

Finding Experts in Social Media Data using a Hybrid Approach

Seamus Brady - seamusbrady.ie

21st August 2015

Contents

1	Introduction	7
1.1	Objective	7
1.2	Structure of the Document	7
2	Background	8
2.1	The Need for Experts	8
2.1.1	Experts in the Wild	8
2.1.2	The Expert Finding Problem	8
2.1.3	The Evolution of Human Expertise	9
2.1.4	Expertise and the Global Corporation	9
2.1.5	What is Expertise?	9
2.2	Expert Finding Using Content Analysis	11
2.2.1	Introduction	11
2.2.2	Content Analysis and Adequate Knowledge	12
2.2.3	Content Analysis and Social Profile	12
2.2.4	Content Analysis and Expertise Credibility	12
2.2.5	Content Analysis and Effective Identification of Experts	13
2.2.6	Commercial Expert Finding Software Using Content Analysis	13
2.3	Expert Finding Using Social Graph Analysis	13
2.3.1	Introduction	13
2.3.2	Social Graph Analysis and Adequate Knowledge	14
2.3.3	Social Graph Analysis and Social Profile	14
2.3.4	Social Graph Analysis and Expertise Credibility	14
2.3.5	Social Graph Analysis and Effective Identification of Experts	14
2.4	Expert Finding Using Semantic Web Technology	15
2.4.1	Introduction	15
2.4.2	Semantic Web and Adequate Knowledge	16
2.4.3	Semantic Web and Social Profile	16
2.4.4	Semantic Web and Expertise Credibility	16
2.4.5	Semantic Web and Effective Identification of Experts	16
2.5	Summary	16
3	Strategy and Requirements	18
3.1	Definition of the Hybrid Approach	18
3.2	Primary Requirements	18
3.3	Secondary Requirements	18
4	Design Choices	20
4.1	Computer Programming Expertise as Domain for Expert Finding	20
4.2	Data Source Constraints	20
4.3	Web Scraping	20

4.4	APIs	21
4.4.1	Twitter API	21
4.4.2	GitHub API	22
4.4.3	DBPedia Linked Data Interface	22
4.5	Using Functional Programming Language - Clojure	23
4.6	Twitter Bootstrap	23
4.7	OpenNLP	23
4.8	General Design Approach	24
4.9	Summary	24
5	Detailed Design and Implementation	25
5.1	Overall System Workflow	25
5.2	ExpertQuest Namespaces	25
5.2.1	Twitter Client	25
5.2.2	GitHub Client	26
5.2.3	DBPedia Client	27
5.2.4	Text Analysis Component	27
5.2.5	Search Component	30
5.2.6	Web Interface	31
5.3	Mutable Elements	33
5.4	Expert Ranking	33
5.5	Challenges and Compromises	33
5.5.1	Matching Twitter Accounts with GitHub Accounts	34
5.5.2	Avoiding Non Programming Homonyms in the Search Results	34
6	Testing and Evaluation	35
6.1	Text Comparison Testing	35
6.2	Testing and API Limits	35
6.2.1	Test Runs	35
6.2.2	Defining an Expert for the Test Runs	36
6.3	Test Results	36
6.3.1	Precision	36
6.3.2	Recall	37
6.3.3	Cosine Similarity	38
6.3.4	Comparison of Different Programming Languages	38
6.4	Achievement of Requirements	40
6.4.1	Primary Requirements	40
6.4.2	Secondary Requirements	40
6.5	Summary	41
7	Conclusions and Future Work	42
7.1	SWOT Analysis	42
7.1.1	Strengths	42
7.1.2	Weaknesses	42

7.1.3	Opportunities	43
7.1.4	Threats	43
7.2	Future Work	43

List of Figures

1	ExpertQuest Workflow	26
2	Stemming and Feature Hashing Algorithms	29
3	Search Page	31
4	Searching Modal Popup	32
5	Search Results	32
6	Expert Ranking Code Extract	33
7	Average Precision	37
8	Average Recall	38
9	Average Cosine Similarity	39

List of Tables

1	The Three Properties of Expertise	11
2	Cosine Similarity Results for Example Strings	36
3	Test Runs	36
4	Programming Language Results from Test Run 3	39

Abstract

Several approaches to the problem of expert finding have emerged in computer science research. In this work, three of these approaches - content analysis, social graph analysis and the use of Semantic Web technologies are examined. An integrated set of system requirements is then developed that uses all three approaches in one hybrid approach.

To show the practicality of this hybrid approach, a usable prototype expert finding system called ExpertQuest is developed using a modern functional programming language (Clojure) to query social media data and Linked Data. This system is evaluated and discussed. Finally, a discussion and conclusions are presented which describe the benefits and shortcomings of the hybrid approach and the technologies used in this work.

1 Introduction

The original idea for this report arose from my frustration at endeavouring to find particular expertise in search engines and a growing realisation that expert finding was actually a much discussed and researched computer science topic. After some research, the initial aim was to build an expert finding system using Semantic Web technologies but the research suggested that this topic had already been covered in some depth. What was missing from the research was a practical examination of what sort of expert finding system could be developed with the data and tools as existing on the Internet today.

This report is an attempt to build this system, a practical, usable prototype expert finding system called ExpertQuest. This system will collect data from Twitter, GitHub and DBPedia to help users find computer programming experts.

1.1 Objective

The objective of this report is to see if it is possible to build and test a usable prototype expert finding system within the following constraints:

- Using a modern programming language.
- Using only data that is available on the Internet via social network APIs and Linked Data.
- Uses a mixed or hybrid approach to expert finding.
- Fulfil the requirements as outlined in Chapter 3.

1.2 Structure of the Document

- Chapter 2 discusses the need and value of expertise, as well as how expertise is defined and how expertise can be managed via technology.
- Chapter 3 outlines the particular strategy and requirements used in the system implemented for this report.
- Chapter 4 discusses various technology and design choices for implementing the expert finding system.
- Chapter 5 contains a detailed description of the actual design and implementation of the expert finding system.
- Chapter 6 outlines the testing and validation carried out on the results of the expert finding system.
- Chapter 7 outlines conclusions and some ideas for future work.

2 Background

This chapter will discuss the need for and value of expertise, as well as how expertise is defined and how expertise can be managed via technology.

2.1 The Need for Experts

2.1.1 Experts in the Wild

In 1960 the famous primatologist Dr. Jane Goodall observed that chimpanzees used specially chosen twigs to fish for termites [1]. This was the first time a scientist had observed this behaviour and it helped open new avenues of research into animal tool use.

Since then, tool use has been observed in seven classes of the animal kingdom, including primates and crows [2]. One of the most interesting things about the research into chimpanzee tool use is that some of the tools that chimps use are “culture-bound”. Only certain tools are known to a particular populations of chimps:

“Many differences can be found in tool behaviour for which neither genetic nor ecological explanations are suitable.” [3]

The use of tools is quite often taught by one generation to the next generation, so that in a sense these chimps have a local tool culture [4]. This also implies that certain individual chimps in a troop have more knowledge about certain tools than some other chimps. The use of tools allows these chimps to broaden their ability to adapt and thrive by opening up new sources of nutrition. These local chimp “experts” are helping to keep the troop alive.

This is much the same as our species case - human beings are also reliant on the skills and knowledge of multiple individuals for our survival.

2.1.2 The Expert Finding Problem

In a large organisation, perhaps even global, finding someone who is an expert may be a challenge - especially when compared to finding the local expert in a small neolithic community.

The problem can be called the *expert finding problem* or *expertise location problem*. We shall stick with the term *expert finding*.

As may be expected, there is a large body of software and research available on the expert finding problem and even the researchers do not agree on an exact definition.

This problem is defined by Lappas (2011) quite simply:

“... given a task at hand and a set of candidates, one wishes to to efficiently identify the right expert (or set of experts) that can perform the given task.” [5]

In their research, Yimam-Seid and Kobsa (2003), also found that expertise is sought out for two different reasons. They identified:

“... two main motives for seeking an expert, namely as *a source of information* and as *someone who can perform a given organizational or social function*.” [6]

These two definitions can be combined into the following single definition of the expert finding problem which will be used in this report: *Efficiently identify the right individual (or group) from a field of candidates that has the expertise to provide desired information or complete a desired task.*

This definition gives a viewpoint from which to survey the various strategies that have been used to solve the expert finding problem using software. For the sake of this report, it will be assumed that returning an individual (rather than a group) is sufficient.

2.1.3 The Evolution of Human Expertise

The story of human evolution is also wrapped up in the evolution of human expertise. Indeed, there is some evidence to suggest that our ability to think and communicate socially as human beings is intimately wrapped up in the evolution and increasing complexity of our tool use:

“Lower Palaeolithic technologies clearly do increase in hierarchical complexity through time, raising the possibility of important interactions with the evolution of human cognitive control and socially supported skill acquisition.” [7]

Human expertise has had a profound impact on our own species, even to the point that increasing human expertise in tooling has improved our cognitive and social abilities.

2.1.4 Expertise and the Global Corporation

As the global economy has become more complicated, expertise has become more and more valuable in the modern corporation. The tangible results of having experts working within your company include patents, copyrighted works, trade marks and other forms of what are known as *intellectual property*. In fact, the true value of intellectual property within a business may be worth more than all of the physical assets combined:

“During 2000, the market-to-book ratios of Fortune 500 companies increased to 6.3:1, suggesting that for every dollar of physical assets on the balance sheet, the market recognized \$6.30 worth of other intangible assets.

On average, in successful organizations, brands, intellectual property and the like are two to three times the value of physical assets. Intellectual property holdings are valuable corporate assets, and may make up a great portion of the total worth of an organization.” [8]

With this new level of value attached to expertise, and the importance of experts to business continuity (such as after a major disaster), there now exists an even greater desire to identify, manage and co-ordinate the expertise within an organisation.

2.1.5 What is Expertise?

So experts and expertise have had a profound impact on human beings and their closest primate relatives and experts are also important in the economy. But what exactly does the word “expert” mean?

The word “expert” generally means someone who knows much about a specific subject matter or discipline. Expertise is the noun describing this quality. However, the concept is more subtle than that.

