LOCAL EXPERTS IN SOCIAL MEDIA

A Thesis

by

VANDANA BACHANI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	James Caverlee
Committee Members,	Thomas R. Ioerger
	John B. Mander
Head of Department,	Nancy Amato

December 2013

Major Subject: Computer Science

ABSTRACT

The problem of finding topic experts on social networking sites has been a continued topic of research. This thesis addresses the problem of identifying local experts in social media systems like Twitter. Local experts are experts with a topical expertise that is centered around a particular location. This geographically-constrained expertise can be a significant factor for enhanced answering of local information needs (What is the best pub in College Station?), for interacting with local experts (e.g., in the aftermath of a disaster), and for accessing local communities. I developed a local expert finding system – called OLE (online local experts) – that leverages the crowdsourced location-topic labels provided by users of the popular Twitter service. Concretely, I mine a collection of 108 million tweets for evidence of local topics of discussion occurring with user-mentions and location pairs; based on this collection, I developed a learning-to-rank approach that incorporates topic-location entropy and a local expert perimeter for varying the expertise focal window. In comparison with alternative expert finding approaches, I find that OLE is quite effective in finding local experts and achieves a 37.72% increase in mean average precision and a 16.8% increase in NDCG scores, across a comprehensive set of queries.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Prof. James Caverlee, for his continued guidance for the work I have done so far. He has been a great source of inspiration and continues to motivate me to help me achieve my research objectives. I would like to express my sincere gratitude to my thesis committee members, Prof. Thomas Ioerger and Prof. John Mander, for their advice, encouraging words and feedback on various milestones of this thesis. I would also like to thank Zhiyuan Cheng, Dr. Krishna Kamath and my other colleagues from infolab for their support, technical advice and feedback during the various phases of my work. Lastly I would like to thank Swayambhoo Jain, a friend who is currently doing his PhD. in University of Minnesota for reviewing my thesis.

NOMENCLATURE

LSTS	Location Specific Topic Score
LSTSM	Location Specific Topic Score Matrix
MAP	Mean Average Precision
NDCG	Normalized Discounted Cumulative Gain
NLP	Natural Language Processing
OLE	Online Local Experts
SNW	Social Networking Websites
ULM	User Location Mention

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Social networking websites (SNWs) like Facebook, Twitter, Google Plus etc. are used by people to connect to other users of the network for various social and information needs. One can use the SNW's search and recommendation services to expand his/her social circle. In order to connect users to the right set of people for answering their particular information needs, for engaging in topic-relevant conversations and growing their interest circles, its important to be able to find topic experts. The problem of finding topic experts in social networking sites is a challenging problem and a continued topic of research [1, 2, 3]. Current approaches of finding topic experts in Twitter rely on user's profile and tweet features, flow of expertise in the network (pagerank-style) and crowdsourced features like user-curated lists. But many times just finding topic experts doesn't suffice, as certain topics are time-sensitive or very strongly influenced by location. These topics require the experts to be better at topics in the context of either time or location or both.

The phenomenon of soliciting local expert opinion is prevalent in current day social networks. Figure 1.1 provides a glimpse of location specific Q&A behavior on Twitter.

This thesis addresses the problem of finding local experts in social media systems like Twitter. A local expert is a user with a topical expertise that is centered around a particular location. This geographically-constrained expertise has a number of applications such as,

- Enhanced answering of local information needs, for e.g. "Which is the best pub in College Station?"
- Interacting with local experts in time of need, for e.g. in the aftermath of a

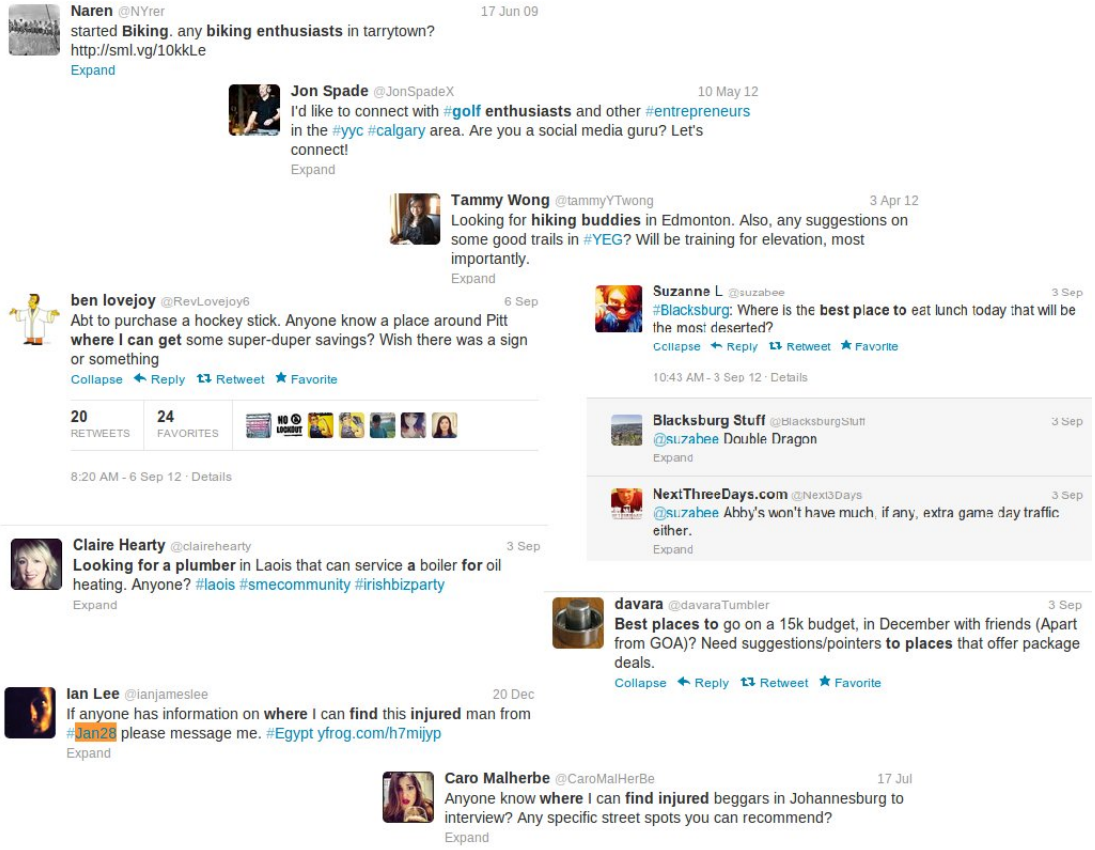

Figure 1.1: Examples of Local Q&A behavior on Twitter

disaster or unfortunate event.

