-system per-topic scores are calculated for systems and then scores are aggregated to obtain a single overall performance score over a set of query topics.

One of the bottlenecks of this evaluation task is assessing the relevancy of documents to a topic. This is a costly step since it relies on human assessors to perform the relevant judgments, this incurs expenses in terms of hiring expert judges and the time spent in the assessment. With increase in number of documents in document corpus, for example, TREC expert search corpus, it becomes impossible to have complete relevance judgments. Thus, the relevance judgment step is commonly generated based on a pooling method of retrieved documents. Furthermore, performing relevance assessments suffer from several disadvantages, (i) it is time-consuming and (ii) it is costly.
intensive, (ii) over the period of time assessors can becomes less precise which may
effect their judgment and, (iii) assessors may disagree in declaring the document
relevant or not depending on the subjectivity of the topic. We will talk about these
methods in the subsequent sections covering both high-cost and low-cost evalua-
tions.

High cost relevance judgment  The pooling method has been the de-facto ap-
proach since it produces sufficient number of judgments for achieving reliable re-
results. The considered relevance scale in this approach maybe in a binary or a
graded-format. Binary judgment means the retrieved document is either relevant
or non-relevant. Graded format formalizes retrieved document as highly relevant
(2), relevant (1) and non-relevant (0). This is also called as a three-point relevance
scale. In pooling, partial judgments are performed by selecting a set of top \( d \)
doc-
uments retrieved for each topic from the runs, and pool of documents is created
for assessment. This reduces the number of judgments that need to be performed
for the assessment. All the non-judged documents outside the pool are considered
as non-relevant and the same process for evaluation is then carried out on this set
of documents. Pooling proves disadvantageous when systems contributing to the
pool are not equally accurate and there is a defective system, which may lead to
adverse effects. Moreover, the systems contributing to the pool and otherwise are
all assumed to be equal and are scored with same level of reliability.

Low cost relevance judgment  Moghadi et al. [7] in their survey of low-cost
evaluation techniques, categorized the low-cost retrieval evaluations into categories
described below. We will consider one approach under each category and will use it
to explain the concept.

- **Calculating robust evaluation measures to cater for incomplete judgments**  Various approaches are proposed to deal with incomplete judgments, one of them
is to use an evaluation metric that handles missing judgments. BPref metric is
introduced which performs evaluation based only on judged documents.

- **selecting the best set of documents to be judged**  Move-To-Front (MTF) pooling
is one approach under this category. The idea is that the documents are
judged in order of rank, and the systems are also prioritized according to the
previously judged documents.

- **statistical inference of evaluation metrics**  Here, documents are first sampled
for judgment and then the evaluation measure is statistically estimated. One
technique is to employ uniform random sample of the pooled documents and
infer the pooled average precision (AP) to create a new metric, infAP.
• inference of relevance judgments Identifying relevant documents amongst unjudged documents using existing judgment is done in this category.

• topic selection There exists a problem of high number of topics while having a small judgment set. It is shown that evaluations with more topics and lesser judgments are more reliable, thereby reducing assessor’s efforts. Hence, the greater the number of topics in the test collection, better the possibility of having higher system rankings.

• techniques to test the reliability of the evaluation and reusability of constructed collections One of the key techniques is, confidence interval generation for evaluation metrics on the basis of logistic regression. Using this method, existing relevance judgments are deemed appropriate for evaluation of a new system if interval is tight. But if the interval is wide, more documents need to be judged.

• alternative methods to pooling These include methods which do not use any relevance judgment, or use judgments generated from alternative methods such as search results retrieved by other effective retrieval system (like crowdsourcing). The ones not using relevance judgments have also gathered considerable interest. One of the first examples is by Soboroff et al. [8], they suggested generating pseudo-relevance judgment by declaring random selection of documents as being relevant and observed it produced good approximation judgment set and reliable system rankings. This is also referred to as no-cost evaluation. This approach sparked an interest in this field and inspired many other solutions. Aslam et al. [9] formalized a simple way to measure the similarity between two retrieval systems by computing the ratio of the number of documents in their intersection and union. This approach also ranks the best performing system with poor performers, similar to Soboroff’s approach. They hypothesize that these two systems are assessing the systems based on "popularity" instead of "performance". Their analysis suggests that "popularity" effect is caused by considering all the runs submitted by a retrieval system, instead of only selecting one run per system.

Due to design constraints of our work, we will use alternative methods of pooling technique of low-cost evaluation to examine our expert finding systems. Its details are discussed in Chapter 5.

### 2.3 Existing Expert Finding Systems

In this section, we begin by describing the design for an expert finding systems and then, provide an overview of existing expert finding systems based on two dimen-
2.3 Existing Expert Finding Systems Background and Related Work

sections which are relevant for our work. We first introduce systems segmented on the basis of type of data sources they utilize as their document collections. Secondly, we introduce types of existing EF systems, along with their employed computing mechanisms.

**Expert Finding Systems: Design** One of the key elements of an expert finding system is the user profile, which stores information about each person in the organization. It is one of the most common ways to find out about other people. Hence, a user's profile needs to put forward information a seeker would need to know about another person to establish their credentials, find commonalities and know their availability.

According to Ehrlich et al. [10], there are three key factors a seeker should be able to extract from a profile: person's credentials, the likelihood of him responding to an unsolicited query, and accessibility. From these factors, we can identify five primary dimensions of a person's expertise. First, expertise is embedded in an organizational context. For this reason it is important for a seeker to have information about where and how the expert resides in the organization, called **demographics**. Second, a person's credentials are established through a combination of **credibility**, **observed behavior**, and **reputation**. Finally, accessibility acknowledges the need to represent information about **availability** and preferred modes of communication as well as general access to the person. Hence, for a good system design, these expertise dimensions in form of sources of information about a person should be present in the system.

We will use these expertise dimensions against our data set for analysis, and for assessing the extent to which the required sources of information are fulfilled by our data sets.

### 2.3.1 Based on Data Sources

In this section, we classify EF systems based on the data sources they utilize for building their document collections. Here, data sources refer to enterprise data sources and external public social media. Enterprise data sources could be private or public within the enterprise network. Public data sources within the enterprise comprise of content sharing sites, social networking and wikis. Whereas, private sources include emails and personal chats.

Experts are always in demand, be it for conferences looking for reviewers, HR personnel for forming project teams or recruiters looking for talent. Large scale social interactions inside and outside the enterprise have made the conventional methods of document browsing and manual expert identification irrelevant. Thus,
expertise finding systems which focus on these data sources and their combination are emerging. In the following sub-sections we will examine such systems.

**Public Social Networks** Judging expertise of users in online social networks is a key challenge. One of the recent examples is Aardvark [11] which routes the query to an expert within user's social networks. It has the capability to import a user's social network like Facebook, LinkedIn, Twitter, Webmail program etc. and also has the option to invite other users by email. The data gathered from these data sources is then used to populate a person's Social Graph and for computing user's expertise. Liao et al. [12] argue that topical expertise of a user is not just dependent on the tweets and re-tweets but benefits the most from other user-related information (bio and user-lists), their results are backed up by user study and feedbacks.

**Enterprise Private Networks** In Enterprises, people from Human Resources, managers etc. are constantly looking for experts, usually this is done by interaction with people, or through a chain of referrals. SmallBlue [13] or IBM Atlas tries to simplify this problem by automating the process. Atlas resides on a participant's machine and infers his social network and expertise by going through his private network communications (outgoing emails and chat transcripts). SmallBlue's main components Find, Reach and Net focus on expertise finding, Find ranks people according to their expertise and Reach computes the shortest path to reach those persons through participant’s social network, whereas Net interlinks experts based on topics and helps in finding alternate experts which maybe well outside participant’s social network.

Despite being prone to privacy issues, this method proves to be advantageous since private data sources like emails and chats are rich in content, updated regularly and are used by most of the people.

**Enterprise Public Social Sources** Introduction of social media in the enterprise has encouraged the "outside the mailbox" user interaction within the firewalls.

Guy et al. [14] made use of this by examining 8 enterprise social media sources for expert and interests mining. Their findings show that profile tags, micro-blogs and forums were more appropriate for expertise finding whereas communities, wikis and bookmarks were more precise for interest inference. Their evaluation of these sources reveals that two relations between a user and a topic, interest in the topic and expertise in it are different things but both relevant for a user exploring that topic.

**Combination of Enterprise Public and Private Sources** Expertise miner systems benefit from evidences emerging from diverse sources including a user’s public
and private network. Guy et al. [15] developed SONAR (Social Network Architecture) API which combined more than ten public and private sources used within IBM intranet. It aggregates the results from these sources using a user configurable weight for each data source and provides an API, which clients like an expertise miner system can use to answer queries such as “who knows about <topic>?” and “Do I know someone who knows about <topic>?” Their evaluation results confirm that this aggregation produces better results and no single source alone is sufficient.