The online Business Dictionary defines “expertise” in the following way:

“Basis of credibility of a person who is perceived to be knowledgeable in an area or topic due to his or her study, training, or experience in the subject matter.”
[9]

This definition offers a slightly more nuanced view of expertise:

- Expertise is a quality of an expert, a person that has studied, trained and gained experience in a particular subject matter. This person has more skills and knowledge than most others in a specific area.
- This person is perceived to have this quality by other people - expertise has a social dimension. A hermit can know everything about growing orchids, but without a society to communicate with, nobody will be able to consider the hermit as an expert (apart from the hermit of course, but this is perhaps tautological).
- It is also implied that this person has credibility or is trusted by others who perceive this expertise. Again, if our hermit suddenly appears in the village square claiming to be an expert orchid grower but refuses to supply any kind of bona fides, the villagers would probably be correct to assume that he or she is a crank.

So we can say that expertise has a knowledge dimension, a social dimension and a trust dimension. To be fully considered an expert, a candidate must have the necessary knowledge, have a group of peers aware of this knowledge and have a level of credibility amongst these peers that this knowledge is valid. These properties are summarised in **Table 1**.

These properties will provide a useful framework for the discussion of expert finding. We shall use these three properties to judge several approaches to expert finding in the following sections. Each of the expert finding approaches we are using can be compared using the following questions:

- (1) How does this approach measure ensure that the expert has adequate knowledge?
- (2) Does this approach measure the social profile of the expert in a meaningful way to show that the expert is regarded as an expert by his/her peers?
- (3) Does this approach validate the credibility expertise of the expert in any way?

There is some crossover between question 2 and 3 (as large groups of peers who regard a candidate as an expert does strengthen the candidate’s credibility). However they will be treated separately as there is enough divergence in other forms of expertise validation (education or research for instance) for the distinction to be useful.

To these three questions we can also add a fourth:

- (4) Does this approach efficiently identify the right individual from a field of candidates that has the expertise to provide desired information or complete a desired task?

Adequate Knowledge	A candidate must have the necessary knowledge
Peer Awareness / Social Profile	A candidate has a group of peers aware of their knowledge
Peer Credibility	A candidate is trusted by their peers

Table 1: The Three Properties of Expertise

The word “efficiently” does have some importance here, as speed and ease of use should be considered when comparing approaches.

2.2 Expert Finding Using Content Analysis

2.2.1 Introduction

Content analysis is the approach used by much of the expert finding academic work found in the Text REtrieval Conferences (TREC)[10]. In this work, content analysis is taken to be a synonym of textual analysis, and the phrases are used interchangeably.

This work was the mainstay of expert finding research until the rise of the social Web (which will be examined in the next section). In his survey of expert finding techniques, Lappas et al (2011) call this approach “Expert Location Without Graph Constraints”.

The content analysis usually amounted to using textual analysis techniques on large data corpora, for instance the email archives of any large organisation. In fact, the email list archives of the World Wide Web Consortium was often used as a basis for academic research as they were publicly available.

This approach concentrates on estimating

“the probability that a candidate... could be an expert with respect to a given topic query”[11]

In order for a data corpus to be analysed, it would need to be put through standard text preprocessing:

- Sanitising the data.
- Tokenising the remaining data.
- Stop word removal and stemming the words in the corpus.

After this, there are two possible models that can be pursued (Balog et al, 2006 [12]):

- A *candidate-centric* model where the various documents associated with a candidate are grouped and then a vocabulary of terms is extracted from the documents to represent the candidate’s expertise. The probability of the set of terms in the query being represented in each candidate’s model can be then generated using Bayesian analysis.
- A *document-centric* model is where the set of terms for each document are extracted and ranked based on the terms in the query. The candidates associated with each returned document are then found and an aggregated set of terms for all documents belonging to this candidate are generated. This set of terms is then analysed against the query to return the top candidate using Bayesian analysis.

Balog found that the document-centric model outperformed the candidate-centric model. One of the main reasons is that an index of candidates did not need to be kept up to date as the list of possible candidates could be generated dynamically from the list of documents returned from the initial query.

There are other variations on this approach but they mainly retain the same emphasis on modelling expertise based solely on the set of terms available in the text of the data corpus. One such variation was produced by Serdyukov and Hiemstra (2008) where terms associated with a candidate were augmented with sets of terms found in various online searches:

“We used various kinds of GlobalWeb search services to acquire a proof of expertness for each person which was initially pre-selected by an expert finding algorithm using only organizational data.” [13]

While this is an interesting avenue of research, it has the same emphasis on text based analysis.

2.2.2 Content Analysis and Adequate Knowledge

Content analysis measures the expertise of each candidate by modelling the expertise of candidates using standard, well understood information retrieval techniques. Various different types of content for example, emails, academic papers, business documents, can be combined into one system for analysis. Other types of text-based information such as Web based searches can be used to augment the system.

2.2.3 Content Analysis and Social Profile

The main disadvantage of content based expert finding is that it only addresses the first of the three properties of expertise as outlined in **Table 1**, *Adequate Knowledge*. This knowledge is captured in the set of terms associated with each candidate. As noted earlier, Lappas (2011) called this approach “Expert Location Without Graph Constraints” and for a good reason. There is very little analysis of the social profile of the expert. There is not much knowledge of where the expert appears in the social graph of his or her peers.

It can be difficult or impossible to gain this type of knowledge based solely on textual analysis of terms. Of course, this knowledge may not be sought (as when a user searches in a search engine, the popularity of an expert is secondary to the expertise being provided). However there are situations where the social profile of an expert may be useful, for instance, language or other cultural factors, when all other criteria are equal.

2.2.4 Content Analysis and Expertise Credibility

For much the same reasons as above, the actual credibility of the expertise of the candidate amongst his or her peers is difficult to assess also, based solely on textual analysis. What if the documents supplied have been falsified, for instance?

2.2.5 Content Analysis and Effective Identification of Experts

This approach can identify an expert from a field of candidates, once the caveats above are considered. However one disadvantage is that the data corpus may have to go through extensive extraction and transformation before it can be used, for example as in Balog et al, 2006 [14]. Once this ETL (Extract, Transform, Load) process is automated, the system may be efficient but there is an upfront development cost for this data munging.

2.2.6 Commercial Expert Finding Software Using Content Analysis

There have been many commercial expert finding systems available for purchase, normally by large corporations with extensive email and document corpora that could be mined for information. Some of these systems were reviewed by Maybury et al (2002) [15]. At the time of writing of Maybury et al's research (2002), most of these systems would have fitted into the content analysis grouping.

2.3 Expert Finding Using Social Graph Analysis

2.3.1 Introduction

Since the rise of social media, much of the academic research on expert finding has been concentrated on augmenting the results of textual analysis with an *analysis of the social graph* to discover if information about expertise can be gleaned from the links between people. Lappas et al (2011) call this approach "Expert Location with Score Propagation". They describe it as a two stage process:

"1) using language model or heuristic rules [to] compute an initial expertise score for each candidate... and 2) using graph-based ranking algorithms to propagate scores computed in the first step and rerank experts..." [16]

This research improves on the results of textual analysis - as Bozzon et al (2013) observed about their own research, it showed:

"the empirical demonstration of the greater contribution of activities of social network members with respect to their profiles for assessing the user expertise. We also found that certain profiles and activities of closest social contacts may provide useful information, thus giving a positive contribution to the expert ranking." [17]

There are several approaches that can be used to do this scoring and each normally revolves around describing the algorithm that is used.

A variation of the PageRank algorithm was proposed by Kardan et al (2011) [18]. They called their variation the SNPageRank Algorithm. The PageRank algorithm is famously used by Google and can be simply described as a link counting algorithm. The more links a website has, the greater the importance of the website. Kardan et al extended this idea to social networks, so rather than counting links, their algorithm counted posts, likes and connections between people in a social network (Friendfeed). The more messages a person had, the more important they were in the network. They extracted data from social networks

and analysed it using their modified algorithm. They found that they could achieve positive results by using this approach.

The Hyperlink-Induced Topic Search (HITS) Algorithm was developed by Jon Kleinberg for ask.com [19]. This was examined by Dom et al (2003) [20] as a possible expert finding algorithm. This algorithm is normally used to measure Web page “authority”. A Web page becomes an authority if it has many links from what are called “hubs”, or Web pages that links to a lot of other authority Web pages. They used the HITS algorithm on an email corpus. They considered those who received lots of email enquiries to be “authorities” and those who forwarded on queries to be “hubs”. In this particular instance, the algorithm was not as successful as others, possibly because email communication is not naturally structured in a strong hub versus authority way.

2.3.2 Social Graph Analysis and Adequate Knowledge

An initial phase of textual analysis generally provides a measure of the expertise of the candidate. This is used to generate profiles for each candidate based on the results of the textual analysis.

2.3.3 Social Graph Analysis and Social Profile

Social graph analysis can provide insight into the social profile of the expert, which may lead to further insights into their expertise (for instance, a particular age profile or profession may be associated with one expert and not another). The main impact of social graph analysis is that, depending on the algorithm, one can measure the number and strength of links to the expert, allowing comparison of link scores.

2.3.4 Social Graph Analysis and Expertise Credibility

The main advantage of using social graph analysis is that two of the main properties of expertise can be analysed. An extensive social profile can be a good indicator of expertise. However, an extensive social profile does not guarantee expertise per se. For instance, a mediocre expert may gain followers on social media due to good public relation skills, rather than actual expertise. Other factors may also need to be assessed for ensuring expertise credibility, for example peer reviewed science papers.

2.3.5 Social Graph Analysis and Effective Identification of Experts

The social graph can indeed provide information about experts. The ironic thing about link analysis and expert finding is that outside the world of computer science research, social media is already used to find expertise everyday:

“Social networking plays a significant role in expert finding. Often we require a more general expert to suggest the specialised expert we need by referring to colleagues in his or her social network. Social networking also helps to find an expert by providing a group of people within a community and perhaps links of people outside of the community as well.” [21]

As with the textual analysis approach, there may be an upfront cost in terms of data extraction and transformation e.g. Kardan et al (2011) [22]. Also, when compared to the textual analysis approach where a data corpus may be available on a local network, the distributed nature of the social graph may introduce its own complications in terms of network latency.