- Accessing local communities or interest groups in a locality, for e.g. hiking enthusiasts in Seattle.

The existing approaches to find topic experts do not consider the geographical constraints imposed by the problem of finding local experts and hence cannot be used in their current form. Also the problem of finding local experts has its own set of challenges such as ambiguous or unknown location associations of users of the social network, lack of direct mechanisms (like lists) to find topic-specific location associations of users and lack of methods for evaluating local expertise.

As part of this thesis, I developed a local expert finding system – called OLE ("Online Local Experts") – that leverages the crowdsourced location-topic labels provided by users of the popular Twitter service. From a random location-mentions sample of tweets from Twitter (108 million tweets), it was found that 45% (48.6 million) of these tweets associated users with locations and often times discussed local events, topics, etc. These pairs of user-location mentions by third-parties (other tweeters) collaboratively serve as a better indicator for predicting location associations of users. Based on this collection comprising of 48.6 million tweets containing pairs of user-location mentions, I developed a learning-to-rank approach that incorporates distance-weighted LSTS ("Location Specific Topic Scores) in a local expert perimeter and topic-location entropy for varying the expertise focal window. The system suggests top 'k' experts for a given local query, comprising of a topic and the area of local influence. An analysis of local topics of discussion was done to identify the subset of queries, called "local queries" which are particularly suited for a local expert system.

Experiments were conducted to compare the system with existing topic expert methods and an adaptation of list-based topic expert finding method to find local experts. As hypothesized, topic expert methods in their current form were not able to match the accuracy of local expert systems for queries with location constraints. OLE outperformed the adapted list-based approach with 37.72% increase in the mean average precision and 16.8% increase in the NDCG scores across a comprehensive set of local queries.

The problem of finding local experts in social media is pretty new, and no work has been done in this domain till now. The following chapters of this thesis discuss the problem and the proposed solution in greater detail. Chapter 2 provides an overview of the related work. Chapter 3 discusses the problem of finding local experts and the

associated challenges. The proposed model for solving the problem of finding local experts is discussed in chapter 4. Chapter 5 discusses the data collected for this research and analysis of local topics. Chapter 6 describes the implementation of the proposed model as a search engine for finding local experts (OLE). The experiments conducted to evaluate the proposed model and the results of the evaluation are presented in Chapter 7 and Chapter 8 concludes the work with future directions of research on this topic.

2. LITERATURE REVIEW

The research which motivates the work in this thesis and provides a context for the problem of finding local experts, can be divided into three main categories: social search and social Q&A, finding topic experts in social media and expertise analysis on Q&A websites.

2.1 Social Search and Social Q&A

Social search refers to using social mechanisms to find information online. There are search engines [4] which index publicly available social media data to answer users queries based on the content provided by other users, relevant to that query, in the past. Some researchers have created tools such as HeyStaks [5], a browser plugin, to integrate social information, such as upvotes for a query result, into the traditional search engines, to enhance the search experience of other similar or socially-connected users.

Social Q&A is a part of the social search phenomenon where users explicitly state their information need as questions to their social groups in order to get tailored responses to their complex queries. Honeycut and Herring [6] did an in-depth analysis of tweets containing '@' symbol, and found that around 2% of those were about soliciting information from users. Computationally identifying Q&A behaviour in social media data is a hard NLP problem. K. Dent and S. Paul [7], tried writing an NLP based parser for processing tweets to find tweets which can be categorized as questions, and realized its a hard problem given the idiosyncrasies of Twitter data as questions on Twitter are hardly well-formed. In another paper [8], they tried to analyze if social Q&A is viable, trying to find whether Twitter is a good place to ask questions. Morris et al. [9] did a survey of the status message Q&A behavior to un-

derstand what kinds of questions people ask their social networks, the motivation for asking, type of answers received, and motivations for answering and not answering. In another study [10], they did a comparative analysis of information seeking using search engines and social networks, to check which method users preferred more and what are the pros and cons of the two methods. Jiang et. al [11], studied the cultural differences in the social Q&A behavior. The results of these survey-based papers could be said to be a little biased to the set of users who participated in the survey. Erin et al. [12] tried to investigate a much noble application of social Q&A, i.e. if they can potentially be used as a resource to help aid blind users by creating VizWiz social app for blind iphone users.

Another dimension of the social Q&A research is to understand the feasibility of asking questions of people outside of one's network. Jeffrey Nichols and his team [13] conducted real-life experiments on random Twitter users and did find interesting positive results on feasibility of depending on strangers for some special kind of information needs (time-sensitive, non-personal, opinion seeking queries).

2.2 Topic Experts in Social Media

There has been a considerable amount of work done to identify topic experts among Twitter users. Bernstein et. al designed Collabio [14], a tagging-based Facebook game that encouraged users to tag people in their networks. The metadata collected by the game about users was intended to be used to find experts in social media.Weng et. al [1], proposed a ranking similar to Page-Rank, called TwitterRank, that uses the information from Twitter social graph and information from tweets to identify experts in specific topics. Pal et. al [2] used a set of 15 features extracted from the Twitter graph and tweets posted by the users to estimate their expertness in topics. Ghosh et. al. [3] devised a system called Cognos which used Twitter

Lists feature, which are user-curated lists of people, to identify topic experts and claim to perform better than graph and tweet feature based expert finding systems. Aardvark.com [15] is a commercial social expert finder, which tries to address the challenge of determining the right person for a person's information need. They studied how factors like trust due to intimacy, user's social graph, etc. influenced a person's information need and the quality of answers.