Combination of Enterprise Sources and Public Social Networks Most of the expert finding systems are developed around organizations but the information evidence and quality contained within the intranet is not always sufficient. Hiemstra et al. [16] propose a system which combines data from six diverse Global Web sources including Yahoo! (Global Web Search, Regional Web Search, Document-specific Web search and News search), blog search, Google Blog and Book search API with enterprise data. Their study examines the data acquisition mechanisms from each of these sources, and finally rank these aggregated results by summing the negatives of ranks for a person from each source (also called Borda count). Their experimental results demonstrated that rank results from combination of sources were significantly better than initial results on enterprise corpora. The authors did not consider social networks like LinkedIn and Facebook since they believe not all the employees are active on these media. Taking inspiration from this work and also since there is not a single example which combines the Social Networks and Enterprise sources, this combination is the main theme of our work.

This section segregated EF systems based on their choice of data sources to build test collections. In the next section, we will look at types of expert finding systems and their employed techniques to locate expertise.

2.3.2 Types of Expert Finding Systems

Expertise is a vaguely defined term with a main focus on association of a person with a topic, stronger the association more likely is a person to be an expert. The following sections explores different techniques developed over the years for expert finding.

Q&A systems The aim of a Q&A system is expert location, with focus on: “Given a question q from user u, the system should return a ranked list of candidates c that maximizes score(u, c, q), where the score is the measure of probability of a successful answer to q”.

12
Expert Location without Graph constraints This content based traditional method deals with a task of finding expert from within a large document collection. Text REtrieval Conference (TREC) in 2005 launched a Enterprise Search Track which focused on Expert finding from large set of documents, encouraging many contributions in this field. Balog et al. [5] in their work assess methods and algorithms for finding experts in a scenario where they were given a crawl of World Wide consortium’s website, a list of candidate experts and a set of topics. They came up with two acclaimed models: Candidate model and Document model. Their approach answers the basic expert finding question but instead of directly computing the probability, they apply Bayes’ theorem. The Candidate model differs from the Document model in the sense that candidate is not modeled rather document act as a hidden variable in the process. This approach was also covered in detail in earlier section 2.1, expert finding.

Expert Location with Score Propagation In this approach, result list of probable candidates is re-ranked by also taking into account user’s social graph. These systems mainly used implementations of PageRank [17] and HITS algorithm for a social network setting to locate experts.

In Aardvark [11] the computed probability score has similarity with PageRank but differs as Aardvark searches for relevant users to answer the query rather than documents.

Expertise Recommender Systems In past recommendation systems have been used to recommend news, documents, movies, music and book. Systems that recommend people, or documents(acting as latent variable between the query and the candidate) are of interest to us. Expertise Recommendation tasks have two different dimensions, expertise identification and expertise selection which were not considered until pointed out by McDonald et al. [6] in their field study. These are discussed further below.

Agent-based approach Large part of recommendation systems are based on identification through referrals, agent based approach is one of them. Agent amplified communication [18] employed by AT&T labs in mid 90s and later transformed in to ReferralWeb [19] is an early example. It used user-bots for expertise location and simplified the task of person to person communication. These bots gather their owner’s expertise and close contacts and use this information to automate the task of expertise location and chaining of referrals by filtering the sending of messages to only relevant people. Yenta [20], a multi-agent matchmaking system, introduced people with similar interest to each other. It differs from [18] as it considers all user data available to infer a user’s interest.

Based on Contextual Factors Most of the above systems did not take into
2.4 Chapter Summary

consideration how people seek information while selecting an expert. Heath et al. [21] found that impartiality and experience of the expert plays a role, also as shown by Woudstra et al. [22] quality (reliability and up-to-dateness of an expert) and accessibility (physical proximity) plays a role in selecting the candidate expert. SmallBlue [13] is another example which takes into account social network information of the user and augments “who knows whom?” to “who knows what” furthermore simplifying the problem of reaching the expert. Smirnova et al. [23] suggested a user oriented approach by considering two extra factors the time to contact a person and the difference between the candidate’s knowledge compared to that of user, which they call knowledge gain. Contact time is based on many factors including organizational network, geographical network and collaboration network. The calculated time is the shortest path between the corresponding network nodes. Their work along with Hoffman et al. [24] work show that context based models are better than content based.

Yarosh et al. [2] did a field study of how helper-finding tasks are accomplished in a large enterprise. By interviews and documented diaries maintained by 36 professionals ranging from different disciplines, they found that majority (76%) of the users turned to their colleagues for a helper-finding task, searching the company directory came in second (20%) and less than 1% resorted to any internal tool. Strikingly low usage of technology indicates that people are still using the old ways of expertise finding as described in McDonald’s and Ackerman’s 1996 study [6]. Since expert finding is not always concerned with finding the right answer but can also lead to co-construction of knowledge [25] and development of strong ties, hence apart from considering contextual factors governing the selection criteria, the systems need to consider factors related to candidate and the task at hand. For this purpose authors aligned helper-finding process along three dimensions tasks, topics and helper selection criteria, also the results of conducted interviews confirmed that all these 3 dimensions are not well supported by current systems. They highlight the importance of including experience of the candidate with some evidence, candidate’s similarity with other candidates as a major selection criteria, and since their results show people are usually interested in collaborating for a longer time to solve their tasks, thus the systems should not just focus on providing quick answer platform.

2.4 Chapter Summary

The work in this chapter, serves to inform readers of our research design, which will be discussed in Chapter 3. Here, we would like to briefly mention the connection between the related work and our design. Our research framework is most similar to the work of Hiemstra et al. [16]. Just like their system, we will also use combination
of enterprise sources and public social networks as data source for our system.

We will assess these data sources across different expertise dimensions introduced in this chapter and will finally, use low-cost evaluation method to assess our expert finding system. These choices are important as our system does not make use of relevance judgments and pre-defined queries.
Chapter 3

Research Framework

We define the work done in this study based on a framework, which covers the description of the framework components, their development sequence and the how they communicate with each other. Figure 3.1, provides a high-level overview of this framework, the modules can also be seen as the main contributions of this work. The framework is designed to work in a generalized manner by making each component as modularized as possible.

![Figure 3.1: Process Pipeline of the Research Framework](image)

**Design decisions for the research framework**  Before we explore each component of our research framework in detail, we will describe the design decisions that
Research Framework

helped us to develop this framework. The motivation for this work emerged from a previous study [26] which was also conducted at IBM. This study used IBM’s employee directory (Bluepages) and LinkedIn to comment on how enterprise employees use and perceive social media. The study had 134 participants from IBM, whose profiles were used to examine characterization of organizational information in external professional social networks.

The goal of our study is to build on the same principle, but we aim to both broaden and deepen the scope of our work in comparison to the previous study. We wanted to extend our user base and also enhance the study by analyzing user profile information for expertise assessment. To broaden the scope, this study was conceived as a part of a broader project which also included the game tool [3] project. This tool served as the front-end to obtain user registrations, hence, data availability for our work was dependent on this tool. To deepen the scope, we extended the data sources to include enterprise social media source. IBM connections was chosen for this purpose. It includes various legacy IBM employee related sources of information, and is also the most popular social network used within IBM at the time of this study. IBM Connections offers a developer API \(^1\) to access its resources, and many of them are access limited by OAuth authentication. Hence, our choice of data fields in IBM connections was influenced by two main factors: (1) ease of access, without getting into administrative permissions, and (2) relevance of these fields for expertise assessment. The relevance was decided by manual inspection, and inputs from an IBM study [14]. This study reported that user profiles, user blogs and bookmarks are indicators of user’s expertise.

While the game tool was being developed, we developed our process pipeline for the framework and tested the first prototype. This prototype is introduced in next section. This early prototype helped us identify problems in our methodology and gave us time to correct it, before application to the target data set.

First prototype: data collection and analysis module

The decision choices mentioned in the previous section, helped us in designing our framework’s process pipeline shown in Figure 3.1. The availability of our user data set was dependent on the game tool, to account for the waiting period, we developed the first prototype of our framework. This prototype had only data collections and analysis module implementations, these modules are described in detail in subsequent sections. The goal of this prototype was to carry out the study until the data analysis phase. We used user data from the study [26] introduced in previous section. We wanted to emulate the data that we would receive from the game tool, hence we only used the LinkedIn

\(^1\)IBM Connections 4.5 API Documentation http://www.ibm.com/developerworks/lotus/documentation/connections/
data set of the previous study. And, crawled IBM connections for the set of users reported in LinkedIn data set.

The results of the analysis are reported in more detail in Chapter 4. Here, we only mention that they point towards the usefulness of addition of an external data source, and provided motivation for continuation of our investigation. We do not use the data set of first prototype in our main study because the LinkedIn data set is from an older crawl, and a new crawl is not possible without grant of permission from the users.

### 3.1 Data Collection Component

Data collection component is one of the key elements to carry out a data intensive study. Thus, making the processes involved in it, very crucial for the success of the study. The main challenge was to implement a system which allowed access to user’s enterprise and social profiles. Hence, this study was conceived as part of a broader project which facilitated obtaining required permissions for the data sources. The data collection component includes the game tool, authorization engine and implements the process of social networks’ modeling and their data extraction. User registration was secured through the game tool. Also, the game back-end was separate from our system back-end. As the data collection component was closely tied to the game tool development, we chose same development environment as used for the game-tool. This meant use of PHP as development language and choice of PostgreSQL for database server. Further details of the data collection component and its processes are described in following sections.