The success of link or social graph analysis may also depend on the choice of algorithm. As Dom et al (2003) found, certain algorithms may not fit the “shape” of certain data corpora and may not produce conclusive results. Also, privacy may become a major issue when using a social network data to find expertise.

2.4 Expert Finding Using Semantic Web Technology

2.4.1 Introduction

There is also a subset of expert finding research that uses *Semantic Web technology*. This approach overlaps with the content analysis and social graph analysis approaches but it is worthy of examination in its own right. Notably Lappas et al (2011) [23] do not mention any of this research in their excellent survey.

In the literature there are very few outlines of expert finding systems that actually use Semantic Web technology. One of the few is Li et al (2006) in which they describe a system called FindXpRT. This system was developed to use the FOAF ontology (Friend of a Friend) [24] with an extra series of rules implemented in RuleML [25], an XML based rule deduction markup language. The experts’ profiles in the system were implemented in FOAF as FOAF facts. Various rules in RuleML were then used to augment these facts - rules to allow clients to find an expert to collaborate with or rules that will allow an expert to pick another expert to collaborate with.

In their survey paper, Titus Schleyer et al (2008), mention a couple of challenges of using Semantic Web technology in an expert finding system, specifically for finding biomedical experts. Among these problems were the fact that Semantic Web technology does not meet all the requirements of building an expert finding system and it would likely be:

“a useful technological infrastructure for implementing expertise location systems, not as an end-to-end architecture” [26].

However Schleyer et al do recommend Semantic Web technology for keeping an expert’s profile up to date by linking it to their activities across the Web. They also suggest connecting disparate users and expertise domains in useful ways using specially adapted ontologies and using data aggregated by other Semantic Web technologies to augment existing data stores.

This use of Semantic Web as an augment to other approaches to expert finding is also echoed in Stankovic et al (2010). They specifically discuss Linked Data (LOD is defined as Linking Open Data):

“Traditional approaches tend to retrieve their data from closed or limited data corpuses. LOD on the other hand allows querying the whole Web like a huge database, thus surpassing the limits of closed data sets, and closed online communities. We believe that this opens new possibilities for traditional expert search and profiling systems which usually only rely on data from their local and limited

databases or on unstructured data gathered from the Web. LOD also stands up for a great promise to deliver multipurpose data that can be used to find experts in many domains and with many different expertise hypotheses." [27]

However, Stankovic et al also mention that there are challenges with merging duplicate data sets, verifying the authenticity of expertise and standardising how expertise data, for example science papers, are captured. All of these challenges arise out of the distributed nature of Linked Data.

2.4.2 Semantic Web and Adequate Knowledge

One of the main advantages that Semantic Web technologies offer is a way out of the information silos that expertise data tends to end up in. With the correct use of common ontologies and Linked Data, expertise from multiple disciplines could be merged and queried. Semantic Web technologies offer the possibility of using richer data sources to generate candidate profiles.

2.4.3 Semantic Web and Social Profile

Several Semantic Web technologies can be used to measure the social profile of a candidate, for example the FOAF protocol.

2.4.4 Semantic Web and Expertise Credibility

As was pointed out above, the possibility of rich connected data sources to generate candidate profiles could allow several different checks to be made against a candidate's bona fides, thereby increasing the credibility of each candidate profile.

2.4.5 Semantic Web and Effective Identification of Experts

Semantic Web technologies do not identify experts on their own. Semantic Web technologies show major potential for linking disparate areas of expertise for analysis and searching. However, as Titus Schleyer et al (2008) [28] point out, there needs to be a technological infrastructure in place to take advantage of these technologies and as yet, these types of developments are not common. So the potential of Semantic Web technologies in expert finding remains in the area of academic study, unlike for instance content analysis, where several commercial systems exist using the content analysis approach.

2.5 Summary

- Experts are a vital part of both human culture and the modern corporation.
- Expert finding can be defined as the problem of identifying the right individual from a field of candidates that have the expertise to provide desired information or complete a desired task.
- Expert finding is a challenging but necessary problem in managing intellectual property.

- Expertise can be said to have a knowledge dimension, a social dimension and a credibility dimension.
- The content analysis approach offers great strengths in providing the knowledge dimension of a candidate.
- The social analysis approach can be used to measure the social profile of a candidate.
- Semantic Web technologies can be used to augment the other two approaches, including increasing the credibility of a candidate by querying multiple, rich data sources.
- No single approach provides a complete solution to the expert finding problem.

3 Strategy and Requirements

The rest of this document will outline the design, development and testing of a prototype expert finding system that will be called *ExpertQuest*. This chapter outlines the particular strategy and requirements used in the system implemented for this report.

3.1 Definition of the Hybrid Approach

ExpertQuest is inspired by, rather than based on, the approach by Metze et al (2007) when they designed the Spree System. They describe the basic system as:

“a system to facilitate exchange of information by automatically finding experts... [our] objective is to provide an online tool, which enables individuals within a potentially large organization to search for experts in a certain area, which may not be represented in company organization or reporting lines” [29]

The actual Spree system [30] consists of a Python Web application that uses ontological data derived directly from the Web (Yahoo Search API) in conjunction with an algorithm that uses this data to classify experts. The website then uses a community of users to further classify expertise as they interact with each other. This is what is meant by the *hybrid approach*. Spree uses all three of the approaches that were outlined in the previous chapter in one system - content analysis, social graph analysis and Semantic Web technology. This gives us a design approach that can be used for the ExpertQuest system. The system should be Web based and use all three approaches to expert finding together, in order to fulfil the primary and secondary requirements outlined below.

3.2 Primary Requirements

As part of this development, the ExpertQuest system will attempt to fulfil the key requirements that Maybury et al [31] outline for an expert finding system:

1. The system should be able to identify experts from a field of candidates.
2. The system should be able to classify the level of expertise of each candidate.
3. The system should be able to validate the expertise of each candidate.
4. The system should be able to rank candidates on multiple dimensions.

These will be regarded as the primary requirements to evaluate the system.

3.3 Secondary Requirements

Alongside these core requirements, ExpertQuest should also meet the following secondary requirements:

1. The system should use the hybrid approach as defined above and measure in some meaningful way, all three dimensions of expertise as outlined in the last chapter - the knowledge dimension, social dimension and credibility dimension.

2. The system should be based on real-time data available on the Internet as much as possible. This will help to avoid the problems associated with stale expert profiles. Any information on an expert should be up to date. The system should also minimise any need for expensive extraction and transformation of data such as outlined by Kardan et al (2011). There will be no delays for batch processing data or anything similar, at least in the prototype.
3. The system should also allow a way to contact the expert if possible.
4. The system should be as real-world as possible, not just a theoretical outline - basically a usable prototype.
5. If possible, the system should be user friendly and reasonably fast and should have a Web interface.

These secondary requirements will also be used to evaluate the system. The design choices and constraints arising from these requirements will be discussed in the next chapter.

4 Design Choices

This chapter discusses various technology and design choices for implementing ExpertQuest.

4.1 Computer Programming Expertise as Domain for Expert Finding

Many areas of expertise were considered as the domain for ExpertQuest and even making ExpertQuest handle multiple domains. However, the multiple domain idea was quickly dismissed as too complicated for this document. Moreover, several domains stood out in terms of available data, and the following domains were isolated as possibilities:

- Biotechnology
- Medical and genetic research
- Computer Science

Each of these would be an excellent candidate as a topic, but in the end the the author's experience in computer programming made this the obvious choice as a domain of expertise. A knowledge of the technical jargon in the field and a rich availability of data sources made this an easy choice. It was also decided that confining the system to one subset of a domain (computer programming versus all of computer science) would make the system easier to build and evaluate in the time available.

However, there were other factors in making computer programming an obvious choice when the data sources available online were reviewed.

4.2 Data Source Constraints

Because ExpertQuest is required to work with real-time data as much as possible and avoid data extraction, this implies that the data that will be required for finding experts will most likely be via one of two methods:

- Scraping data (text extraction) from websites.
- Application Programming Interfaces (API) that are exposed by various organisations on the Internet.

There are others (purchasing formatted data for instance) but they remain outside the scope of this document.

4.3 Web Scraping

This choice was considered as an excellent possibility for extracting data. For instance, *Google Scholar* contains hundreds of academic papers and citations. It is an excellent resource for finding experts in multiple fields, but particularly in the medical and biotechnology fields. There even exists a Python library for extracting data via Web scraping from Google Scholar [32].

However, on further examination, this type of Web scraping breaks the terms of service of Google Scholar (and multiple other websites of this type) and the legality of this activity, even for research, became questionable. Therefore the idea was abandoned.

Unfortunately Google Scholar does not expose an API, so it was not an option as a form of data for ExpertQuest. Alongside this decision, the domains available for ExpertQuest were diminished as there are not many publicly accessible data sources for academic papers as reliable as Google Scholar. Other vendors in this area were contacted, but all they provided were multi-gigabyte data dumps of data that are several years old.

4.4 APIs

Full lists of what APIs are available on websites like the Programmable Web [33]. On this website alone there are 13,000+ APIs listed for everything from 3D to zip codes. However, these APIs break down into several main categories:

- APIs provided by commercial companies at a cost.
- APIs provided by commercial companies available for free but with limits on usage.
 - A subset of this category are social media APIs.
- APIs provided by non-commercial or academic organisations available for free.
 - A subset of this category includes most of the Linked Data APIs.

Unfortunately, the availability of API access is also quite constrained as many of the commercial organisations who were offering free API access have removed this access as part of the drive to control their data as “big data” which is now commercially valuable. *LinkedIn* is one such business. This service is a professional social network (“Facebook for work”) and would be an excellent source of expertise data. Many professionals effectively use it as an online curriculum vitae. However, LinkedIn have closed their API access to all but a couple of partner businesses [34].