This thesis concentrates on finding local experts in social media sites like Twitter, who are different from topic experts in terms of the geographical constraints posed due to the local criteria. A topic expert may not be a local expert on that topic in an given location as he may lack knowledge on that topic from the particular location's perspective. For example, if a person wants to know where to get fresh "asiatic lilies in chicago" with all the details and options available, a local expert whose expertise is "lilies in chicago" would be able to assist him better than a world famous expert on lilies. Besides, the problem of local experts comes with additional set of challenges which are highlighted in later chapters.

2.3 Expertise Analysis on Q&A sites

Several Q&A websites like Yahoo! Answers, Quora, etc. and Q&A forums like stackoverflow, etc. are extensively used by internet users to ask and answer questions online. The question answering behavior on Q&A sites has been well-studied. [16] and [17] studied the types of questions users posted on Q&A websites and classified them into information-seeking, advice-seeking, opinion-seeking, and non-information seeking kinds of questions. Lada et. al [18] studied Q&A threads on Yahoo! Answers and analyzed various factors which affected the quality of answers such as user's entropy, type of question, etc. and also the factors which determined the expertise of users depending on the question domains for which they answer questions. Paul et. al

[19] studied who are the authoritative users on Quora and what features distinguish them from the rest of the users. In a recent study [17], a group of researchers proposed and evaluated methods, to predict the satisfaction levels of a web searcher from the existing community-based Q&A sites like Yahoo! Answers, Baidu Knows, etc. Though the research in this section is not directly related but motivates our methods of analysis of features of local experts.

3. PROBLEM FORMULATION AND IMPORTANT RESEARCH QUESTIONS

This chapter provides a formal definition for the problem of finding local experts, defines the terms and notations and discusses the key challenges associated with the problem and the intuition for solving the problem of local experts.

3.1 Problem Definition

Consider a social network consisting of a set of $|\mathcal{U}|$ [1] users denoted by $\mathcal{U} = \{u_1, \cdots, u_{|\mathcal{U}|}\}$, a set of $|\mathcal{T}|$ topics denoted by $\mathcal{T} = \{t_1, \cdots, t_{|\mathcal{T}|}\}$, and a set of $|\mathcal{L}|$ locations denoted $\mathcal{L} = \{l_1, \cdots, |\mathcal{L}|\}$. The problem of *finding local experts* for an input query q is defined as follows:

Definition 1 *[Local Expert Finding] Given a query q consisting of tuple $[T(q), l(q)]$ [1], where $T(q)$ is list of topics and $l(q)$ is the location, find the set of top k local experts from the set of users, $\mathcal{U}$, with knowledge about the query topics within distance $\epsilon \geq 0$ from the query location $l(q)$.*

Figure 3.1 captures the notion of ϵ radius around a location. ϵ denotes the importance of geographical proximity to the query location in the results. $\epsilon = 0$ denotes that the user wants to find experts exactly at the given query location and a very high value of ϵ (close to max distance between two locations possible), means the user is least concerned about the local expertise and mostly concerned about the topic expertise.

[1] $|\mathcal{U}|$ denotes the number of elements in set $\mathcal{U}$.

[1] Assuming that the set of topics, $T(q)$, and location, $l(q)$, can be extracted from the query text or they may be explicitly provided.

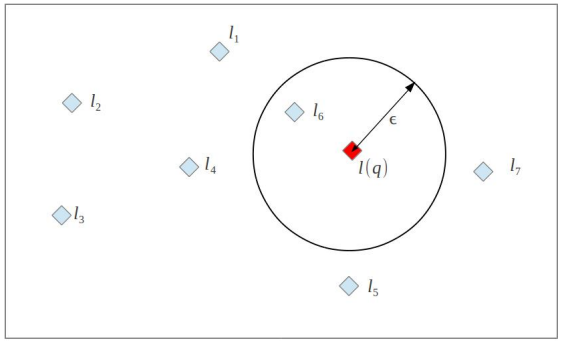

Figure 3.1: Relevant locations within ϵ distance from query location $l(q)$

3.2 Important Research Questions

The problem of finding local experts in social networks introduces the idea of location constrained expertise which presents a relatively new research avenue in the domain of social network research. Being a new problem it comes with a new set of challenges and poses important research questions. Some of these questions are directly related to the motivation and importance of solving this problem and some are about feasibility and the methodology which one should use to solve this problem. Following are some of the important questions which have a direct bearing on the work done in this thesis:

1. What is the motivation behind solving the problem of local experts? Why is it important?

2. What are the challenges associated with the problem of finding local experts?

3. What mechanisms are provided by the social media to identify local experts?

3.3 Motivation

Topic expert systems serve to connect social media users with the right set of people for answering their particular information needs; for engaging in topic-relevant

conversations and expanding their interest circles. Many times finding topic experts doesn't suffice, as certain-topics are time-sensitive or the user information needs may be constrained by locality. These scenarios require the experts to be better at topics in the context of either time or location or both. For example, if a user is interested in knowing about the current wait time in the security queue at the Houston-Bush International Airport by asking the people (suppose there was an app which one could use), a person who is currently waiting or is at the airport or a frequent traveler from Houston are the best people to ask this query. Similarly a person interested in "hiking in Seattle" would like to get in touch with hiking experts in Seattle to tap into the community in Seattle. In the current scenario, where there is a shift in the web user's information seeking behavior with a upward trend in local search[2] and local Q&A, there is a need to focus attention on topic experts with geographical constraints. This geographically-constrained expertise can be a significant factor for enhanced answering of local information needs (e.g., Which is the best pub in College Station?), for interacting with local experts (e.g., in the aftermath of a disaster), and for accessing local communities. The above mentioned benefits and the need to consider the local aspect while trying to find topic experts motivate this research. This research paves the way for closing the gap between the online and physical worlds further.

3.4 Challenges

The challenges associated with the problem of finding local experts include:

- **Faulty or in-comprehensive association of locations to users**

  Profile location as indicated by user profiles is the only way of knowing the association of a user to a location on Twitter. But often these locations are

[2]As per a study conducted in September 2012 (http://chitika.com/insights/2012/local-search-study), 24% of Google search queries were found to be local in nature.

incorrect or ambiguous. Much research has been done to predict a user's true location. In today's dynamic lifestyle, where people often move between locations due to several reasons related to education, work, leisure travel, etc. a user usually can be associated with several locations. Facebook tries to capture this by introducing location types in profile, namely "Place of Birth" and "Current location", but those are also not sufficient to capture all the location associations of a user. The type of association of a location to a user is also important as it affects the knowledge of a person about specific aspects or topics in a location.