#### 3.1.1 Social Authentication System

When the user lands on the system’s landing page, the first step is to authenticate him for his identity. Following steps were followed as part of implementation of this process.

The first step, is to implement an authorization system which can handle OAuth2 requests as both of our external data sources LinkedIn and Facebook use them. We used HybridAuth PHP library, which provides OAuth2 implementation and easily extendable functions. As the second step, we used Facebook Graph API and LinkedIn Developer API to obtain user profile details. HybridAuth stores user session token keys corresponding to each user, which can be utilized to query profile details of the user. For this step, it is also required to have a developer account with these social networks, this enables obtaining developer API keys which are required for authenticating the developer while making queries to the social network API.
The usual flow of the implemented authentication system is: once the user lands at the main page of the game-tool, he is prompted to authenticate using his LinkedIn account. User’s adherence to this triggers two processes, (i) user’s HybridAuth session details gets stored in our main user database, (ii) this user is searched inside the IBM Connections using his full name for a profile match, if a match is retrieved, user is asked to confirm the returned match, there are cases where more than one user profile match is returned, in that scenario user was asked to confirm the correct result. Everytime the user logged in again or returned to the system his HybridAuth session key was refreshed and could be used to retrieve the updated profiles. Also, every new entry in the HybridAuth session key database triggered the crawl process which is described in the next section. Facebook was authenticated by fewer users as it was an optional step, and only the first of the above two triggers get activated. Next, we will describe permissions module, which is necessary to obtain access to user data residing in external social media sources.

**Permissions** One of the key concerns of collecting user data is privacy and security. We took care of this concern by requesting permission of the user to the data sources only appropriate for our study. This is important for data sources like LinkedIn and Facebook, which offer many profile attributes and each of their access require gaining permission from the user. The users were also made aware through the game tool about the purpose of this data collection and how their data will be used. In addition, they were given a choice to revoke rights to their data access
whenever requested. When a revoke was requested, use session keys and associated profile corresponding to the user were deleted. Furthermore, it was also made sure that the data resided securely within the enterprise.

3.1.2 Data Extraction

As mentioned earlier in the authentication step, registration or logging of a user triggers a crawl of their social network profile and IBM Connections profile. Figure 3.2 gives a general overview of functionality of the crawlers. The following section describe the crawlers developed for data collection from IBM Connections, LinkedIn and Facebook social networks.

IBM Connections The enterprise social network used within IBM is called IBM Connections. It offers an IBM Connections API\(^2\) which allows developers to retrieve information from it. Through this API, we were able to retrieve the following information for each IBM user who participated in our study:

- **Profiles**: This is a directory of the people in the organization, it includes the following information:
  - *user role*: indicates if the employee is a regular employee, contractor or a non-employee.
  - *title*: gives information of the designation of the employee within its company e.g. Application architect.
  - *organization*: home or mobile
  - *is-manager*: indicating employee’s Manager status
  - *country code*: of the employee’s assigned location, more location based data is also included in form of fields, *building and floor*.
  - *office number* which indicates the affiliation of the employee to the a Business Unit (BU)\(^3\)
  - *profile tags*: employee’s social-tags assigned by peers and user himself, includes skills, industries and clients.

- **Colleagues**: this contains information about employee’s colleagues

- **Blogs**: Includes posts authored by different users.

---


\(^3\) IBM business units: [http://www-03.ibm.com/employment/us/ibm_major_units.shtml](http://www-03.ibm.com/employment/us/ibm_major_units.shtml)
Bookmarks: contains information about employee’s saved bookmarks including their description of title, subtitle and link.

Two instances of IBM connections Crawler was used, first to gather users which registered with the tool and the other one was used to gather all the IBM Netherlands user records, so that we can compare our sample of the users with the user population of IBM Netherlands employees, first one was a periodic crawler that ran end of every day to collect user data and the latter one was run once at the end phase of the data gathering step.

The next step of the data gathering process is to extract the information from the collected data and store it into a data store, as depicted in figure 3.2, we use PostgreSQL as our data store and Figure 3.3 and Figure 3.4 show the schema of the extracted data. There are structural and semantic heterogeneity in the schema. Through a manual study of the schema we note that the overlap of user profile fields is minimal at 13%. This measure indicates that the inclusion of an external source is providing additional profile features of the employees. This already is an indication towards our RQ3, showing that the duplication of profile field is only 13%.
LinkedIn is a business-oriented social network which started in 2008, it allows developers to query its user database considering they have required permissions. As, the first step of the authorization is to authenticate user through his LinkedIn account, we asked for full-profile, connections and network permissions. For users who complied to the access request, information was gathered using LinkedIn API functions. We can see from the 3.3 that we were able to gather all the profile information from LinkedIn. Since, this social network is pretty vast, it took sometime to understand and model the entity-relationship process. Social networks liked Linkedin have well written API documentation but the challenge is the constant change in API. During the study-run there was a change in LinkedIn API, and the system stopped collecting a field after the API update. This remained unnoticed until the analysis step, to tackle this it is advised to perform regular sanity checks of the collected fields.

3.2 Data Preparation

A folksonomy is a collection of set of users, set of tags and a set of resources or objects, and there exists a ternary relationship between these fields. Since folk-
sonomies do not explicitly follow formal taxonomies, they reflect the vocabulary of the users of that system, they are easier to use, develop and maintain compared to developing a precise controlled vocabulary which is development intensive. But they present a trade-off as they do not include any classification and all terms in it belong to a flat hierarchy. The data sources used in our study, LinkedIn and IBM Connections Profile, are social networks and their folksonomies contains much larger and non-standard vocabulary. Furthermore, in these systems the concept of tagging is transferred to people where people assign tags to each other referring to their expertise, interests or job-roles. The LinkedIn data contains partially controlled vocabulary, although the system started with uncontrolled vocabulary but with learning curated its own vocabulary specially for skill terms and thus has a more generalized skill representation. Current implementation of LinkedIn recommends tags, based on the ones already used by others, but users are not restricted to just these and can also use tags of their choice.

On the other hand, IBM Connections profile has uncontrolled tags vocabulary and use free user input. IBM Connections’ profile tags contain IBM related concept tags, client tags which convey with whom the person has worked, and general skills tag. This makes the IBM Connections profiles tag a highly noisy data set. To conduct a comparative tag study of these systems, we need a data preparation step, which will bring the systems to similar vocabulary standards.

In the subsequent section, we describe the steps followed for data preparation of social-tagging dataset of both LinkedIn and IBM Connections. The process follows the steps shown in figure 3.5, raw tags are accessed from the data store and altered by means of scripts developed in R scripting language.
3.2 Data Preparation

3.2.1 Syntactic variation filter

Syntactic variation in tags refer to the inflectional and morphological variations in tags. Inflectional variation includes singular and plural forms of same tags, basically the inflectional of same lexical word, and orthographic variations include spelling differences, misspellings and other variations which cannot be addressed with lemmatization or stemming, eg, ABNAMRO vs. ABN AMRO vs. ABN which all refer to the same thing. This kind of variation is a result of tagging systems where users are not allowed to increment the current tag count, and thus different users perform tagging using more variations of the same tag. Such limitation is not present in LinkedIn.

To tackle this variation, we do the following steps:

- lowercase all the terms, and replace punctuations with whitespace and trim extra whitespaces, remove English stop-words and commonly occurring verbs e.g. developing, maintaining etc.
- perform Google search feature ‘did you mean’ to further improve tag spellings and disambiguation.

The second step of the above process also delivers plenty of NULL values in case the tags are not recognized by the Google engine. This happen in cases where the tag is a personal tag or enterprise specific tag not known to outside enterprise parties or the tags are abbreviations which cannot be disambiguated or identified. We deal with this NULL value challenge by keeping the original tag, but others failing to match the criteria are discarded. Also the results given by Google filter can be non-null but still be not representative of the original tag, in this scenario we compute the cosine similarity between the result and the original tag, if the value is higher than the threshold value of 0.75 we use the result, else we keep the original tag. Furthermore, we do not discard the abbreviations tag as it is a possibility the same abbreviations maybe used by the user across different tag-spaces as both the tag-spaces are related to professional social networks.

3.2.2 Semantic variation filter

Semantic variations in tag refer to synonyms and abstraction of the same concept. Synonyms can be dealt by looking at the dictionary or thesauri like Wordnet, but since the social-tagging data is noisy data and includes user’s personal interpretation of concepts and use of non-standard words, this solution would throw away a lot of terms.

To overcome this we try to use our own algorithm, which is described below. Also, we try to identify synonyms and near-synonyms based on combination of Lev-
enshtein distance and cosine similarity. We follow the following approach to achieve the ‘collapsing into same tag’ phenomenon:

- We first compute the Levenshtein distance between all the pairs of tags in the tag-space. Since we have tags of many different length we use normalized Levenshtein distance.

- Secondly, we compute cosine similarity between all the candidate pairs identified in the previous step.