After reviewing available API data sources and deciding on computer programming as a good choice for the domain of ExpertQuest, there were three APIs available that seemed to be useful:

- *Twitter API*
- *GitHub API*
- *DBPedia Linked Data Interface*

4.4.1 Twitter API

Twitter has an extensive API [35] and as Bozzon et al note, it is an excellent choice for searching for experts:

“Twitter appears [to be] the most effective social network for expertise matching, as it very frequently outperforms all other social networks (either combined or alone)... Twitter appears as well very effective for matching expertise in domains such as computer engineering, science, sport, and technology & games.”[36]

Twitter offers easy access to data that can be mined for experts and the API is free, so it is ideal as one of the central data sources for ExpertQuest.

4.4.2 GitHub API

GitHub is the most popular open source code sharing website on the Internet today and also offers an extensive free API to interact with the site [37]. As was outlined above, computer programming is the domain of expertise that ExpertQuest will cover. GitHub makes an obvious choice as a data source - the site operates almost as a social media site for developers in many different languages, so the data stream available in the API is rich and extensive.

4.4.3 DBPedia Linked Data Interface

Linked Data is a Semantic Web technology and is an important part of the hybrid approach as was defined earlier. Linked Data is defined as a way to use the Web:

“to create typed links between data from different sources... ...Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to form external data sets.” [38]

Linked Data is important because it allows the meaning of the data to be explicitly defined so it can be used as reference material by systems such as ExpertQuest. A system can extract meaningful content quite easily from the machine readable Linked Data.

Unfortunately after much research, it appears that much of the Linked Data [39] is not suitable for ExpertQuest as it falls under the following categories:

- Data is available on the Web but the server is not reliable enough to base a system on (for instance, SPARQL end points can disappear and reappear).
- The data that is made available is in the format of large data dumps of RDF triples. This is very useful for most research, but is explicitly disallowed for ExpertQuest as there will be no complex extracting or processing of data in this prototype.
- Much of the data was not in the domain of computer science.

The best example of Linked Data available on the Web is also the most useful for ExpertQuest. *DBPedia* provides a Linked Data interface to the data that is contained in Wikipedia.

DBPedia provide an API that can be queried in several ways and the JSON (JavaScript Object Notation) enabled DBpedia Linked Data Interface [40] was chosen as the best way to get interesting data into ExpertQuest via the web. The server connection to DBPedia is relatively reliable and the interface is well documented.

4.5 Using Functional Programming Language - Clojure

Several different programming languages were examined for implementing ExpertQuest, but in the end the decision was made to use the Clojure programming language [41]. Clojure is a modern functional programming implementation of Lisp that is written to run on the Java Virtual Machine (JVM). Clojure offers the following advantages for writing an expert finding system like ExpertQuest:

- As it runs on the JVM, it is stable and performs well on multiple operating systems. All of the extensive Java libraries are also available to use in Clojure.
- As Clojure is a Lisp with the “data as code” philosophy that this entails, processing data is extremely easy. Clojure comes with a rich standard library that makes programming data analysis software very straightforward. It is also a small language and easy to learn as a result.
- Clojure is a modern functional language with data structures that are immutable by default. This allows the software developer to avoid many bugs associated with managing state in applications. However, the language is designed for practical software development and also comes with an assortment of mutable data types that can be used as needed.
- The tooling around Clojure is excellent and it comes with support for several text editors, integrated development environments (IDEs) and build/deployment tools.

The various Clojure libraries that were used will be discussed in the next chapter.

4.6 Twitter Bootstrap

It was originally envisaged that a complex JavaScript based user interface would have been desirable for ExpertQuest. There are many programming languages that compile to JavaScript, including an implementation of Clojure called ClojureScript, that were considered for this purpose.

After some research it was concluded that this would complicate the project beyond the needs of this document. For this prototype a simple clean and usable Web interface can be produced using the excellent *Twitter Bootstrap* library [42]. This is a library of HTML and Javascript components that can be used to build a Web interface quickly and easily without the major time sink which would be involved in writing a user interface from scratch.

4.7 OpenNLP

For the text analysis portion of the ExpertQuest system, the choice of libraries was obvious. The *Apache OpenNLP* [43] is the one of the best machine learning/natural language processing toolkits available and it is available on the JVM. As Clojure is also hosted on the JVM, OpenNLP is available to be used in ExpertQuest without any difficulty.

4.8 General Design Approach

The general design approach for ExpertQuest will be the following:

- Initially identify candidates via Twitter.
- Use Linked Data and GitHub data to verify and rank the candidate experts.

4.9 Summary

- ExpertQuest will use social media data from the Twitter and GitHub API for finding experts.
- ExpertQuest will also extract content from the Linked Data interface to DBPedia.
- ExpertQuest will be written in Clojure, with a Web based interface using the Twitter Bootstrap library. The OpenNLP libraries will be used for text processing.

The next chapter will go into more detail as to how ExpertQuest is implemented.

5 Detailed Design and Implementation

This chapter contains a detailed description of the design and implementation of ExpertQuest.

5.1 Overall System Workflow

Figure 1 exhibits a diagram for the overall system workflow for ExpertQuest. The system goes through the following steps to find experts for a user:

1. The user chooses a search term in the Web interface. In the prototype, the user selects a programming language from a drop down list in the Web interface. The languages presented are the top 50+ languages from the Tiobe programming language popularity index [44].
2. The search term is passed to the Twitter Search API.
3. The Twitter Search API returns a number of search results.
 - a) All Twitter usernames in the results are collated and are passed on to the GitHub API to ascertain if there are matching usernames on GitHub.
 - b) All matching GitHub accounts are queried against the GitHub API.
4. All Twitter accounts with matching GitHub accounts are queried using the Twitter API to obtain a number of most recent tweets.
5. The abstract for the programming language being queried is loaded from DBPedia via the Linked Data API.
6. The Tweets for each Twitter account are concatenated into a string and analysed against the DBPedia abstract using a feature hashing algorithm.
7. All of the results are collated and the expert candidates are ranked according to several criteria as outlined below.

5.2 ExpertQuest Namespaces

Code is abstracted into namespaces in Clojure, which are generally confined to individual files, similar to Java packaging, each containing a number of functions. The components mentioned below are each implemented as individual namespaces in ExpertQuest.

5.2.1 Twitter Client

A simple client for Twitter was written using the recommended Clojure Twitter API library [45]. There are multiple API calls exposed by the Twitter API. Only two of these APIs are used in ExpertQuest:

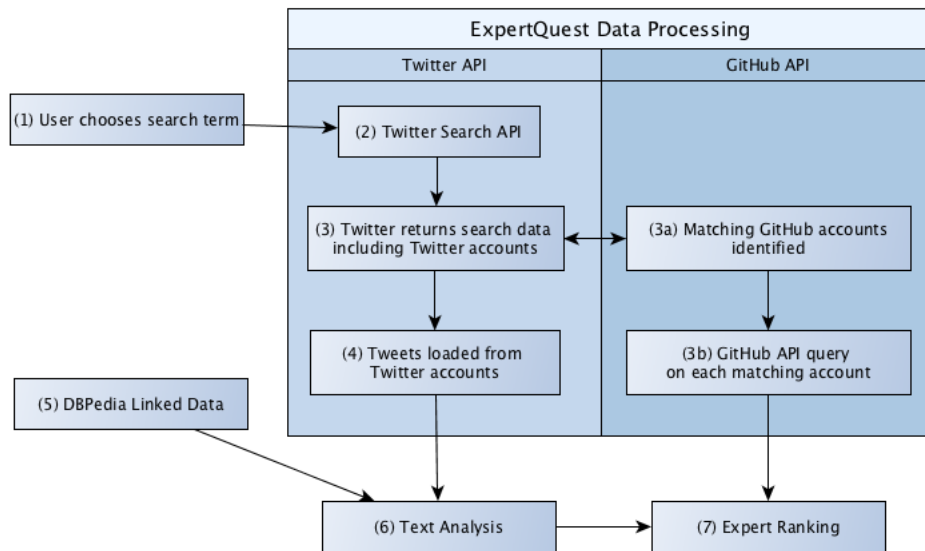


Figure 1: ExpertQuest Workflow

- The *Twitter Search API* which will return a number of search results for a specific search term. This is used in step 2 above. The search term used is the programming language plus the string “github”. This is used to disambiguate searches where the programming language name has also an everyday meaning e.g. Python or Ruby.
- The *Twitter User Timeline API* which will return a specified number of most recent tweets for a specified username. This is used in step 4 above.

The Twitter client component wraps all tweet data into a Clojure data structure and returns it to the search component. The values returned by the component includes the Twitter user’s name, their account name and the number of followers the user has on Twitter.

5.2.2 GitHub Client

A wrapper for the GitHub API was also created using the best available Clojure GitHub API library [46]. There are again multiple individual APIs exposed by the GitHub API but there are only two needed by ExpertQuest:

- A call to the GitHub User API to extract all information pertaining to one user account. This is used in Step (3a) and (3b) above.
- A call to the GitHub User Repository API which returns all repositories belonged to the user. This is used in Step (3b) above and is used to sum the total bytes of code that are contained in a user’s repositories for a specific language.

The GitHub client component wraps all returned GitHub data into a Clojure data structure and returns it to the search component. The values returned by the component include the number of followers the user has on GitHub, and the sum total number of bytes of code in a specific language that the user has in their Git repositories.

5.2.3 DBPedia Client

The Linked Data API of DBPedia provides a REST (Representational State Transfer) interface that will return JSON data based on a query to a specific URI. ExpertQuest uses the Clojure HTTP client and JSON libraries to query this interface for a provided query. The resulting data is automatically converted to a Clojure data structure and the abstract (text data) pertaining to the search parameter is extracted.

A call to extract disambiguation data from DBPedia was also created, but in the end was not necessary for the prototype.

This component is used in step 5.