- **Lack of direct mechanisms like twitter lists for local expertise**

  The most challenging part of solving the problem of finding local experts is identifying the signals and mechanisms provided by the social network which can help us find local experts in Twitter. Mechanisms used for topical expertise can be divided into 3 categories: the signals generated by a user which constitute the profile information, the tweets, the user's activity, etc., the signals generated by the network features such as ones used in [1], and endorsements such as user-curated lists. As claimed by [3], the user-curated lists work better than the network and user's self-endorsement features for identifying topic experts. Unfortunately there are no direct endorsements like lists available which can help find local experts easily.

- **Lack of methods for evaluating local expertise**

  Evaluation of topic expertise is relatively easy due to availability of ground truth in terms of known celebrity or domain experts (e.g. Guido is a known python expert). It has been observed that local experts are non-celebrity users which makes it hard to evaluate them.

3.5 Solution Approach

Current approaches of finding topic experts in Twitter rely on user and profile features, flow of expertise in the network (pagerank-style) and crowdsourced features like user-curated lists. The existing approaches to find topic experts do not consider the geographical constraints imposed by the problem of finding local experts and hence cannot be used in their current form. Alternative approaches need to be devised to handle the additional requirements and challenges associated with the problem of finding local experts.

3.5.1 Crowd-Sourced Location Topic Assignments

I mined a collection of 108 million tweets and found 48.6 million tweets containing occurrences of pairs user-mentions and location. An initial look at the data clearly signaled the value of these occurrences as user-location mention pairs associate users with locations. The tweets in table 3.5.1 show how user-location mentions associate @lancearmstrong with Austin and @jack (Jack Dorsey, Founder of Twitter) with San Francisco. This phenomenon can be described as "crowd-based location assignment" as these locations are assigned by third-parties (other tweeters) to users, and may not be related to user's self-identified locations. These location assignments provide a good estimate of a user's location. One major advantage of this approach is its ability to associate multiple locations with a user. An important feature of the user-location mention tweets is the topic of discussion of the tweet. When a user is referenced by an "@mention" in a tweet in context of a location and topic (derived from the tweet), it creates a signal associating the user with the location in context of the topic. When several users associate the mentioned user with the topic in the given location in their tweets, it acts as a crowd-based endorsement for the mentioned user. These location-topic associations of users come with varying degrees of strength depending

@lancearmstrong - austin
@lancearmstrong we're skating across America and are headed for Austin, Texas!...
@lancearmstrong First biz trip to Austin, visiting a customer Jan 23rd; I have 24...
@lancearmstrong lance do u still bike in the northwest hills in Austin? I recently...
@lancearmstrong it was great living next to that place for a year. Miss Austin so...
@lancearmstrong did you just turn off of 11th street in Austin? If so, I totally saw...
@jack - san francisco
This is what it looks like when coffee is ground in the middle of San Francisco...
@sfciti: We're working with @jack, @biz, @bchesky and more to imagine a better...
iPhone 5, poor signal. RT @jack: A beautiful morning in San Francisco for an...
The #TRSalon in San Francisco had many highlights - here's @jack on creating...
@jack Hi Jack! Im very soon in San Francisco... Do you have tips About trendy...

Table 3.1: User Location Mention Tweets

on the number of user-location-topic mentions. Topic based location profiles of users represented in form of a "Location Specific Topic Score Matrix", LSTSM, can be constructed. The topic-location profile of **@TimTebow** created based on number of user-location mentions for the topic "football" as shown in table 3.5.1 clearly depicts the major locations he has been associated in his life till date as more number of tweets are associated with the locations of the teams he has been with. Similarly, I found that well-known tattoo artist **@amijames** who mentions New York and Miami as his twitter location, is also associated with London for the topic "tattoo artist", according to his topic-location profile created as described above and is due to the fact that he has a tattoo boutique in London which he visits often. The location

denver	**florida**	**new york**	**jacksonville**	**colorado**	**dallas**	**philippines**
443	89	71	44	24	17	14

Table 3.2: Topic-Location Profile of @TimTebow based on user-location mentions

specific topic scores prove to be a good indicator of local expertise as they act as

an indirect endorsement mechanism and help delineate the space of potential local experts. I developed a unified single-step model based on LSTS for solve the problem of finding local experts where the scores calculated by the model are representative of the local topic expertise as a whole.

4. PROPOSED MODEL

Location specific topic expertise of a user, $u_i \in \mathcal{U}$ depends on three important factors derived from the user-location mention tweets:

1. Content-based Expertise Score, captured by tf-idf score on user-location mention tweets.

2. Local Expertise Score, captured by the crowdsourced endorsements of local topic expertise given by Distance-Weighted Location Specific Topic Scores.

3. Distribution of knowledge of a user in a topic across locations, captured by the Topic-Location Entropy.

The following sections discuss the above factors in detail.

4.1 Content-based Expertise Score

Cosine similarity based on tf-idf vectors due to the content of the user-location mention tweets establishes whether the person has the local expertise in the topic or not. As described in section 6, Apache Lucene returns documents from its index based on this score. This score helps delineate the space of potential experts by weighing the scores with considering all user's topics but does not capture the relation between user's topics and location associations.

4.2 Local Expertise Score

The local expertise score is used to quantify the location specific topic expertise of a user for the given query. This score should take into consideration the variation of expertise with distance from query location and also what is the best radius around a location to find experts. For example, if a user is interested in finding local

experts for "tourist places" in college station, an expert from Austin with knowledge of "tourist places in college station", must be ranked a little low compared to a person from college station. For the same query system should consider potential experts from nearby locations such as Bryan. These two challenges are addressed using distance-weighted location specific topic scores and local expert perimeter for locations.