The final measure for computing whether the tags are near synonyms or not is to multiply the values obtained in the above two steps. Here, we introduce a measure and we call it $\tau$. $\tau$ should be greater than a certain threshold. This threshold is identified to be 0.5 after performing several trial and error runs of maximizing collapsed pairs. Also, we need to take care that our tag-space has many acronyms, therefore, while computing $\tau$ we discount cases which have same number of characters and length less than 3.

After performing these two steps, we can collapse the tag pairs, and hence achieve semantic filtering. Application of both syntactic and semantic filter will in most cases reduce the vocabulary size of the tag-space but also align them closer to standard word representation.

This data preparation module is a necessity before making the extracted data available for consumption to other modules. The major challenge here is understanding the vocabulary of terms or the folksonomy of the data-set. This module comprises of several R-language scripts which were developed in an iterative manner, as many permutation and combination of preparation steps were tried out before finalizing the best possible mechanisms. Even before, exploring the mechanisms to prepare the data, it is important to explore the data-set in question to a great extent. Otherwise, this step can prove to be the most challenging step of the entire framework. We spent more than one month finalizing these mechanisms and to automate the entire module.

As part of the framework, this step can be triggered with a script which takes raw data from the database as an input and produces csv files corresponding to each data source. This module/step of framework is common for most data analysis tasks but is unique to each data source. Hence, in most scenarios, same preparation steps can not be used for more than one data source. The result of this module can also be further utilized by IBM to standardize their vocabulary of tags.
3.3 Data Analysis Engine

The purpose of this module is to receive the prepared data from the previous module and perform analysis on it. Here, analysis could be understood as extracting meaning from the collected data sources to either prove or disprove that the external data sources adds value to the expert finding task. As described in the next chapter, we will perform this analysis task with focus on expert finding dimensions namely, credibility and reach. Also, this kind of analysis helps us understand the similarity and differences between the target data sources. Therefore, this module comprises of several SQL and R language scripts, which perform pre-processing, querying, aggregating, summarizing, hypothesis testing on data and finally producing these results visually to be interpreted by the readers. To carry out these processes on data, one major task we do is to make a data model with all possible fields from the data sources that will support in carrying out these activities.

Also, this module caters to exploration of data sources to first understand their properties before data analysis, moreover, we used this module to perform sanity check on the data points, when we started crawling the social network data, to ensure good quality data.

Data Analysis is a highly context based task and the steps involved widely differs based on the domain of analysis. Our main challenge was formalizing how two social networks can be studied/analyzed to understand expertise. For this, we looked at the schema representation of both the networks and tried to identify fields that were common and also contributed to employee’s expertise. Next step, was to identify which dimensions of expertise will be used that are representative of the identified fields. This process required us to look back into published research related to expertise. This step remained a challenge until the above two steps were identified, but before we could identify them, it was required to explore the data sets thoroughly so that we have a grasp of properties of each of those networks.

3.4 Expert Finding Module

In this section, we will first discuss the existing expert finding methods employed at IBM, secondly, we will introduce our implementation of the expert finding system, and lastly we will explain the challenges that were encountered in implementation.

3.4.1 Expert Finding at IBM

Before EF module development, it was a paramount requirement to understand the data processing and implementation details of the EF system used at IBM. This was hindered firstly, by the presence of multiple expert finding (EF) systems in the or-
ganization with varying usage in terms of visitors per day. This was overcome by getting in touch with their product owners and through similar communications, we were able to identify the most relevant EF system for our study. We found that EF system named *IBM Expertise* was the one most supported by the organization. Moreover, through this communication, we were able to understand some details of this system, these details are shown as system overview in figure 3.6. The system uses IBM Connections Profile database as the data source. IBM Connections database is fed by legacy systems like Career framework, self-assessment data and Bluepages (employee directory). One endpoint to this database is Connections Profile API, which we crawled as mentioned in previous sections, and other endpoint is feeding data to an indexer, which caters to IBM Expertise Search API.

### 3.4.2 Expert Finding System: Implementation

A standard expert finding system is an information retrieval system which is supposed to answer the question ‘What candidate experts are associated with topic query X’. There are two major identified language models for EF systems in enterprise corpora to answer this question. These models are discussed in detail in ‘expert finding’ section of Chapter 2. Here, we mention them to indicate the reasons behind our choice of using one over the other.

- **Candidate Model**: it uses textual representations of individuals’ knowledge according to documents with which they are associated. This representation is used to assess how probable the query topic is to rank candidates.

- **Document Model**: it ranks documents according to a query, and then determines how likely a candidate is an expert by considering the set of associated documents.

We are not aware of the technique used by *IBM Expertise* to find experts. But, for our system we use *Candidate model*, as our data is in a representation similar to the one required for this model. Hence, the crawler for indexing and the index were based on *candidate-model* implementation. The process of building the index was automated by the crawler, which crawled for candidate experts and documents from the prepared data and stored the index in an index directory. This index directory was utilized by software implementation of the expert finding system. We encountered several challenges during this implementation, which are discussed in detail in Chapter 5.
3.5 Discussion

This chapter describes the framework used to carry out this study. The framework is composed of modules which communicate with each other to accomplish the goal of this work, i.e., how can the combination of enterprise internal sources and external social sources improve expert finding. This is to be achieved by first observing whether the external data source (LinkedIn) contributes any extra value to the expertise dimensions, secondly, this claim needs to be validated. For validation purposes, we develop an expert finding system from the collected data and perform experiments to validate the results.

Most of the components and their sub-components were developed in a modularized manner, they communicate with each other through scripts. The data collection component which comprises of authorization component, social network modeling and data extraction can be seen as one big module. This also includes the game tool which is only a front-end for our scenario, allowing user registration. The entire process of registering a user, authorizing and crawling their social networks and finally storing it as raw data in the database is an automated process. The data preparation module takes the stored raw data as input and performs cleansing and preparation processes to it. The output of this step is stored as comma-separated (csv) files which is further utilized by data analysis and expert finding module. The data analysis engine is a stand-alone module which uses cleansed data from .csv files, it performs several hypothesis testing to prove the worth of our data. The cleansed .csv files consisting of user’s social network information is utilized to de-

Figure 3.6: System Overview of *IBM Expertise*
velop test collection for consumption by the expert finding module. The expert finding component is also a stand-alone software module which has the capability to conduct automated experiments.

There is one major identified challenge in the implementation of user authentication module. It is the identification of the user in the IBM connections database. Identification was performed using IBM Connections Search API to look for profiles with given first name and last name retrieved from LinkedIn login details. The challenge is when the user is given a choice to confirm his IBM details among the retrieved results, many users would choose wrong details for some reason. This led to aggregation of wrong IBM profile during the crawling process, ultimately leading to inconsistency in the database records. This inconsistency was hard to catch as the profiles would have similar names, and their other details would also be difficult to match. For example, email can prove to be an identifier but majority of the users use different email for their professional and social records. To tackle this challenge early sanity checks that are done during initial data analysis step. The sanity check for this identification problem tries to match first their full names, their emails, then country across IBM and LinkedIn profiles in that order.

This framework is generalized, it can be used to perform similar case study for other enterprises. This is possible, if the organization has internal data sources comprising of employee profile information and other fields specific to their expertise and network. The framework pipeline will remain the same, only the component implementations will have to be changed. This framework can also account for scalability, addition of more participants or additional data sources is possible. To achieve this, data gathering component can be extended to crawl more sources. Major implementation change will be required in analysis engine, as additional source signifies an extra data exploration phase to identify relevant expertise sources.

We will use the framework described in this chapter as a foundation for performing data analysis, described in Chapter 4. Moreover, the results of this data analysis will be used to evaluate our expert finding system, described in Chapter 5.
Chapter 4

Expertise Characterization inside and outside IBM

Companies face the challenge of sharing their employee’s expert knowledge and uses tools to register, communicate and search employees as a knowledge resource. Expert finding tools act as a way of mapping knowledge by condensing employee’s demographic, credential, organizational and accessibility information. A survey [27] involving 25 companies, asked its participants to evaluate their expert finding systems on these characteristics. The findings of this survey insinuates that there is a correlation between number of profile elements and satisfaction with the system. Furthermore, it also establishes the importance of richer expertise descriptions and networking to build and bridge connections.

As discussed earlier in Chapter 2, section Design of an expert finding system, according to Ehrlich et al. [10] an expertise finding system should have three primary goals: establish credentials of the person who is found, increase the likelihood that the person will respond to an unsolicited query and establish the accessibility of the person. To achieve this, the system needs five sources of information about the person: demographics, credibility, behaviour, reputation, and accessibility. In this chapter, we will investigate the extent to which these required sources are fulfilled by expertise characterization within IBM and how the addition of external data sources can further improve this expertise characterization. This will help us in answering our RQ3 and RQ4.

The data corresponding to external and internal data sources are collected in the period between May 2014 and July 2014. During this period, 211 IBM employees participated in this study. The expertise characteristics of these 211 users will be discussed in this chapter. In addition, we will also introduce the result of our first prototype analysis, which establishes our motivation to perform this expertise characterization.
4.1 First prototype: analysis

In this section, we will discuss analysis details of the first prototype. This prototype was introduced in Chapter 3, it is an implementation of data collection and analysis module of the framework. As mentioned in its introduction, we will use data set from previous user study [26]. But, we will replace the employee directory (Bluepages) data set with IBM connections crawl for the same set of users.