5.2.4 Text Analysis Component

The text analysis component of ExpertQuest is used to measure how much the expert candidate has been tweeting about the specific programming language. This component uses the Clojure wrapper library for the OpenNLP library for this purpose [47]. In order to get a measurement of how much an expert candidate has been tweeting about a specific language, the following steps are taken:

1. The tweets for one specific user that were collected from the Twitter API in Step 4 are concatenated into one long string and passed to the text analysis component.
2. The abstract text data from DBPedia (Step 5), about the specific programming language, is passed to the text analysis component. This text is generally a few hundred words long and is loaded with descriptive phrases and terminology pertaining to the programming language. For instance, here is the Clojure entry:

“Clojure (pronounced like "closure") is a dialect of the Lisp programming language created by Rich Hickey. Clojure is a functional general-purpose language, and runs on the Java Virtual Machine, Common Language Runtime, and JavaScript engines. Like other Lisps, Clojure treats code as data and has a sophisticated macro system. Clojure's focus on programming with immutable values and explicit progression-of-time constructs are intended to facilitate the development of more robust programs, particularly multithreaded ones.”

3. Both pieces of string data are transformed into vectors using the feature hashing algorithm described below and are compared using cosine similarity to give a value between 0 and 1, where 1 is more similar. This value is passed back to be used as part of the expert ranking.

Feature Hashing Algorithm: It was originally envisaged that a simple comparison such as the Sorensen / Dice coefficient or the Jaro-Winkler distance would have been sufficient for the text comparison task in ExpertQuest. However these algorithms are generally used on much shorter strings than those ExpertQuest is comparing. Further research was carried out to determine a fast way to compare text documents.

In the field of machine learning and natural language processing, input data is normally transformed into a *feature vector* which is an array of integer values where each value represents an item of a specific type or a measurement of a specific property (a “feature”). A standard procedure to vectorise a document is to extract and count the occurrence of each word. This vector is generally mapped to a document dictionary which is represented as an associative array where the actual word can be looked up. ExpertQuest needs to be reasonably fast, and converting text documents into a vector form for comparison can be slow and use a large amount of RAM [48].

Fortunately there is a technique called *feature hashing* available that will work perfectly for ExpertQuest. Feature hashing is a machine learning technique (also known as the “hash trick”) that bypasses the associative array element of the feature vector. The Wikipedia article on feature hashing explains this quite succinctly:

“[feature hashing] is a fast and space-efficient way of vectorizing features, i.e. turning arbitrary features into indices in a vector or matrix. It works by applying a hash function to the features and using their hash values as indices directly, rather than looking the indices up in an associative array.” [49]

So rather than store the word in an associative array, we hash the word and use that resulting value as the index for the word in the feature vector. This is updated in the vector as appropriate. As Foreman and Kirshenbaum observed of their implementation of this technique “SpeedyFX”, the results are compelling and accurate when compared to the standard method:

“We have shown that using SpeedyFx integer hashes in place of actual words is faster, requires less memory for transmission and use of multiple classifiers, and has an effect on classification performance that is practically noise compared to the effect of other common parameters in model selection.” [50]

And again in Weinberger et al, feature hashing is praised for its benefits:

“The benefits of the hashing-trick leads to applications in almost all areas of machine learning and beyond.” [51]

There is no native implementation of the feature hashing technique in Clojure, so an implementation was written. This implementation works as follows:

1. The incoming string parameter is passed to a function which removes punctuation, filters out everything but nouns and verbs and stems the remaining words (using the Clojure OpenNLP library).
2. Creates a vector of the remaining words and hashes each one using a Clojure hashing function.

```

1 (defn conj-stemmed-nouns-verbs
2   "Conj sentences by nouns and verbs, stemming them at the same time"
3   [input-text]
4   (->> input-text
5     (clean-string)
6     (nouns-verbs-filter)
7     (map #(porter-stem-text (first %1)))
8     (string/join " ")))
9
10
11 (defn get-fvec-for-text
12   "Transforms a string into a feature hash vector"
13   [input-text]
14   (clear-fvec-atom)
15   (let [stemmed-text (conj-stemmed-nouns-verbs input-text)]
16     (let [word-vec (string/split stemmed-text #" ")]
17       (doseq [x (range 0 (count word-vec)) :let [word (get word-vec x)]]
18         (let [index (mod (digest/crc32 word) fvec-size)]
19           (let [old-value (get @fvec-atom index)]
20             (swap! fvec-atom assoc index (inc old-value))))
21         (identity @fvec-atom))))))

```

Figure 2: Stemming and Feature Hashing Algorithms

3. These hash values are stored in an atom (a mutable Clojure data type) that is updated with each word in turn.
4. The finished vector is passed back as a return.

The actual implementation is displayed in **Figure 2** above.

The algorithm performs well enough to be usable within a Web interface. For instance, the text of the novel *Anne of Green Gables* by Lucy Maud Montgomery, downloaded from Project Gutenberg, contains about 105,000 words. This was vectorised by the algorithm on a MacBook Pro with 8gb of RAM in under two minutes. The resulting vector looks like this:

```

1442 367 386 50 157 546 193 76 81 379 68 178 103 211 324 153 210 243 38
99 109 66 197 41 130 148 131 262 89 214 54 89 165 210 226 30 153 93 72 51
26 127 123 95 82 949 331 93 109 142 55 41 107 106 37 195 106 72 57 73 59
44 210 95 656 280 46 92 128 425 132 156 53 153 35 67 55 68 141 132 93 519
72 49 62 111 278 49 101 74 103 74 121 124 17 124 95 102 105 113 83 95 116
245 104 165 54 21 224 37 216 122 357 103 297 1336 68 71 215 83 91 34 100
292 42 82 55 86 112 50 118 345 116 589 264 222 23 76 27 58 24 155 89 263
38 114 172 57 139 179 181 311 67 218 60 181 287 33 154 103 71 218 221 900
195 87 117 62 468 149 109 96 116 20 147 244 133 177 88 158 60 47 83 62 95
32 146 329 22 30 60 239 127 31 101 320 193 50 85 67 164 99 170 126 173 215
57 135 353 277 103 131 202 70 680 114 237 526 163 216 168 304 234 104 298
125 298 121 64 82 156 144 59 457 115 89 37 43 54 118 142 75 27 126 329 72
74 97 129 63 149 128 78 99 366 207

```

As most of the text documents will be much smaller than this within ExpertQuest, this level of performance was deemed sufficient for the prototype.

Cosine Similarity The resulting vector for each text document (either a DBpedia abstract or a string of concatenated tweets) need to be compared for similarity. The cosine similarity was chosen as it is generally quite fast over sparse vectors. The actual implementation of the cosine similarity algorithm was taken from the Clojure Incanter statistical package. The cosine similarity value is returned back to be used in the ranking of the expert candidates.

5.2.5 Search Component

The search component contains the main search function that is called when the user submits the form in the Web interface as described below. This function calls the various components in order as described in the workflow as set out in **Figure 1**.

Search Algorithm Constants Two constants are defined in the ExpertQuest search component. These values behave how the search algorithm behaves:

- *Twitter Search Count*: This is the amount of search items to load from Twitter. This defaults to 50.
- *User Timeline Count*: This is the amount of tweets to load from a user's Twitter timeline for text comparison. This defaults to 25.

This means that for each of the 50 search results that returns a Twitter account that has a matching GitHub account, 25 tweets will be loaded from this account for the text analysis. This means that the search algorithm in ExpertQuest performs at $O(n^2)$.

ExpertQuest Helping you find coding experts on the web

[Home](#) | [Search for Programming Experts](#) | [About](#)

Search for coding experts in the Tiobe Top 50+ programming languages...

Search for experts on

© Corvideon 2015

Figure 3: Search Page

5.2.6 Web Interface

The ExpertQuest Web interface is developed using the following Clojure libraries:

- The Ring library which provides clean abstracted access to a Web server from within the application [52].
- The Compojure library which is used to route URLs internally in the application [53].
- The Hiccup library which is used to output HTML to the client [54].

Twitter Bootstrap is used to provide clean cross-browser typography and layout (albeit a simple one).

The Web interface is illustrated in the following figures:

- **Figure 3**, above, exhibits the search drop down that presents a list of programming languages. The user can search for programming experts by selecting a language and clicking the search button. The data for this drop down list is loaded from a configuration file as described later in this chapter.
- **Figure 4**, overleaf, shows the Twitter Bootstrap modal popup that the users see when the system is searching. It was decided that rather than using a more complex asynchronous callback based system, this user interface was sufficient for the prototype despite its inelegance.
- **Figure 5**, overleaf, illustrates the search results as displayed once the search has returned with results. The results are ranked according to the criteria described later in this chapter. One thing to note is that the cosine similarity value is listed as the *Twitter Mentions* column to give the user some indication of how much this expert candidate has been tweeting about the specific programming language. This value is displayed as a progress bar, rather than as a number, for increased visual feedback.

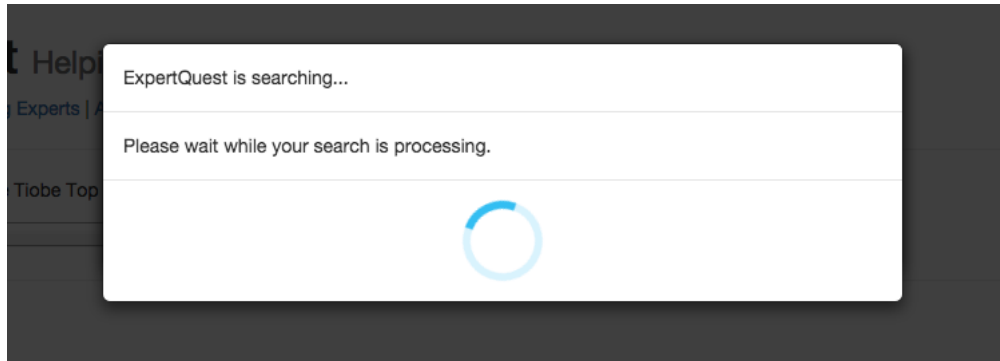


Figure 4: Searching Modal Popup

ExpertQuest Helping you find coding experts on the web
[Home](#) | [Search for Programming Experts](#) | [About](#)

ExpertQuest found 10 expert candidates for Clojure.