4.2.1 Distance-Weighted Location Specific Topic Scores

Building upon the approach mentioned in section 3.5, location specific topic profiles of users are created. Expertise of user $u_i \in \mathcal{U}$ in topic $t_j \in \mathcal{T}$ at location $l_k \in \mathcal{L}$ is quantified by *Location Specific Topic Score* (LSTS) $s_i(j,k)$. The location specific topic scores for users are calculated using the user-location mention tweets collected as mentioned in 5.

$$LSTS(t_j, l_k, u_i) = s_i(j,k) = \begin{array}{l}\#\text{ of mentions of topic, } t_i \text{ and location, } l_j \text{ in the} \\ \text{user-location-topic map,}\end{array} \tag{4.1}$$

. Larger value of LSTS ($s_i(j,k)$), denotes user $u_i's$ higher expertise about topic t_j at location l_k. The LSTSs of a particular user u_i are arranged in a *Location Specific Topic Score Matrix* (LSTSM) denoted by $\mathbf{S}_i$

$$\mathbf{S}_i := \begin{bmatrix} s_i(1,1) & \cdots & s_i(1,|\mathcal{L}|) \\ \vdots & \vdots & \vdots \\ s_i(|\mathcal{T}|,1) & \cdots & s_i(|\mathcal{T}|,|\mathcal{L}|) \end{bmatrix}. \tag{4.2}$$

As per the definition of problem of finding local experts, a query q consists of a tuple $[T(q), l(q)]$, where $T(q)$ is list of topics and $l(q)$ is the location, a local expert

finding system needs to find the set of top k local experts from the set of users, $\mathcal{U}$, with knowledge about the query topics within distance $\epsilon \geq 0$ from the query location $l(q)$.

In order to impose geographical proximity constraints in query results of the local expert finding system, the notion of ϵ *thresholded distance weight* has been introduced. The formula for this weight is given by,

$$w_\epsilon\left(l\left(q\right), l_i\right) := \begin{cases} w\left(d\left(l\left(q\right), l_i\right)\right) & \text{if } d(l(q), l_i) \leq \epsilon \\ 0 & \text{Otherwise} \end{cases}, \tag{4.3}$$

where $d(l(q), l_i)$ is the physical distance (using haversine formula) between location l_i and query location $l(q)$, and $w(d(l(q), l_i))$ is any monotonically decreasing function of $d(l(q), l_i)$. For instance it can be

$$w(d(l(q), l_i)) = \left(\frac{d_{min}}{d(l(q), l_i) + d_{min}}\right)^{\alpha}. \tag{4.4}$$

A more sophisticated choice of the weight function is,

$$w(d(l(q), l_i)) = \left(\frac{d_{min}}{d(l(q), l_i) + d_{min}}\right)^{\frac{\alpha}{g(\epsilon)}}. \tag{4.5}$$

where $g(\epsilon)$ is a monotonically increasing function of ϵ. This choice captures the notion of ϵ more nicely because a user providing higher ϵ does not attach much importance to the specificity of geographical proximity. The distance d between two locations on earth with latitudes ϕ_1, ϕ_2 and longitude λ_1, λ_2 can be calculated by using the haversine formula as follows

$$d = Rc, \tag{4.6}$$

where R is earths radius (mean radius $= 6,371$ Km and constants a, c can be calculated as

$$a = \sin^2\left(\frac{\phi_1 - \phi_2}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_1 - \lambda_2}{2}\right) \tag{4.7}$$

$$c = 2\text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right) \tag{4.8}$$

The ϵ thresholded distance weight vector $\mathbf{w}_\epsilon(l(q))$ is used to obtain the *distance weighted topic score vector*. The vector $\mathbf{w}_\epsilon(l(q))$ is obtained from $w_\epsilon(l(q), l_i)$ as

$$\mathbf{w}_\epsilon(l(q)) := \begin{bmatrix} w_\epsilon(l(q), l_1) \\ \vdots \\ w_\epsilon(l(q), l_{|\mathcal{L}|}) \end{bmatrix}. \tag{4.9}$$

Using the weight vector $\mathbf{w}_\epsilon(l(q))$ the *distance weighted topic score vector* for user $u_i \in \mathcal{U}$ is obtained as follows

$$\mathbf{s}_i^\epsilon(l(q)) = \mathbf{S}_i \mathbf{w}_\epsilon(l(q)). \tag{4.10}$$

The *distance weighted topic score vector* contains the expertise of user u_i near the query location $l(q)$. In order to extract user's expertise in query topics $T(q)$, the distance weighted topic score expertise of user in all the topics in $T(q)$ is added to get an final estimate of user's local expertise at location $l(q)$ in topics $T(q)$. For this a vector $\mathbf{v}(T(q))$ is defined such that the i^{th} entry in v_i is given by,

$$v_i(T(q)) = 1, \quad \text{if } t_i \in T(q). \tag{4.11}$$

The final distance-weighted LSTS for user u_i for query tuple $[T(q), l(q)]$ is given by,

$$s_i^{final}(T(q), l(q), \epsilon) = \mathbf{v}(T(q))^T \mathbf{s}_i^{\epsilon}(l(q)) \tag{4.12}$$

$d_{min} = 100km$, $\alpha = 4.0$ and $g(\epsilon) = (1 + log_{10}(\epsilon))$ in the implementation of the local expert finding system (section 6).

4.2.2 Local Expert Perimeter

The provision of ϵ radius in the query for the problem of finding local experts was to give the user the ability to select the expertise focal window. If the query contains $\epsilon = 0$, i.e. the user would not like to consider any nearby locations to find a local expert, the system is likely to miss out on relevant results for locations which are closely connected.

An experiment was performed to find the local expert perimeter for 25 locations to find the default value of epsilon to be used with the particular locations in case $\epsilon = 0$ in query. The details of the experiment are provided in section 6.3. In the final model, the optimal epsilon radius as found for the 25 locations in the dataset from this experiment was used as the default distance for the queries for those locations.

The crux of the LSTS based approach with ϵ location preference is that, given a query which contains a set of topics and locations (due to ϵ), for every user we are looking at the query box region in their LSTSM to figure out their level of expertise, which is depicted in figure 4.1. This query box helps filter out a lot of users for whom the sub matrix is insignificant (empty), and once the potential experts have been identified, they can be ranked based on the actual scores in that query box/sub-matrix and other factors.