The study in question was conducted between January 2013 and March 2013, and consisted of 134 participants. But, during our crawl of IBM connections, the data set was reduced to 105 users. This is attributed to users who were no longer part of IBM, and the call to their email address resulted in NULL responses.

The goal of this analysis is to give an early indication about the effectiveness of such an analysis for expertise assessment. Hence, we will analyse the data set of 105 IBM employees across three key information, user demographics, skill representation and colleague connection representation of users in internal and external environments.

Demographics  We begin by looking at the user demographics of the user group under study. 77.14% of the participants are male and 22.85% are managers. Maximum number of participants (39.04%) are from IBM Sales and Distribution business unit. 57.14% (60) of users were from Netherlands, 13.33% (14) from Belgium, 9.5% (10) from U.S.A., and 21% from other European nations.

Skill representation  Table 4.2, shows summary of 105 user’s skill tags count on IBM connections and LinkedIn. The values do not show any clear difference in the skill representation, except the maximum tag count, which is 1.7 times higher in IBM Connections than LinkedIn.

Friend network representation  Table 4.3, shows summary of 105 user’s friends count on LinkedIn, and IBM connections. LinkedIn friends count only considers people from user’s friend list who work at IBM. From this table, we can see that user’s median friend count in LinkedIn is 5.9 times higher than on IBM connections.

From the above discussed measures of our first prototype, we can see that there is value in addition of an external source. In the above study, most value can be derived from friend network representation on LinkedIn. This gives confidence to continue this study further with our data set. Also, this prototype serves the purpose of going through an iteration of the framework until the analysis module. This helped us find challenges earlier in our analysis.
4.2 User Demographics

In this section, we will discuss the first source of information, i.e. demographics, which is necessary to achieve the primary goal of establishing a person’s credentials.
Organizational Demographics  

Organizational demography is defined by Lawrence et al. [28] on various characteristics: demographic unit, attributes, domain and measure. Demographic unit is the level of analysis to which theoretical generalizations can be made. Its definition includes individuals within groups, dyads, groups, and organizations. In our analysis, we will make generalizations on individual and group level.

Attributes are defined at level of analysis at which data is collected rather than level of analysis of the demographic unit. Individual attributes are classified into 3 categories: (1) attributes that describe immutable characteristics such as age, gender and ethnicity, (2) attributes that describe individuals’ relationships with organization, such as organizational tenure or functional area, and lastly (3) attributes that identify individuals position within society, such as marital status. Table 4.4 and 4.5 report on attributes of our 211 participants across gender, country, and functional area (title and business unit). We see that 21.8% of the user base are female and 78.2% are male. Maximum (79.62%) of the participants are from Netherlands as the game tool [3] to collect this user base was developed and advertised in Netherlands. The involvement of participants from other countries (9% Belgium, 6.16% Romania, 1.42% each from UK and USA and remaining 2.84% from Argentina, France, Germany, Mexico, Peru and Poland) can be attributed to the invitation element of the tool [3], which enabled participants to invite their peers for participation in the study. Most of our participants are non-managers (78.19%) and belong to Global Business Services (33.17%) and IBM Sales & Distribution (32.22%) business units.

In (Table 4.5), we examine individuals’ attributes based on their relationship with the organization, which includes their tenure and position. We see that 9.95% of our participants have a tenure of 0-2 years with IBM, 11.37% with 3-5 years of experience, 20.85% have 6-10 years, 24.17% (maximum) have 11-15 years, 14.69% have 16-20 years and 13.24% have 20 years or more experience within IBM. 11 out of 211 participants did not write start and end year on their job positions field, which is used to compute the tenure of the participants. We also notice that we have slightly more participants with total job positions between 0 and 5 years (48.34%) than between 6-10 years (43.12%).

Domain is a context within which a demographic unit is studied. In this analysis, we will consider group-level analysis by measuring certain attributes across business units. These will be discussed further in subsequent sections. Measures are either simple or compositional. Simple measures are defined at the level of analysis of the attribute. Compositional measures are defined at a level of analysis higher than that of the attribute. In our study, we are using both simple and compositional measures. For example, tag intersection ratio is a compositional measure. In next sections, we will introduce the measure that we use to conduct our analysis.
4.3 People Tagging Evaluation

As discussed earlier, to fulfill the primary goals of an expert finding system, the system needs to establish credibility information about the user. Credibility is related to the knowledge of the person, how they gained that knowledge and how well versed are they in it. To achieve this, we will first evaluate the skill set of our user base across the different social networks, namely, LinkedIn and IBM connections. As mentioned in previous chapter's data preparation section, user’s skills sets exist in these networks as social tags assigned to the user. They exist as uncontrolled vocabulary and we followed the data preparation step of previous chapter to improve tags align-
4.3 People Tagging Evaluation

Table 4.5: Demographics attributes of 211 participants (2)

<table>
<thead>
<tr>
<th>Tenure with IBM (in years)</th>
<th>Number of Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-2</td>
<td>21</td>
</tr>
<tr>
<td>3-5</td>
<td>24</td>
</tr>
<tr>
<td>6-10</td>
<td>44</td>
</tr>
<tr>
<td>11-15</td>
<td>51</td>
</tr>
<tr>
<td>16-20</td>
<td>31</td>
</tr>
<tr>
<td>21-25</td>
<td>8</td>
</tr>
<tr>
<td>26-30</td>
<td>15</td>
</tr>
<tr>
<td>31-35</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total Job positions</th>
<th>Number of Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5</td>
<td>102</td>
</tr>
<tr>
<td>6-10</td>
<td>91</td>
</tr>
<tr>
<td>11-15</td>
<td>14</td>
</tr>
<tr>
<td>16-21</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4.5: Demographics attributes of 211 participants (2)

ment before conducting the evaluation step. In the subsequent section, we will first describe the vocabulary of the social-tags in these networks and the achieved improvement after data preparation step. We will also look at the similarity between user tags in LinkedIn and IBM connections. Further, we will evaluate data enrichment as a result of addition of LinkedIn skill tags. These will help us in answering our RQ3 and RQ4.

4.3.1 Vocabulary Evaluation

As the first step of evaluation, we look at the distinct tag terms in both the networks. Table 4.6 shows the vocabulary terms of the networks in the raw state, then after applying the syntactic variation filter and lastly after application of semantic variation filter. We note that, since LinkedIn is a partially controlled vocabulary, the syntactic filter does not have major effect on the number of tag terms. Moreover, IBM connections system show constant decline of tag terms after each filter step, this is an indication of multiple variations of tags collapsing into one and also due to the fact that this system contains specific terms and acronyms which are not comprehended by the filters.

For further vocabulary evaluation, we define tag intersection. It measures the number of distinct tags that exist in both LinkedIn and IBM Connections profile divided by total distinct tags in both the systems. Table 4.7 shows this measure cal-
4.3 People Tagging Evaluation

Table 4.6: LinkedIn and IBM Connections vocabulary terms

<table>
<thead>
<tr>
<th></th>
<th>IBM connections</th>
<th>LinkedIn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>1985</td>
<td>1157</td>
</tr>
<tr>
<td>syntactic variation</td>
<td>1851</td>
<td>1157</td>
</tr>
<tr>
<td>semantic variation</td>
<td>1668</td>
<td>983</td>
</tr>
</tbody>
</table>

culated on raw tags, and after filter application. We see that this measure is lowest for raw tags, indicating that in the raw form, both profiles have least similar set. Whereas, both syntactic and semantic variation intersection measure has similar value, showing that with both variations, we achieved same profile similarity sets.

Table 4.7: LinkedIn and IBM Connections profile tags vocabulary intersection

<table>
<thead>
<tr>
<th></th>
<th>Tags vocabulary intersection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>0.0458</td>
</tr>
<tr>
<td>syntactic variation</td>
<td>0.0651</td>
</tr>
<tr>
<td>semantic variation</td>
<td>0.0623</td>
</tr>
</tbody>
</table>

4.3.2 User Tag Semantics

User-Tag Statistics Both IBM Connections Profile and Linkedin give users the ability to list their skills, which can be endorsed/added by their peers and themselves. Since Linkedin API does not let us retrieve the number of endorsements per skill, we will not look at this factor. Our main purpose here is: To find whether there is any significant difference in skills count per person listed on both networks.

We start by looking first at the summary statistics of the tags, which are shown in Table 4.8, it can be seen from the table the median and mean tag counts are higher in LinkedIn. Also, it is important to note that maximum value of tag counts in Linkedin is capped at 50. In our user base 13 (6.1%) users are affected by this, which can be seen in both Table 4.8 and Figure 4.1b.