Name	Accounts	Twitter Mentions	Bytes of Code	GitHub Followers	Twitter Followers
'[[[hugs true]]]	http://www.twitter.com/otfrom http://www.github.com/otfrom		65431	85	2663
Matt Cottingham	http://www.twitter.com/mattro http://www.github.com/mattro		61059	30	851
Isaac Johnston	http://www.twitter.com/superstructor http://www.github.com/superstructor		49019	27	223
Shayan Mohanty	http://www.twitter.com/shayanjm http://www.github.com/shayanjm		227	15	17607

Figure 5: Search Results


```
1 (sort-by
2   (juxt :doc :github-followers :cosim :twitter-followers))
```

Figure 6: Expert Ranking Code Extract

5.3 Mutable Elements

As Clojure is a functional programming language, the emphasis is on minimising mutable state within an application. It is interesting to note that all the areas where state can be mutated can be listed quite simply for ExpertQuest as a result (compared to say, an application written in typical Java, which would have mutable state spread right through each class).

The mutable state in ExpertQuest consists of the following

- A Clojure atom that is used when processing the data within the feature hashing algorithm.
- Various Twitter and GitHub security configuration files in the Clojure EDN format (Extensible Data Notation, basically Clojure data structures stored as strings in files).
- An EDN file containing the list of the *Time Top 50+ programming languages* [55] with the corresponding correct search string for DBPedia. This was necessary as carrying out a search for “Java” in DBPedia returns the island nation rather than the programming language. The programming language is listed as “Java (programming language)”. The author developed code to return disambiguation pages from DBPedia, but this made the Web interface very slow and the idea was left to one side for the prototype.

5.4 Expert Ranking

The expert ranking within ExpertQuest is achieved quite simply thanks to the rich data handling ability of Clojure. A vector of maps is passed back to the Web view from the search function and before it is displayed it is sorted by the following criteria:

- *Number of Bytes* - The number of bytes of code that the expert candidate has in their GitHub repositories in the specific programming language being searched for.
- *GitHub Followers* - The number of GitHub followers they have watching their GitHub repositories.
- *Twitter Mentions* - The cosine similarity of their most recent tweets when compared with the DBPedia abstract on the specific programming language being searched for.
- *Twitter Followers* - The number of Twitter followers they have in their Twitter account.

This code is extracted and displayed in **Figure 6** above.

5.5 Challenges and Compromises

This section is a discussion of some the challenges that were encountered during developing ExpertQuest.

5.5.1 Matching Twitter Accounts with GitHub Accounts

The system only presents results to the user where the Twitter and the GitHub name are exactly the same. This is a compromise for the prototype as it obviously ignores results where the same person has two different account names on each service.

An early attempt was made to cross reference Twitter and GitHub accounts using the Yahoo Search API. However, there was no easy way to verify that any results returned were about the same person - this would be a project in itself.

The compromise within ExpertQuest is a reasonable one as each expert candidate is checked to make sure they were tweeting about a specific programming language and also that they have code in repositories in this language.

5.5.2 Avoiding Non Programming Homonyms in the Search Results

Ruby, Python, Java and several other languages are all homonyms. Avoiding these non programming homonyms in the Twitter Search is a real issue, so a simple compromise was to also use the word "github" in the search query which helped narrow the results. This is not very elegant, but it works reasonably well to avoid the homonyms.

In the next chapter, the ExpertQuest system is evaluated against the primary and secondary requirements.

6 Testing and Evaluation

This chapter outlines the testing and validation carried out on the results of ExpertQuest.

6.1 Text Comparison Testing

The feature hashing / cosine similarity algorithm was tested for accuracy using some basic tests on the Clojure REPL as illustrated in **Table 2** overleaf. In each case, string A and B were passed through the feature hashing algorithm and the resulting vectors were passed through the cosine similarity algorithm. The higher the score, the more alike the strings, with the highest possible value being 1.

The scores were predictable and accurate as each test is described below:

1. This test shows that the algorithm only compares on nouns and verbs as designed. These strings are equal as the noun “dog” is all that is being measured.
2. There are three out of four verbs that match in this test with an expected value of 0.75.
3. The only difference in these strings is the extra verb “barking”, so the score is less than one.
4. String A has two nouns and string B has one noun and one verb, hence the 0.50 score.
5. The difference here is one word out of three (“mouse ran clock” versus “elephant ran clock”) so the score is close to two thirds similar.
6. String A and B are here seen as the same, as word order is irrelevant.

These simple tests show that the text comparison is reasonably accurate. ExpertQuest only requires reasonable accuracy as the cosine similarity score is only one factor amongst others for expert ranking (as the “Twitter Mentions” measure in the Web interface).

6.2 Testing and API Limits

6.2.1 Test Runs

Three different tests were run to measure the performance of ExpertQuest. The tests took the following format:

- A data dumper was written that rotated through all 53 programming languages in the system (see Appendix 1 on page 45) and ran a search against Twitter and GitHub. The data from these sessions was recorded and analysed.
- Each test only differed from the other in the values of the search component constants as mentioned earlier. Column A in **Table 3** overleaf is the amount of search items to load from Twitter. Column B in **Table 3** overleaf is the amount of tweets to load from a user’s Twitter timeline for text comparison.

Test No.	String A	String B	Cosine Similarity
1	"white dog"	"black dog"	1.00
2	"run jump play hide"	"run jump play seek"	0.75
3	"running dog"	"running and barking dog"	0.82
4	"runner jump"	"running jump"	0.50
5	"the mouse ran up the clock"	"the elephant ran over the clock"	0.67
6	"the mouse ran up the clock"	"the clock ran over the mouse"	1.00

Table 2: Cosine Similarity Results for Example Strings

Test Run No.	A - Twitter Search Count	B - User Timeline Count
Test Run 1	10	5
Test Run 2	30	15
Test Run 3	50	25

Table 3: Test Runs

- The upper and lower values for these tests were discovered by trial and error. The lower value represents the approximate threshold where some useful values are returned. The upper value represents the approximate point at which the Twitter and GitHub API limits get hit by the system. Beyond this point search queries may start failing.
- The values used in each test run are displayed in **Table 3** above.

6.2.2 Defining an Expert for the Test Runs

ExpertQuest will display all of the data that it returns but will rank all candidates according to the criteria outlined in section 5.4. Some candidates will have Twitter and GitHub accounts and have been actively tweeting about the specific programming language in question. These candidates may be of interest to the user. However, for the sake of the test runs, only candidates who also have GitHub repositories containing code in the programming language in question are considered "full" experts. These candidates are counted as correct returns in the tests as they would be of the most interest to the user.

6.3 Test Results

Please note, more detailed test results are contained in the appendices on pages 45-49.

6.3.1 Precision

The precision of each test was calculated using the number of experts found divided by the total number of candidates found. As each search for a programming language is independent, the average precision for each programming language was worked out across all 53 searches. As expected, when more data was pulled into the system using the larger search constants,

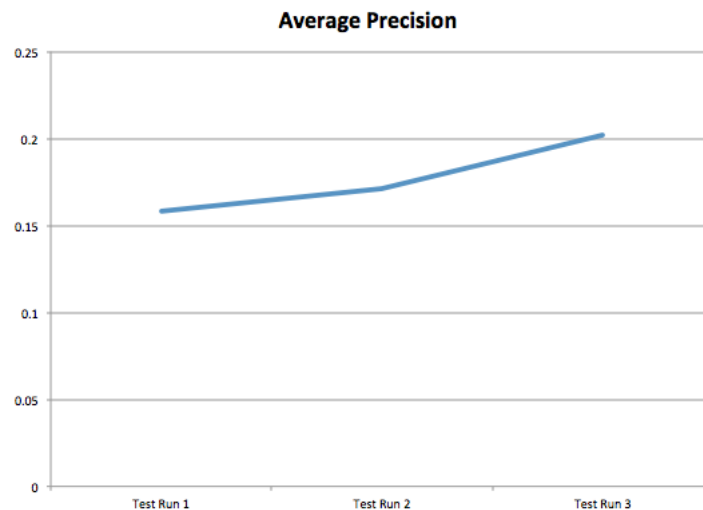


Figure 7: Average Precision

the precision went up steadily. The average precision for each test run was as follows (also see **Figure 7** above):

1. 0.158265948
2. 0.171069182
3. 0.20215256

6.3.2 Recall

Each of the test runs passed in a search constant representing the number of Twitter search results to return (10, 30 or 50). Under ideal conditions, each of these search results would have contained the details of an expert. So the maximum number of experts that can ever be returned is equal to the number of Twitter search results.

Using this information we can calculate the recall of each test by dividing the total number of experts found by the Twitter search constant. As illustrated above, the results of each test were calculated and then averaged across all 53 programming languages. The average recall for each test run was as follows (also see **Figure 8**, overleaf):

1. 0.052830189
2. 0.020754717
3. 0.050943396

The recall dips in the second test run even though the search constants were increased, the number of experts returned was not more much than in first test. The third test was much

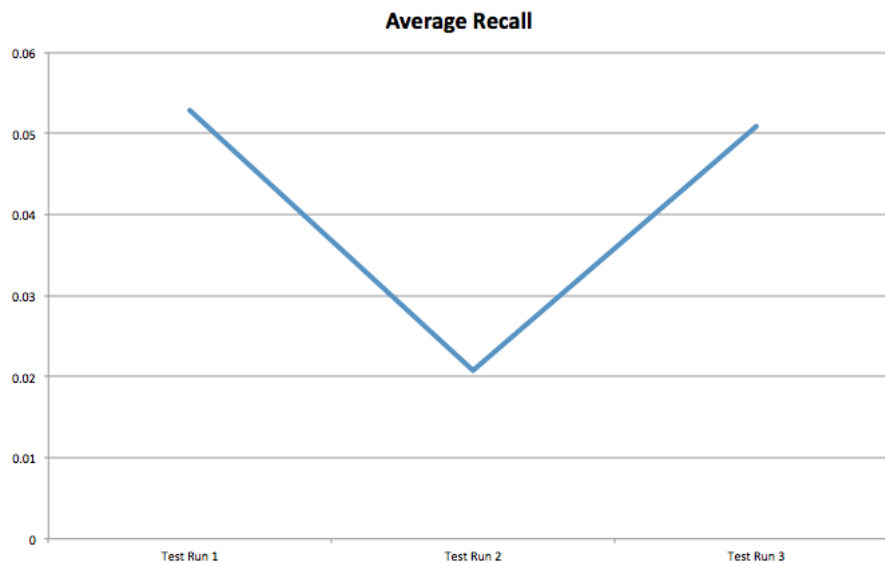


Figure 8: Average Recall

better, though surprisingly at a level similar to the first test. A lot more candidates were returned in the third test, but the ratio of experts to non-experts does not increase. This suggests that even with more data being added to the algorithm, the percentage of relevant returns (experts in this case) does not increase significantly.