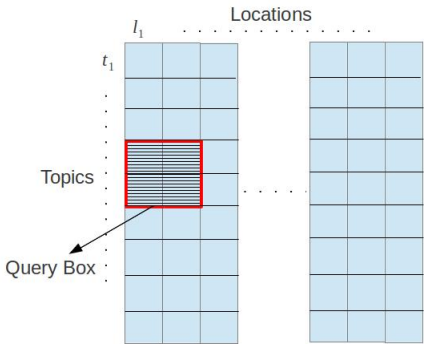

Figure 4.1: Relevant query box for query $T(q)$,$L(q)$

4.3 User's Topic-Location Entropy

One of the important assumptions of a local expert finding system as described in section 3.5 is that a potential expert can have knowledge about a topic across locations and maybe associated with same location on various topics. A user u_i's topic-location entropy for a given topic t is a measure of the distribution of his/her expertise in a topic across locations. Topic-location entropy $I(t, u_j)$ for topic $t \in \mathcal{T}$ and user $u \in \mathcal{U}$ is measured as,

$$I(t, u_j) = \sum_{i}^{nl} p(l_i|t, u_j) \log p(l_i|t, u_j) \tag{4.13}$$

where $nl = no.\ of\ locations\ the\ expert\ is\ associated\ with$ and $p(l_i|t, u_j)$ is the proportion of an expert's total LSTS in location l_i i.e.,

$$p(l_i|t, u_j) = \frac{s_j(t, l_i)}{\sum_i^{nl} s_j(t, l_i)} = \frac{LSTS(t, l_i, u_j)}{\sum_i^{nl} LSTS(t, l_i, u_j)} \tag{4.14}$$

For a query comprising of set of topics $T(q)$ and location $l(q)$, a user u_i's topic-location entropy is measured by taking the average of the topic-location entropies

for the set of topics $T(q)$ as given by,

$$I(T(q), u_j) = \frac{\sum_i I(t_i, u_j)}{|T(q)|} \tag{4.15}$$

where $t_i \in T(q)$.

In order to understand how a user's topic-location entropy affects expertise, an entropy based analysis of the data using top 20 users from the dataset in section 5 by number of user-location mentions for a small set of topics, $\mathcal{T}$ = "restaurant", "museum", "beer", "iphone", and "nba", across a set of 25 locations was done. Following table shows the average entropies of users for the topics across locations in decreasing order of entropy:

Topic	**Avg. of top 20 entropies**	**Avg. of top 10 entropies**	**Avg. of top 5 entropies**
restaurant	0.29	0.35	0.57
museum	0.334	0.386	0.56
beer	0.96	1.34	1.95
nba	1.404	1.417	1.698
iphone	1.42	1.42	1.572

Table 4.1: Average entropies of top 20 users for topics

The distribution of entropies of users for the given topics was plotted as a heatmap. We found that potential experts for topics like "restaurant", and "museum" are characterized by lower entropy scores with majority in range(0,0.6) and as depicted in 4.2 the distribution of entropies for experts for these topics are centered around the mean with a majority of the entropies within 1 standard deviation of the mean. Potential experts for topics like "nba" and "iphone" are characterized

by extremes very low (close to 0) or very high entropies (in the range(2,3)) with majority being on the higher side. As observed from the entropy heatmaps in the figure 4.2 the distribution of entropies is highly variable with very few entropies closer to the mean, suggesting the mean value was due to the extremes.

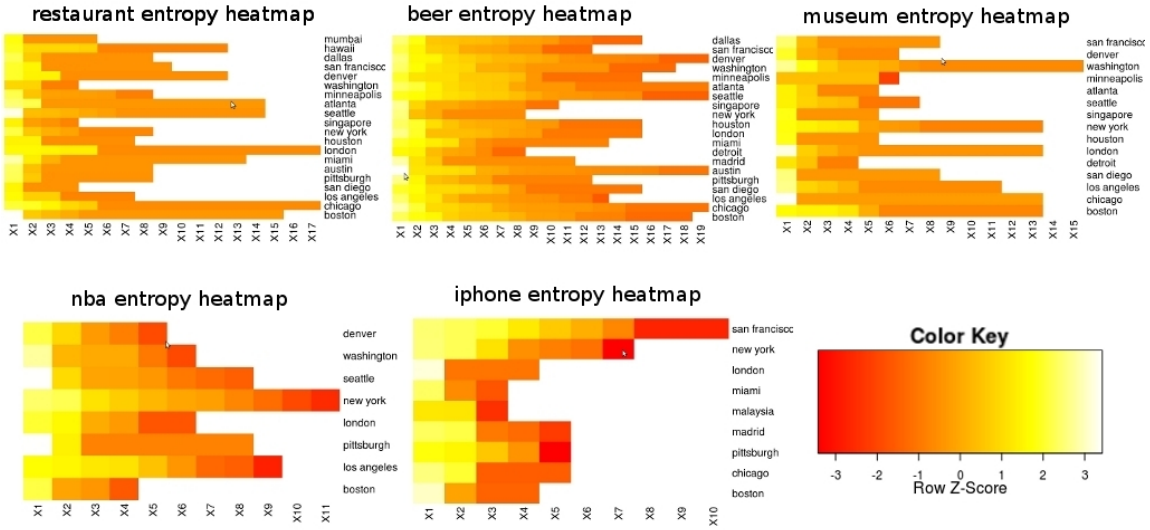

Figure 4.2: Entropy heatmap for topics

This analysis reveals that the topic-location entropies play an important role in the determination of local expertise of users. It depicts the characteristics of a topic and helps identify the localness of a topic, i.e. topics like restaurant and museum being local tend to have users who have low entropy with high LSTS and topics like "nba" and "iphone" have users who tend to have either very high or very low entropy. The topic "beer" in this case shows this behavior because some of the top users for topic "beer" were the local beer brands which are famous across locations (higher entropy).