Table 4.8: Summary statistics of Tags Count

<table>
<thead>
<tr>
<th></th>
<th>IBM connections</th>
<th>LinkedIn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Median</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>Mean</td>
<td>15.38</td>
<td>22.61</td>
</tr>
<tr>
<td>Maximum</td>
<td>112</td>
<td>50</td>
</tr>
</tbody>
</table>
We also examine the tag distributions in both networks per user to understand how often our users are tagged. Figure 4.1 shows the distributions, from Figure 4.1a, we can see that high percentage (25.1%) of our user base has only 1 tag in their IBM connections profiles and there are some users (5.68%) which have more tags (> 45) than others. Whereas, from Figure 4.1b, we see that user tag counts are more distributed in LinkedIn. This observation leads to the main question "Is the skills representation equal on both networks?" which is a way of answering our RQ3.

First, we look at the distribution and summary statistics of the difference of tag counts of user, shown in Figure 4.2 and Table 4.9, we see that there are some users for whom we see high difference in tag counts, and the maximum difference is limited by the LinkedIn limit of capped tags. We perform a hypothesis testing for significance of difference in tag counts of users on both networks, to arrive at an answer.

Table 4.9: Summary: Difference in tag count of LinkedIn and IBM connections tags

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Median</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>-81</td>
<td>10</td>
<td>7.22</td>
<td>49</td>
</tr>
</tbody>
</table>

We use Shapiro-Wilk normality test to find the normality of skill count difference distribution. The results report a p-value of 3.325e-09. From this, we know that the distribution of difference in skill count on both networks is non-normal, therefore, we can use two-sided one-paired sign test to test the significance. Hypothesis testing: $H_0$(null hypothesis) - Average difference in the skills count on two networks is zero. $H_A$(alternate hypothesis) - Average difference in the skills count on two networks is not zero. Reported p-value : 4.35e-13 with 95% confidence interval : (7, 12) Since
the p-value is less than .05 significance level, we can reject $H_0$ and conclude that the representation is not similar.

Table 4.10: Difference in Tag count across BUs

<table>
<thead>
<tr>
<th>Business Unit (BU)</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Business Services</td>
<td>8</td>
<td>7.7</td>
<td>14.76</td>
</tr>
<tr>
<td>Global Technology Services</td>
<td>8</td>
<td>11.3</td>
<td>13.67</td>
</tr>
<tr>
<td>IBM Sales &amp; Distribution</td>
<td>6.5</td>
<td>5</td>
<td>20.55</td>
</tr>
<tr>
<td>IBM Software Group</td>
<td>2</td>
<td>14</td>
<td>30.66</td>
</tr>
<tr>
<td>Integrated Operations</td>
<td>10.67</td>
<td>15</td>
<td>20.25</td>
</tr>
</tbody>
</table>

Since, the tags representation is not similar, it is worthwhile to understand the data points which contribute to the differences. For this purpose, we will inspect this measure (difference in tag count) across IBM Business Units (BU) in Table 4.10, we note that it is difficult to point out whether there are any significant differences in tag-count across BUs. This is also evident from Figure 4.3. Here, we will be doing a Kruskal Wallis test to confirm our hypothesis. This test is comparable to one-way ANOVA for normal distributions. We are not doing pair-wise comparison between business units as for that we would need to perform MANOVA test and there is no equivalent test available for non-parametric distributions. The test reports that $p-value = 0.1476$, hence we cannot reject our null hypothesis that there are
no significant differences in tag count of users on both networks across different Business units.

**Tag Intersection Ratio** We first want to check for the user’s tag intersection ratio across LinkedIn and IBM Connections Profiles, for this we compute a measure

$$\alpha = \frac{|t_{li} \cap t_{ibm}|}{|t_{li} \cup t_{ibm}|}$$

, this measures the number of distinct tags for a user that occur both in LinkedIn and IBM Connections Profile divided by total distinct tags in both the networks. We computed this value for each of our 211 users’ filtered tags, and the mean value came out to be 0.0199. For the raw tags, $\alpha$ was computed to be 0.0117, therefore we see slight improvement in tag intersection ratio of the filtered tags. This finally indicates that the LinkedIn and IBM Connections profile tags are quite diverse and have hardly any overlap.

**Tag similarity** In the previous approach, we computed the tag intersection ratio for each user which only indicated the vocabulary overlap of the tag-spaces. In this step, we will include the tag semantics and compute the tag similarity using cosine similarity for each of our 211 user’s. This cosine similarity is assumed to reflect text similarity of the user’s IBM Connections and LinkedIn tag-space. For this, we
first construct document term matrix, considering each tag vector as document and tags as its terms. So for each user there is a document corresponding to LinkedIn tags and another for IBM connections tag. We take this document term matrix to compute cosine similarity between the tag vector-space. Based on this approach, we compute the average tag-similarity measure to be 0.183, which is very low and indicates highly dissimilar tags in both networks. Figure 4.4 shows the tag-similarity for our user base, it is also useful to check these values across business units as shown in Figure 4.4b. From these figures, again the distinctive differences are not evident, but we can see that the median tag-similarity for Software Group is the highest, indicating that users of this BU use somewhat similar tags compared to users of other BUs.

**Tag categorization** In the above sections, we have looked at the tag-space similarity between the two networks, but until now we have over-looked semantic context of the tags used in the networks of study. The aim of this section is to first connect tags to their semantic concepts, we will call this process as tag-categorization. To achieve this, the meaning and the context of the tag needs to be derived beforehand.

- Map an input tag to a concept existing in a knowledge base (KB), in our case we will use YAGO taxonomy [29], which covers WordNet [30] and significant part of Wikipedia.
- expanding concept towards its taxonomic ancestors until reaching a reference ancestor. We follow an approach similar to the one mentioned by Cantador et al. [31].
Expertise Characterization inside and outside IBM 4.4 People Network Evaluation

(a) Tag-Category similarity across BUs

(b) Tag-Category similarity across experience of user

Figure 4.5: Cosine Similarity for LinkedIn and IBM Connections tags-categories

Using the above approach, we were able to identify 232 unique categories across LinkedIn and IBM Connections tags. The average tag-categories count per user in LinkedIn was calculated as 6.62 and for IBM Connections as 4.83. As we did for tags, we can also conduct similarity measures for tag-categories. For our user base, there are lesser data points than 211 as there are users for whom tag-categories were not identified from the knowledge base. The average tag-categorization similarity measure was computed to be 0.254 and the 3rd quartile was 0.72. This is already an improvement from tag-similarity measure, as this approach looks at the meaning of the tags by categorizing them into higher-level concept. We can also try to explore this measure with respect to business units and experience level of the employee. As seen from the Figure 4.5, we are not able to identify any trends or distinctive features. Furthermore, the similarity measure is low, indicating that the tag-spaces even after being identified by their high-level concepts, are dissimilar to each other. Therefore, we can say addition of an external data source will enrich the enterprise network. Furthermore, it also means that more diverse and increased quantity of user tags will establish increased credibility dimension of the user with respect to an expert finding system.

4.4 People Network Evaluation

Another important dimension of an expert finding system is the reach of the concerned candidate. Hence in this section, we will focus on evaluating whether an addition of an external source like LinkedIn helps in improvement of the reach dimension. We define this dimension as a measure of a candidate’s accessibility through
4.4 People Network Evaluation

Table 4.11: Summary: Friends Count in different networks

<table>
<thead>
<tr>
<th>Friends Count</th>
<th>LinkedIn</th>
<th>LinkedIn (IBM only)</th>
<th>IBM Connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Median</td>
<td>430</td>
<td>35</td>
<td>32</td>
</tr>
<tr>
<td>Mean</td>
<td>450</td>
<td>58.5</td>
<td>48.26</td>
</tr>
<tr>
<td>Maximum</td>
<td>1378</td>
<td>285</td>
<td>234</td>
</tr>
</tbody>
</table>

Figure 4.6: Distribution of Friends Count in both Networks

his connections, connection to more people and more diverse people can also be said to be a reach related measure. Therefore, we will begin by looking at the total friends count measure of our user base in both the networks, hence, the average friends count per user in LinkedIn is $450$ and in IBM Connections is $48.2$. It should be noted that LinkedIn also caps its total friends count limit to 500 to be returned by the API, so the friends count of these users are only visible to the users itself. Another point to consider is that the user’s friends count in LinkedIn also includes connections to people outside their own enterprise and hence for evaluation purposes, we will consider only those friends of LinkedIn users which also currently work within IBM. We achieve this by looking at the `current-positions` field of user’s LinkedIn profile, and filtering those users which report their `current-company` field as IBM. Figure 4.6 shows the distribution of the friends count in both networks, needless to say that they follow non-normal distributions. Table 4.11 shows the summary statistics of friends count in these networks, with LinkedIn clearly marking high numbers. It is also interesting to note, the statistics of users with only IBM friends in LinkedIn network, the near values hint at similarity in representation.

Further on, we will not use the user’s overall LinkedIn friends count but only the count of friends which are user’s colleagues from IBM land connected to them on LinkedIn.
We will further investigate the friends count measure by computing the difference in friends count per user. As seen in Figure 4.7a, we also do Shapiro normality test which gives p-value of 0.0004932. Hence rejecting the normality of the distribution. The average difference in friends count of both networks is 9.6, showing that the representation might be similar. Therefore, as a next step we perform Hypothesis testing for the significance of difference in friend count on two networks, since friend count on different network is compared for same set of users, we can use the significance test for matched pair. Thus, in this case of non-normal distribution, we use one-paired sign tests, which gives us a p-value of 0.7723 with a 95% confidence interval of (-4.262155, 9.000000). Therefore, we cannot reject the hypothesis that an average difference between the number of friends on two networks is zero.