6.3.3 Cosine Similarity

ExpertQuest measures the cosine similarity of each candidate's tweets and the DBpedia abstract on the programming language, as laid out in section 5.2.4 on page 28. This is used to represent how much each candidate was tweeting about the specific programming language. As expected, the average cosine similarity goes up across each test run as the amount of tweet data is increased. This can be seen in **Figure 9**, overleaf.

6.3.4 Comparison of Different Programming Languages

The results for Test Run 3 are displayed in **Table 4**. These results represent the ExpertQuest at its most useful. Some results that are revealing:

- ExpertQuest was only able to find experts in 21 out of the 53 programming languages.
- ExpertQuest was most successful at finding experts in programming languages whose development is actually hosted on GitHub. The precision here was much higher and the results much more relevant. The top four languages where experts were found are JVM (Java Virtual Machine) languages.

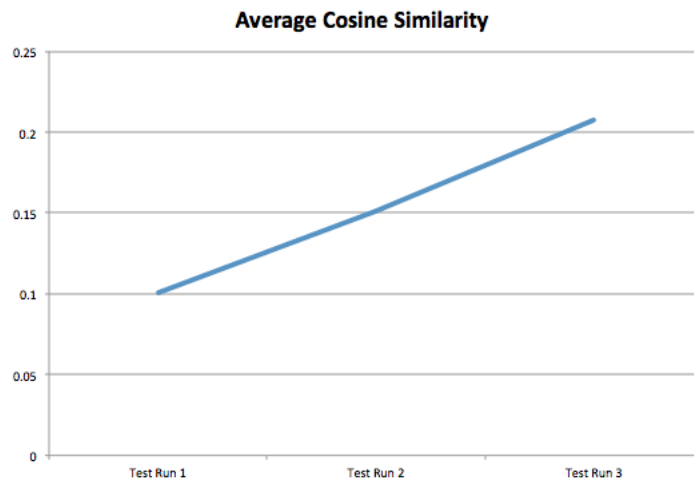


Figure 9: Average Cosine Similarity

	Total Candidates	Experts
Scala	18	14
Clojure	22	14
Groovy	24	14
Java	17	12
Erlang	18	11
Python	16	10
Ruby	15	9
Haskell	13	8
Perl	12	7
Dart	16	7
Lua	18	7
C	11	4
Go	28	4
R	9	3
C++	11	3
C#	8	2
D	23	2
AWK	1	1
MATLAB	1	1
Prolog	2	1
Smalltalk	5	1

Table 4: Programming Language Results from Test Run 3

6.4 Achievement of Requirements

ExpertQuest can now be evaluated as to whether it meets the primary and secondary requirements as outlined in Chapter 3.

6.4.1 Primary Requirements

These requirements are the key requirements that Maybury [56] outlines for an expert finding system.

- **The system should be able to identify experts from a field of candidates:** ExpertQuest meets this requirement. For the test runs, experts were defined as candidates with GitHub repositories in the desired programming language (see section 6.2.2 page 37) and ExpertQuest will rank these individuals higher than those without such repositories.
- **The system should be able to classify the level of expertise of each candidate:** ExpertQuest meets this requirement. The display of candidates offers several measures that can be used to classify the candidates - the number of Twitter and GitHub followers, the amount of code in their repositories and the amount they have been tweeting about the desired programming language.
- **The system should be able to validate the expertise of each candidate:** ExpertQuest meets this requirement. GitHub followers are a sign that a candidate's repositories are useful and is a reasonable way of assessing the candidate's credibility. A link to the candidates GitHub account is also supplied in the interface, so the user can go and take a look at the code for themselves.
- **The system should be able to rank candidates on multiple dimensions:** ExpertQuest meets this requirement. Expert candidates are ranked simply in descending order according to the criteria outlined in section 5.4 (on page 34).

6.4.2 Secondary Requirements

- **The system should use the hybrid approach:** As outlined in section 3.1 (on page 19), the hybrid approach uses all three approaches in one system - content analysis, social graph analysis and Semantic Web technology.
 - Content analysis is done on the Twitter stream and DBPedia abstracts via feature hashing.
 - Social graph analysis is provided via information on followers in the Twitter and GitHub APIs, albeit in a very trivial form.
 - Semantic Web technology is used in the form of the Linked Data interface with DBPedia.
- **The system should measure in some meaningful way all three properties of expertise:** ExpertQuest meets this requirement as follows:

- A candidate can have tweets, repositories and GitHub followers, demonstrating knowledge and credibility within the programming language in question.
- A candidate can have followers both on Twitter and GitHub, demonstrating that their social peers are engaged with their expertise.
- **The system should be based on real-time data available on the Internet and should also minimise any need for expensive extraction and transformation of data:** ExpertQuest only uses live data from Web based APIs.
- **The system should also allow a way to contact the expert if possible:** ExpertQuest provides links to the candidate's Twitter and GitHub profiles.
- **The system should be a usable prototype:** The finished system returns experts in approximately 40% of the languages used in the test runs. This is a weak but positive result.
- **If possible, the system should be user friendly and reasonably fast and should have a Web interface:** The system is not fast but it is Web based and easy to use.

6.5 Summary

- As more Twitter data was added to ExpertQuest the average precision and average cosine similarity went up. Recall was not increased by adding more data.
- ExpertQuest returns useful and relevant results, but testing suggests that this happens only in a specific subset of languages which are hosted on GitHub.
- ExpertQuest meets the requirements outlined in Chapter 3 with some caveats.

7 Conclusions and Future Work

In Chapter 2, expert finding was defined as efficiently identifying the right individual (or group) from a field of candidates that has the expertise to provide desired information or complete a desired task. As we have seen, ExpertQuest did provide an efficient way to find experts in specific programming languages across Twitter and GitHub. ExpertQuest will now be analysed as the prototype for a possible commercial system.

7.1 SWOT Analysis

7.1.1 Strengths

- ExpertQuest met the requirements for an expert system as set out in Chapter 3.
- ExpertQuest demonstrates that one can build a reasonably useful system that returns experts in a specific field using only data available via API. For programming languages that have large communities on GitHub, the experts identified are often the programmers who play major roles in their respective language ecosystems as library authors and the like.
- The Clojure language and tooling infrastructure is an excellent choice for this kind of Web enabled data extraction project. The Lisp based “data as code” philosophy and the immutable data structures makes the code succinct and concise. Furthermore, the fact that Clojure has built-in concurrency means that ExpertQuest could scale if needed. The available of Java based libraries such as OpenNLP is also a major advantage.

7.1.2 Weaknesses

- ExpertQuest is pushing the limits for how slow a Web interface can be and still remain useful. Queries may take several minutes. A different design may be warranted if commercial development was taken further - batching and processing results in a database perhaps. This would break one of the requirements for the prototype (no data extraction).
- Over 60% of the programming languages returned no useful results (though this is a factor of the data sources chosen, some programming languages are just not discussed on Twitter). The precision and recall results are poor for many of the languages that do return results also.
- The first large compromise in the design as outlined in section 5.5 (matching Twitter and GitHub accounts, see pages 34-35) is a major weakness. In reality a commercial software product with this flaw would not be useful as it would miss large chunks of data. Even the solution used here has a small chance of associating two different people in the same profile. This would need to be remedied before a commercial product could be released.

7.1.3 Opportunities

- The general design approach defined in section 4.8 (page 25) could be reused for other areas of expertise i.e. Twitter can be used as a way of identifying expert candidates that could then be filtered and ranked via information from other sources.

7.1.4 Threats

- ExpertQuest is built on the APIs of a company that is notorious for cutting off access to developers. Twitter has killed multiple startup businesses in the last few years in this way.
- API limits are also a factor and if the application became popular this could become a real bottleneck.
- Privacy and data protection could also become major issues if a commercial version of ExpertQuest was released.

7.2 Future Work

There are several areas of possible future work that arise out of this research:

- Researching how one authenticates the accounts of one individual across multiple social networks. Perhaps a technology such as FOAF could be used to solve this problem.
- The general approach of augmenting Twitter and other social network data with other sources of data could also be explored in other areas such as biotechnology. If a reliable source of up-to-date scientific research was made available, the hybrid approach used by ExpertQuest could then be used to create a system for expert finding using this data source to rank expertise.
- ExpertQuest could be further developed as full application using a data store of some kind. This would mean that the data would be slightly out of date but the queries would run a lot faster. There would also be opportunities for doing further analysis across various programming languages communities, for instance, seeing how experts and their followers are related to one another across the social graph. There are multiple projects of this type that could be carried out using a few weeks worth of ExpertQuest queries.