An important factor while considering user's topic-location entropy is the concentration of the distribution of a user's knowledge geographically, i.e. if the distribution of a user's knowledge is concentrated in several locations close to each other v/s if the distribution is widely spread. One could use distance-weighted entropy scores to incorporate the effect of concentrated vs wide-spread distribution. This factor has not been explored and incorporated in the current system but will be an interesting avenue for future work.

When a query arrives, the local expert system first finds the set of potential local experts using the LSTSMs of the users and the topics and locations contained in the query and calculates the topic-location entropy for users for the set of topics in the query. These potential experts are then ordered using the ranking model described in 6.2. The model described in section 4 based on content expertise scores, crowdsourced endorsements of local topic expertise and topic-location entropy of users works well in practice as is depicted in the results in Chapter 7.

4.4 Aside: Probabilistic Interpretation

The distance-weighted location specific topic scores as described in section 4.2.1 present a very constrained view of the problem with a lot of assumptions such as, the users being equally likely to be experts, every location or topic being equally likely in the query, no topical hierarchy, etc. Hence the representation in section 4.2.1 is restrictive in the sense of the different ways in which we may want to combine the effect of the LSTSs for a user and incorporating other factors which might contribute to the selection of users (the distribution), etc.

This section describes a more general probabilistic approach to calculate the local topic expertise of a user. Suppose the query q contains a single topic t_j and location

l_k. Let $p(u_i|t_j, l_k)$ be defined as follows,

$$p(u_i|t_j, l_k) := \text{ probability that user } u_i \text{ is an expert in topic } t_j \text{ at location } l_k. \quad (4.16)$$

Using Bayes Theorem,

$$p(u_i|t_j, l_k) = \frac{p(t_j, l_k|u_i)p(u_i)}{p(t_j, l_k)} \quad (4.17)$$

Considering a simple case in which $p(u_i)$ is same for all the users, i.e. uniform distribution and it can be observed that $p(t_j, l_k)$ is same for all the users for a given query q we have the following relation

$$p(u_i|t_j, l_k) \propto p(t_j, l_k|u_i) \quad (4.18)$$

This shows that the probability that a user u_i is an expert in the topic t_j at location l_k, is proportional to the conditional probability which estimates the expertise/support of a user in topic t_j at location l_k for a given user, u_i from the data. By calculating the relative number of mentions of particular user u_i from its LSTSM one can get an estimate of $p(t_j, l_k|u_i)$. The LSTS described in the previous section exactly does this. This suggests the fundamental importance of LSTS as a measure of users expertise in local topic. In addition to this fundamental insight the probabilistic interpretation helps us to calculate LSTSM in the situation when it can't be explicitly calculated. Using chain rule,

$$p(t_j, l_k|u_i) = p(t_j|l_k, u_i)p(l_k|u_i) \quad (4.19)$$

$$= p(l_k|t_j, u_i)p(t_j|u_i) \quad (4.20)$$

[1]. Other sources of data, such as check-in data from various check-in services like Foursquare, Gowalla, etc., prior topic expertise knowledge about users, etc. can be used in place of user-location mention tweets to estimate these probabilities.

Generalizing equation (4.17), the probability that a user u_i, is an expert in a topic event $T(q)$, and a location event $L(l(q), \epsilon)$, where $l(q)$ is the location mentioned/inferred from the query and a given distance measure ϵ can be defined as follows:

$$p(u_i|T(q), L(l(q), \epsilon) = \frac{p(T(q), L(l(q), \epsilon)|u_i)p(u_i)}{p(T(q), L(l(q), \epsilon))} \tag{4.21}$$

The flexibility this representation provides is that, the local expert finding system has the freedom to compute the way topics are combined (equally likely, ANDed, ORed, weighted, dependent or independent, etc). A topic model capturing the distribution of user's knowledge in a topic across locations can also be used as a factor in defining the topic event. Similarly, the notion of location events which are derived using the epsilon measure, can be incorporated by modeling the decrease in the overall probability of a user being an expert with the distance from the query location using an exponential distribution which depends on ϵ and the monotonicity constant α, used in the formulation in equation (4.12). One can even choose a probability distribution for users, if there is some prior information/bias available about choosing a particular set of users over others, instead of the uniform distribution used in the formulation in equation (4.12). The probabilistic approach though very flexible and more intuitive, I have not developed a concrete understanding of the same and I include this as part of my future work in this thesis.

[1]These apply only in certain conditions such as domain experts, the type of topics (local vs global), etc.

5. DATASET AND ANALYSIS

5.1 Data Collection

I collected 12 months of Twitter data from Jan 2012 to Dec 2012 using the random sample public streaming API from Twitter providing geo-tagged tweets, from which I filtered and kept tweets which had one or more location mentions as part of their text. I used Stanford's Named Entity Recognizer [20] and a locations list[1] consisting of top 1300 locations around the world to identify location mentions in the tweets. The filtered location-mentions dataset contains 108 million tweets. From this dataset, I filtered tweets which contained one or more user mentions (approx. 49 million tweets). This dataset is referred to as D1 in the thesis.

Location names come with their own set of ambiguities. A few examples of the ambiguities include:

1. Interpretation of location names. I do not make any attempts to distinguish between instances where a mention of the location name "houston" could be in the context of "Houston Street" in downtown Manhattan or the city Houston, TX. The granularity of location names is restricted to the names of list of cities/states/countries and major location regions which are part of the fixed location set.

2. A location name maybe analogous to the common name of a person or some other entity and the Named Entity Recognizer may also fail to identify that.

3. Same location might have different names such as New York, NYC, etc. The system was restricted to the names in the locations list. There are several

[1]Location names were taken from the geonames data set and restricted to top 1300 locations from US and important locations in world, by population

dictionaries available which map same location names to ids and those could be used in a sophisticated implementation of the current system.

Due to lack of data and to minimize the effect of above mentioned ambiguities, the set of locations was reduced to top 501 locations of the world[2] and tweets corresponding to those locations were filtered from the D1 dataset. Reducing the set of locations also helped reduce the complexity of processing the information for solving the problem of finding local experts.