Next, we explore the difference in friends count (friends on Linkedin - Friends on IBM connections) across Business Units. Figure 4.7b and table 4.12 shows the measures, from both we can observe that the median difference friend count is more for Sales and Distribution, and Integrated Operations BUs, which is probably due to the nature of their job which requires more connections within the company. On the other hand, the Software Group BU has very low mean and median indicating their use of IBM Connections more than LinkedIn for connecting to others within the company. We also do a significance test to evaluate the differences in friends count across business units, testing the hypothesis, the average difference in friends count on both networks across different business units is zero. We again conduct Kruskal wallis test here, which gives us a p-value of 0.01337, hence we can reject this hypothesis.

<table>
<thead>
<tr>
<th>Business Unit</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Business Services</td>
<td>9.8</td>
<td>-0.5</td>
</tr>
<tr>
<td>Global Technology Services</td>
<td>20.12</td>
<td>1.5</td>
</tr>
<tr>
<td>IBM Sales &amp; Distribution</td>
<td>16.53</td>
<td>13.5</td>
</tr>
<tr>
<td>IBM Software Group</td>
<td>-45.44</td>
<td>-29</td>
</tr>
<tr>
<td>Integrated Operations</td>
<td>16.91</td>
<td>17</td>
</tr>
</tbody>
</table>

4.5 Discussion

In this chapter, we carried out the exploration phase of our work, with the aim to analyse similarity and differences between the social data sources. We identified limita-
4.5 Discussion

Expertise Characterization inside and outside IBM

(a) Distribution of difference in Friends Count in both Networks

(b) Difference in Friends Count across BUs

Figure 4.7: Difference in Friends count distribution and across BUs

(a) Distribution of difference in Friends Count in both Networks

(b) Difference in Friends Count across Manager status

Figure 4.8: Difference in Friends Count across dimensions

tions in terms of ambiguity in terms coming from these sources. Presently no dis-
ambiguation is done on the tags, for example, Oracle could belong to many classes
including Computing, Places, Publications etc.. Possible solution would be to per-
form more cleaning steps. There is also the dilemma of expanding the acronyms or
not. In this study, we do not expand the acronyms, we take into account that there
are enterprise specific acronyms. These acronyms require IBM specific vocabulary
for expansion. This vocabulary development can be a part of future work. The work
done in this chapter was primarily a means to answer our RQ3, "Do the external
social media sources duplicate the data already contained in enterprise sources?", and RQ4, "Which attributes of social media sources can help in improving expertise
assessment?". We labeled the dimensions related to expertise as demographics,
credibility, behaviour, reputation, and accessibility. We saw in the user-tag evalua-
tion that the additional source had dissimilar properties to the IBM connections. This answers our RQ3, by concluding from this evaluation that the overlap of user skill tags is minimal, and hence duplicacy is minimal.

The network evaluation contributed to the reach dimension of expertise. This attribute also showed fairly dissimilar properties to the IBM Connections network, indicating an enrichment in terms of reachability/accessibility dimension of the employee expertise. More parameters of credibility dimension, like, how well the person knows a topic, how he acquired that knowledge can be answered by LinkedIn’s summary and job positions fields. Presently, LinkedIn users can list out their projects and summary fields, but these are not yet available through their developer’s API. These project fields can also add more value to the credibility dimension, further hinting that external source LinkedIn enriches the data. Hence, we can say that external social source, like LinkedIn, does not duplicate the data already contained enterprise sources. This is an answer to our RQ3 corresponding to IBM as an enterprise. But can be generalized only when similar case study is carried out for more enterprises.
Chapter 5

Improving Expert Finding with External Social Sources

In the previous chapter, we described the benefits of an additional data source and explored how it can enrich the expertise finding task further for any organization. We also describe expertise dimensions of these data sources that contribute more towards this. These dimensions were first introduced in Chapter 2. In this chapter, we will evaluate the addition of an external social source to the system with current social sources.

This chapter will describe the approach we applied to accomplish such a system, then we will explain the system, developed in order to achieve the evaluation. Further, we will discuss the experiment setup used for our case, followed by discussion of results.

5.1 Approach

There are two major formal language models for Expert finding in enterprise corpora, the candidate model, and the document model. We discussed the decision to use candidate-model for our systems in Chapter 3.

We use the basic approach of candidate model to build textual representations, since our data originates from a user profile, it is easier to build these representations as the document-candidate associations are already established by the data. We follow below steps to build the representations.

- Create a user profile for each user containing all appearances of that person including name and email-address
- Extract data from documents for users according to their user profile and the document type, and
- Combine these description into an expertise profile

**Challenges** In this section, we will discuss the challenges that were identified for expert finding phase of the research, which corresponds to the final research question, RQ5, of our study.

- First challenge, is we do not use the entire user base of the EF system deployed at IBM and have our own limited data set of 211 users for which we have both IBM Connections data and LinkedIn data. To tackle this challenge, we only requested IBM expertise’s development support to provide the remaining data sets for our user base. This way we could build our own version of their system for our user-base, which will be used in our experiments.

- Another pressing challenge is to obtain the topics for this system, against which we will run our tests and the relevance judgments. IBM was again helpful in providing the top 10 queries for their Expertise system, but the challenge here was, that the queries were skewed towards the expertise engine. By this skew, we mean that to adapt to the system the user starts furnishing queries for which he is certain that he will retrieve a response, this is a gradual process. This is evident from the top-5 query topics which were AIX, Citrix, Java, TSM and Puresystems, these terms are not easily identified by external knowledge bases and hence using these terms as topics without doubt will give poor results for the external data source.

Hence, we tackle this, by collecting most frequent terms from user’s skills in IBM connections and LinkedIn as IBM Connections Profile skill tags are also indexed by the expertise system. The last challenge is absence of relevance judgments, to overcome this we make use of evaluation techniques that can be performed on test collections without relevance judgments.

In the next section, we will discuss the experiment setup used in our EF system. The setup is designed to accommodate the above mentioned challenges.

### 5.2 Experiment Setup

In this section, we describe the setup of our experiment. A standard way in Information Retrieval, to measure the effectiveness of the experiments is through a test collection consisting of three things:
5.2 Experiment Setup

Improving Expert Finding with External Social Sources

- a collection of documents
- a suite of information needs in form of search queries
- a set of relevant judgments

Due to company and data specific choices, we will have deviations from the standard approach of creating a test collection which are described in the upcoming sections.

5.2.1 Document collection

Candidate experts  In our study we had 211 employees for whom we had both IBM Connections and LinkedIn profile information. We considered these employees and created a collection of 211 candidate experts. We used unique field of each employee as IBM email-address, LinkedIn Id and Full name.

Documents  We consider two sets of documents for our system: (i)IBM Connection Profile (title, user-role and office number), IBM Connection Profile tags, and, (ii) LinkedIn Profile (headline, summary, specialities) and LinkedIn tags.

5.2.2 Query topics

We gathered top-10 queries for IBM Expertise system, but the queries were skewed towards this EF system. By this skew, we mean that to adapt to the system the user starts furnishing queries for which he is certain that he will retrieve a response, this is a gradual process. This is evident from the top-5 query topics which were AIX, Citrix, Java, TSM and Puresystems. These terms are not easily identified by external knowledge bases, and hence using these terms as topics will in most cases give poor results for the external data source.

It is possible to create queries for which one can be certain that only a limited part of the collection will contain relevant documents. Therefore, we collected 20 most frequent terms from user’s skills in IBM connections and LinkedIn. These terms are listed in table 5.1.

5.2.3 Relevance judgments

Our system does not have a set of relevance judgments. We examine the possibility of using just the raw pool as relevance judgment set with no manual assessor effort. Since our document collection has a rich quantity of relevant documents, this method can prove helpful.
5.3 Performance metrics

In this section, we will list and use metrics to measure the performance of our retrieval systems. With these metrics we wanted to answer two key questions:

- How does the solution with an additional data source perform compared to the current approach used by IBM expertise?
- How does this solution perform compared to Expert Finding approaches?

As mentioned in earlier chapters also, we are not aware of the expert finding techniques used by IBM Expertise and hence we try to emulate their system in terms of data sources and calculate performance based on this setup.

Standard information retrieval research uses precision, variants of precision, and recall to compare systems across approaches. Our system does not use relevance judgments, and hence will use metrics specific to low-cost evaluation techniques introduced to the reader in related work, Chapter 2. Metrics related to these techniques and the ones appropriate for our system are retrievability measure, and similarity measures comparing two retrieval systems. As retrievability is an important measure for recall oriented applications.

**Retrievability** Retrievability provides an indication of how easily a document can be retrieved using a given retrieval function. Essentially, retrievability is the ease with which it is possible for any document to be retrieved. A document with high retrievability means that a user has a high probability of finding that document by querying. Conversely, a document with low retrievability in a particular retrieval model is likely to be difficult to find by the user, up to impossible for documents showing a retrievability of 0. Clearly, if a document is difficult or impossible to retrieve, in general, then it will be difficult or impossible to retrieve when relevant. It is the inability to retrieve certain relevant documents that will lead to low recall. To use this measure in our case, we will look at the number of document returned by our constructed retrieval systems.