Appendix 1: List of Tiobe Top 50+ Programming Languages

- ABAP
- ActionScript
- Ada
- Assembly language
- AWK
- Bash
- C
- C#
- C++
- Clojure
- COBOL
- CoffeeScript
- D
- Dart
- Eiffel
- Erlang
- F#
- Forth
- Fortran
- FoxPro
- Go
- Groovy
- Haskell
- Inform
- Java
- JavaScript

- LabVIEW
- Lisp
- Logo
- Lua
- MATLAB
- Max
- ML
- Objective-C
- OpenEdge ABL
- Pascal
- Perl
- PHP
- PL/I
- PL/SQL
- PostScript
- Prolog
- Python
- R
- RPG
- Ruby
- Scala
- Scheme
- Scratch
- Smalltalk
- T-SQL
- VB
- VB .NET

Appendix 2: Test Run 1 Results

Programming Language	Total Candidates Found	Total Experts* Found	Precision	Recall
AWK	1	1	1.000	0.100
Dart	2	2	1.000	0.200
Perl	2	2	1.000	0.200
Java	4	3	0.750	0.300
MATLAB	3	2	0.667	0.200
Python	3	2	0.667	0.200
Erlang	5	3	0.600	0.300
Scheme	5	3	0.600	0.300
Go	4	2	0.500	0.200
Prolog	2	1	0.500	0.100
Clojure	7	2	0.286	0.200
Haskell	7	2	0.286	0.200
C++	5	1	0.200	0.100
Groovy	6	1	0.167	0.100
Scala	6	1	0.167	0.100

*See section 6.2.2 for expert definition

Appendix 3: Test Run 2 Results

Programming Language	Total Candidates Found	Total Experts* Found	Precision	Recall
AWK	1	1	1.000	0.033
Dart	1	1	1.000	0.033
Perl	2	2	1.000	0.067
R	3	3	1.000	0.100
Ruby	5	4	0.800	0.133
Erlang	5	3	0.600	0.100
Java	5	3	0.600	0.100
Scheme	5	3	0.600	0.100
Groovy	6	3	0.500	0.100
Clojure	5	2	0.400	0.067
MATLAB	5	2	0.400	0.067
Go	6	2	0.333	0.067
Haskell	6	2	0.333	0.067
Python	3	1	0.333	0.033
Scala	6	1	0.167	0.033

*See section 6.2.2 for expert definition

Appendix 4: Test Run 3 Results

Programming Language	Total Candidates Found	Total Experts* Found	Precision	Recall
AWK	1	1	1.000	0.020
MATLAB	1	1	1.000	0.020
Scala	18	14	0.778	0.280
Java	17	12	0.706	0.240
Clojure	22	14	0.636	0.280
Python	16	10	0.625	0.200
Haskell	13	8	0.615	0.160
Erlang	18	11	0.611	0.220
Ruby	15	9	0.600	0.180
Groovy	24	14	0.583	0.280
Perl	12	7	0.583	0.140
Prolog	2	1	0.500	0.020
Dart	16	7	0.438	0.140
Lua	18	7	0.389	0.140
C	11	4	0.364	0.080
R	9	3	0.333	0.060
C++	11	3	0.273	0.060
C#	8	2	0.250	0.040
Smalltalk	5	1	0.200	0.020
Go	28	4	0.143	0.080
D	23	2	0.087	0.040

*See section 6.2.2 for expert definition

References

- [1] "Tool Use, Hunting & Other Discoveries",
<http://www.janegoodall.org/chimpanzees/tool-use-hunting-other-discoveries>
- [2] Vicki K. Bentley-Condit and E.O. Smith, "Animal tool use: current definitions and an updated comprehensive catalog", *Behaviour*, Volume 147, Issue 2, 2010.
- [3] Christophe Boesch and Hedwige Boesch, "Tool Use and Tool Making in Wild Chimpanzees", *Folia Primatologica*, 54:86-99, 1990.
- [4] Tomasello, Davis-Dasilva, Camak & Bard, "Observational learning of tool-use by young chimpanzees", *Human Evolution*, Vol 2, 175-183, 1987.
- [5] Theodoros Lappas, Kun Liu and Evimaria Terzi, "A Survey Of Algorithms And Systems For Expert Location In Social Networks," in *Social Network Data Analytics*, C. C. Aggarwal, Ed., Springer Science and Business Media, pp. 215-241, 2011.
- [6] Dawit Yimam-Seid and Alfred Kobsa, "Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach", *Journal of Organizational Computing and Electronic Commerce*, 13(1), 1-24, 2003.
- [7] Dietrich Stout, "Stone toolmaking and the evolution of human culture and cognition", *Philosophical Transactions B*, 10.1098/rstb.2010.0369, 28 February 2011.
- [8] "Intangible assets, intellectual capital or property? It does make a difference",
http://klminc.com/branding_brand-strategy/intangible-assets-intellectual-capital-or-property-it-does-make-a-difference
- [9] "Expertise", <http://www.businessdictionary.com/definition/expertise.html>
- [10] <http://trec.nist.gov>
- [11] Theodoros Lappas, Kun Liu and Evimaria Terzi, "A Survey Of Algorithms And Systems For Expert Location In Social Networks," in *Social Network Data Analytics*, C. C. Aggarwal, Ed., Springer Science and Business Media, pp. 215-241, 2011.
- [12] Krisztian Balog, Leif Azzopardi and Maarten de Rijke, "Formal Models for Expert Finding in Enterprise Corpora", 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, 2006.
- [13] Pavel Serdyukov and Djoerd Hiemstra, "Being Omnipresent To Be Almighty: The Importance of the Global Web Evidence for Organizational Expert Finding", *Future Challenges in Expertise Retrieval SIGIR workshop*, Singapore, July 24, 2008.
- [14] Krisztian Balog, Leif Azzopardi and Maarten de Rijke, "Formal Models for Expert Finding in Enterprise Corpora", 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval, 2006.

- [15] M. Maybury, R. D'Amore, and D. House, "Awareness of organizational expertise", *International Journal of Human-Computer Interaction*, 14(2): 199-217, 2002.
- [16] Theodoros Lappas, Kun Liu and Evimaria Terzi, "A Survey Of Algorithms And Systems For Expert Location In Social Networks," in *Social Network Data Analytics*, C. C. Aggarwal, Ed., Springer Science and Business Media, pp. 215-241, 2011.
- [17] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri, Giuliano Vesci, "Choosing the Right Crowd: Expert Finding in Social Networks", *International Conference on Extending Database Technology / International Conference on Database Theory*, 13 March, 18 - 22 2013.
- [18] Ahmad Kardan, Amin Omidvar, Farzad Farahmandnia, "Expert finding on social network with link analysis approach", *19th Iranian Conference on Electrical Engineering (ICEE)*, 2011.
- [19] http://en.wikipedia.org/wiki/HITS_algorithm
- [20] B. Dom, I. Eiron, A. Cozziand and Y. Zhang, "Graph-based ranking algorithms for email expertise analysis", *Proceedings of the 8th ACM SIGMOD workshop on Research issues in Data Mining and Knowledge Discovery (DMKD)*, 2003.
- [21] Jie Li, Harold Boley, Virenda Bhavsar, and Jing Mei, "Expert finding for eCollaboration using FOAF with RuleML rules", *Proc. of the 2006 Montreal conference on eTechnologies*, 2006.
- [22] Ahmad Kardan, Amin Omidvar, Farzad Farahmandnia, "Expert finding on social network with link analysis approach", *19th Iranian Conference on Electrical Engineering (ICEE)*, 2011.
- [23] Theodoros Lappas, Kun Liu and Evimaria Terzi, "A Survey Of Algorithms And Systems For Expert Location In Social Networks," in *Social Network Data Analytics*, C. C. Aggarwal, Ed., Springer Science and Business Media, pp. 215-241, 2011.
- [24] [https://en.wikipedia.org/wiki/FOAF_\(ontology\)](https://en.wikipedia.org/wiki/FOAF_(ontology))
- [25] <https://en.wikipedia.org/wiki/RuleML>
- [26] Titus Schleyer et al. "Requirements for expertise location systems in biomedical science and the Semantic Web", *Proceedings of the 3rd Expert Finder Workshop on Personal Identification and Collaboration: Knowledge Mediation and Extraction (PICKME'08)*, 2008.
- [27] M. Stankovic, C. Wagner, J. Jovanovic, & P. Laublet, "Looking for Experts? What can Linked Data do for you?" *LDOW*. 2010.
- [28] Titus Schleyer et al. "Requirements for expertise location systems in biomedical science and the Semantic Web", *Proceedings of the 3rd Expert Finder Workshop on Personal*

- Identification and Collaboration: Knowledge Mediation and Extraction (PICKME'08), 2008.
- [29] Florian Metze et al, "A community based expert finding system", Proceedings of IEEE Int. Conf. on Semantic Computing, 2007.
- [30] <http://project.askspre.de>
- [31] Mark T. Maybury, "Discovering distributed expertise - Regarding the 'Intelligence' in Distributed Intelligent Systems", MITRE, 2007.
- [32] <https://pypi.python.org/pypi/google-scholar-scrap/>
Google Scholar: <https://scholar.google.com>
- [33] <http://www.programmableweb.com>
- [34] <http://www.pcworld.com/article/2883992/linkedin-restricts-api-usage.html>
- [35] <https://dev.twitter.com/rest/public/>
- [36] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri, Giuliano Vesci, "Choosing the Right Crowd: Expert Finding in Social Networks", International Conference on Extending Database Technology / International Conference on Database Theory, 13 March 18 - 22 2013.
- [37] <https://developer.github.com/v3/>
- [38] Christian Bizer, Tom Heath, and Tim Berners-Lee, "Linked Data - The Story So Far", Special Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS), 2009.
- [39] <http://linkeddata.org/>
- [40] <http://wiki.dbpedia.org/OnlineAccess#h28-13>
- [41] <http://clojure.org/>
- [42] <http://getbootstrap.com/2.3.2/index.html>
- [43] <http://opennlp.apache.org/>
- [44] <http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html>
- [45] <https://github.com/adamwynne/twitter-api>
- [46] <https://github.com/Raynes/tentacles>
- [47] <https://github.com/dakrone/clojure-opennlp>
- [48] "Hashing Language", <http://blog.someben.com/2013/01/hashing-lang/>

-
- [49] http://en.wikipedia.org/wiki/Feature_hashing
 - [50] George Forman and Evan Kirshenbaum, "Extremely fast text feature extraction for classification and indexing", Proceedings of the 17th ACM conference on Information and knowledge management, ACM, 2008.
 - [51] Kilian Weinberger et al. "Feature hashing for large scale multitask learning", Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009.
 - [52] <https://github.com/weavejester/lein-ring>
 - [53] <https://github.com/weavejester/compojure>
 - [54] <https://github.com/weavejester/hiccup>
 - [55] <http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html>
 - [56] Mark T. Maybury, "Discovering distributed expertise - Regarding the 'Intelligence' in Distributed Intelligent Systems", MITRE, 2007.