**Similarity measure** Similarity measure is a simple metric which measures the similarity of two retrieval systems based on the similarity of their retrieved results. It is simplest of system measures indicating similarity in common returned documents. It is expressed as:

\[
\text{SysSimilarity}(Sys_1, Sys_2) = \frac{\text{Ret}_1 \cap \text{Ret}_2}{\text{Ret}_1 \cup \text{Ret}_2}
\]
where $Ret_i$ indicates the set of documents returned by system $i$. In presence of multiple retrieval systems, average similarity score is used, which is defined as:

$$\text{AvgSysSim}(S_0) = \frac{1}{n-1} \sum_{S \neq S_0} \text{SysSimilarity}(S, S_0)$$

### 5.4 Results

We create two retrieval systems: (i) considering document collection from IBM Connections only, emulating the IBM Expertise EF system, and (ii) considering document collection from IBM connections and LinkedIn.

We execute the set of our queries on these two retrieval systems. Table 5.1 shows the number of documents retrieved for each query for both the systems, we observe a higher return count of documents for each query with the system with combined sources. Also, we can compute the average difference in retrieved documents, which comes out to be 44.52381. Table 5.1 also lists the similarity measure of documents metric for each of the queries, the average value obtained for the two retrieval systems is 0.36.

### 5.5 Discussion

With the details and results of this chapter, we try to answer our main Research Question: “how the combination of external social sources and enterprise social sources result in better understanding of the employee expertise?”. The results obtained in this chapter shows higher document retrievability measures for the system with combined sources. The evaluation metric limits us to draw conclusion only related to average documents retrieved per query per system. This measure was higher in each case for the system with combined sources. The similarity metrics for the two retrieval systems reported an average of 0.36 which shows that the retrieval are not similar in the documents they retrieve. Our study lacks a base system against which we could assess our results. But we know from the literature, that this measure produces reliable ranking results when assessed alongside TREC track evaluations.
### Table 5.1: Count of documents returned after running topic-query sets

<table>
<thead>
<tr>
<th>Topics</th>
<th>IBM Connections</th>
<th>Combined</th>
<th>Similarity measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>analytics</td>
<td>25</td>
<td>42</td>
<td>0.59</td>
</tr>
<tr>
<td>application</td>
<td>5</td>
<td>37</td>
<td>0.13</td>
</tr>
<tr>
<td>architecture</td>
<td>8</td>
<td>87</td>
<td>0.09</td>
</tr>
<tr>
<td>benelux</td>
<td>58</td>
<td>61</td>
<td>0.95</td>
</tr>
<tr>
<td>data</td>
<td>17</td>
<td>68</td>
<td>0.24</td>
</tr>
<tr>
<td>design</td>
<td>4</td>
<td>39</td>
<td>0.09</td>
</tr>
<tr>
<td>development</td>
<td>21</td>
<td>91</td>
<td>0.23</td>
</tr>
<tr>
<td>marketing</td>
<td>14</td>
<td>41</td>
<td>0.34</td>
</tr>
<tr>
<td>mobile</td>
<td>23</td>
<td>31</td>
<td>0.74</td>
</tr>
<tr>
<td>process</td>
<td>5</td>
<td>73</td>
<td>0.05</td>
</tr>
<tr>
<td>project</td>
<td>15</td>
<td>85</td>
<td>0.17</td>
</tr>
<tr>
<td>sales</td>
<td>25</td>
<td>93</td>
<td>0.26</td>
</tr>
<tr>
<td>security</td>
<td>8</td>
<td>28</td>
<td>0.28</td>
</tr>
<tr>
<td>service</td>
<td>20</td>
<td>101</td>
<td>0.19</td>
</tr>
<tr>
<td>services</td>
<td>21</td>
<td>72</td>
<td>0.28</td>
</tr>
<tr>
<td>smarter</td>
<td>22</td>
<td>27</td>
<td>0.81</td>
</tr>
<tr>
<td>sme</td>
<td>8</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>social</td>
<td>34</td>
<td>56</td>
<td>0.6</td>
</tr>
<tr>
<td>software</td>
<td>14</td>
<td>90</td>
<td>0.15</td>
</tr>
<tr>
<td>solutions</td>
<td>15</td>
<td>57</td>
<td>0.26</td>
</tr>
<tr>
<td>strategy</td>
<td>14</td>
<td>125</td>
<td>0.11</td>
</tr>
</tbody>
</table>
Chapter 6

Conclusions and Future Work

This chapter gives an overview of the project’s contributions. After this overview, we will reflect on the results and draw some conclusions. Finally, some ideas for future work will be discussed.

6.1 Contributions

This section will discuss the contributions made in this work. The contributions are driven by a general framework which was developed as the first part of the work. This framework contains modules and their description which can be applied to any organization to yield similar kind of results. They also correspond to the major work that is done in this thesis. This includes a system developed for data gathering from social sources of our study, an authorization and permission module which connects to the game tool, a data storage module which was can be constantly synced with events, furthermore, exploration of data sources were done as part of the analysis engine.

Our second contribution is the employee data set collected from his external and internal social media sources. Data is said to be the new currency in the business, with this work we put together a data-set rich in information about the employee and its network from different social sources, this enabled our exploration of employee’s expertise from multiple dimensions. This kind of data set can be used to further explore user/employee dimensions, not just related to expert finding but others like user-behaviour finding etc.

Third and the last major contribution of this work is the exploration study performed on the collected data sets and the evaluation to validate the findings of this exploration. This exploration study is conducted in a way that it can be extended to include more data sources.
6.2 Conclusions

In this section, we will discuss the conclusions of our work and how they relate to the research questions proposed in the beginning of the study.

To answer our main research question, "How the combination of social networks and enterprise social sources can result in better understanding of employees' expertise?". We identified several sub-research questions which are further discussed:

We answer our first two research questions, RQ1 and RQ2, which are related to the understanding of the field of expertise and expert finding through a literature study describe in Chapter 2. The prime learning from this chapter is understanding of the taxonomy of expert finding research field and its diverse implementations in existing systems.

The work done in Chapter 4 is a means to answer our research questions, RQ3, "Do the external social media sources duplicate the data already contained in enterprise sources?" and RQ4, "Which attributes of social media sources can help in improving expertise assessment?". We identified expertise dimensions, and assessed external and internal data sources across these dimensions. The results demonstrated that the external source, LinkedIn, had dissimilar properties to the IBM connections, indicating that addition of an external source can add diversity to the internal source. The network evaluation contributed to the reach dimension of expertise. This attribute also showed fairly dissimilar properties to the IBM Connections network, indicating an enrichment in terms of reachability/accessibility dimension of the employee expertise. More parameters of credibility dimension like, how well the person knows the topic, how he acquired that knowledge can be answered by LinkedIn's summary and job positions fields. Presently LinkedIn users can list out their projects and summary but these fields are not yet available through their API. These project fields can also add more value to the credibility dimension, further, hinting that external source LinkedIn enriches the data.

Results of Chapter 5, helped us in answering our final sub-research question: "To which extent can the combination of the enterprise and external data sources improve the expertise assessment in expert finding system?". The results discussed in chapter 5, show higher document retrievability measures for the system with combined sources. These results give an indication of average documents retrieved per query per system. This measure was higher in each case for the system with combined sources. Therefore, it is fine to say that there is a high percentage improvement in document retrievability with inclusion of an additional data source.
6.3 Future work

As described in Chapter 3, we have collected data from multiple social networks, but in this work have only utilized IBM connections and LinkedIn data sets for our analysis. One of the future goals would be to utilize an additional external data source like Facebook, this could be in terms of extending current work related to accessibility/reach and skill dimensions of an expert finding system.

In our work, we performed data preparation step, which included data cleansing of IBM connections profile data. This step enabled us to perform meaningful analysis on the data set. Although we employed sophisticated methods but were unable to reach significant improvement. One of the main reasons for this was, the presence of tag-terms, which were centric to IBM taxonomy and acronyms, also possibly related to projects and products used within IBM. These terms were not recognized by external Knowledge Bases like Wikipedia, and led to many tags being not standardized. An improvement suggestion would be to develop a dictionary/taxonomy for IBM specific terms, which presently does not exist. This can help in vocabulary development that can eventually lead to partially controlled IBM skills vocabulary similar to that of LinkedIn.

Another future requirement can be to evaluate the current Expertise system implemented within IBM with a standard test collection also obtained from the company and then include an external source in their system to validate it with real users. This can turn out to be a bottleneck obtaining these resources and support from the sources, due to the large and distributed geography of the company.


Appendix

**BU** Business Unit. vii, 35, 36, 39–41

**EF** Expert Finding. 10, 12, 26, 44

**ESM** Enterprise Social Media. 2

**PSM** Public Social Media. 2

**RQ** Research Question. 2, 20, 32, 34, 41, 42