The final dataset referred to as the "user-location-mention" (ULM) dataset, comprised of approximately 24.4 million tweets, 8.5 million Twitter user handles and the top 501 world locations. Besides the tweets information in the ULM dataset, I collected user profile information on a need basis. The current dataset consists of around 351,395 user profiles so far. As user profile information is required only for ranking the relevant results and displaying the same found by the local expert finding system. The user profiles for the Twitter handles were fetched only for users identified as experts based on LSTS. The ULM dataset in the analysis of local topics (section 5.2.1) and constructing OLE, the system which finds local experts among Twitter users.

5.2 Data Analysis

As part of the initial analysis of the ULM dataset, I wanted to understand the distribution of tweets for the locations in the ULM dataset. Figure 5.1 shows the distribution of tweets. As the locations in the ULM dataset are particularly not the top 500 locations as per their proportion of tweets in the dataset, this limited the ability of OLE to find local experts for some of the locations in the dataset (the tail locations). The user mentions by location also follow the power law statistics as

[2]The location set was biased towards US locations comprising of top 200 cities from US and all US states

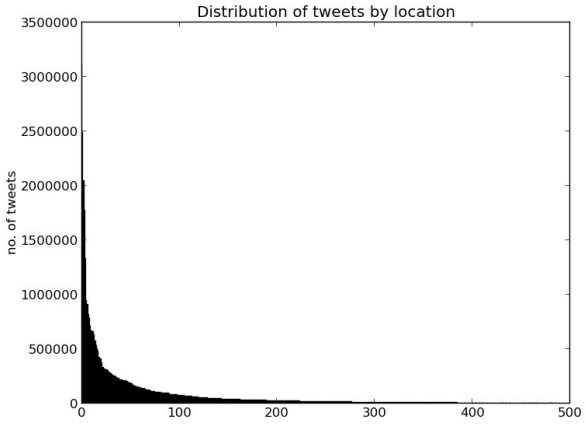

Figure 5.1: Distribution of tweets by location

shown in figure 5.1.

In an initial analysis of the dataset, the top 10 locations by the number of location mentions found (table 5.1) match the statistics about popularity of social networks in countries like Indonesia and Brazil[3]. We also found locations which are famous ports or tourist places or important centers of commerce.

Location	**No. of mentions**
Rome	1324226
Indonesia	809406
London	776266
Brazil	567668
Miami	402524
Chicago	251452
Boston	203248
Houston	178536
Florida	165555
New York	135415

Table 5.1: Top 10 Locations by Mentions

[3]http://royal.pingdom.com/2011/10/21/social-network-popularity-around-the-wo rld-in-2011/

Table 5.2 shows the top 5 users by number of location mentions. Basically these users are celebrities or have a very good reputation in their locations. The presence of foursquare which is a famous location service clearly shows the strength of the mentions based endorsement approach. Table 5.3 which is an extension of the previous analysis, shows the top user-location pairs by number of mentions and confirms the popularity of accounts in their home locations. These analysis showcase the power of crowd-sourcing and the correctness of the user-location associations.

User	**No. of mentions**
@Real_Liam_Payne	214520
@foursquare	193304
@Tweetnesian	82552
@dealprobe	16198
@London2012	10000

Table 5.2: Top 5 Users by Mentions

User	**Location**	**No. of mentions**
@Tweetnesian	Indonesia	77521
@RealMadrid	Madrid	29850
@Real_Liam_Payne	Italy	21778
@London2012	London	8722
@dealprobe	London	8324

Table 5.3: Top 5 User-Location Mention Pairs

5.2.1 Local Topics

An important analysis which aids the understanding of the problem of finding local experts and approach towards solving the same, is finding the topics of local

discussion. Basically it helps to find the subset of queries which are best suited for a local expert finding system and defines the limitations of the local expert finding system, OLE. I use GibbsLDA++ [21] implementation of the well-known topic-modeling algorithm, Latent Dirichlet Allocation (LDA) [22], to find the underlying local topics in tweets of the ULM dataset. First I process the ULM dataset tweets, by case desensitization, followed by removing any user and location mentions and removal of special characters. With the initial configuration of looking for k=50 topics, and α[4] and β[5] equal to 0.5 we find 42 interesting topics. Table 5.2.1 shows some of the topics found with the most describing words in the topic:

Topic	**Top words**
College	state, high, school, university, building, campus, library
Concerts	show, tickets, concert, tour, coming, sold, april, festival, dates
Fashion	blue, jersey, wear, shirt, store, fashion, nyc, hat, shoes, girls
Food	food, eat, bar, beer, taco, dinner, pizza, market, breakfast
Foursquare	mayor, ousted, hotel, college, office, square, building, library
Local crime	news, dead, police, story, shooting, fire, killed, theft, bomb, report
Moving in	live, nice, area, city, country, side, expensive, cheap, europe, living
Music	music, artist, album, cafe, rock, band, songs, record, pop, festival
Places	shopping, grand, mall, theatre, stadium, museum, amc, station
Politics	war, vote, obama, government, romney, bill, election, economy
Soccer	league, match, football, united, team, liverpool, arsenal, season
Travel	trip, international, flight, drive, heading, airport, car, bus, train
US sports	game, team, playing, football, nba, baseball, win, lebron, coach
Vacation	beach, hotel, club, lake, downtown, bay, island, hills, house, pool
Weather	weather, hot, nice, summer, cold, rain, sun, beautiful, lovely

Table 5.4: Local Topics

The topic analysis of user-location mention tweets finds topics which change in perspective with location in real life. The top words of topics found, closely represent

[4]Parameter of Dirichlet prior on per-document topic distributions
[5]Parameter of Dirichlet prior on per-topic word distributions

the local topics and clearly delineate the topic boundaries. Out of the 42 topics found, topics like technology, product reviews, which are discussed more in the context of the online world were not identified. Topics relating to food, travel, places, sports, etc. are clearly local. The ubiquity of spam was observed in local topics too with 3 different types of spam clusters. The topic "Foursquare" consisted entirely of tweets about who ousted whom as the "mayor" of a check-in location. The user-location mentions in the foursquare cluster inform about location association of a person and sometimes topic too, but it was a small set and not applicable to all users (hence was not considered for local topic expertise features). The most important learning from this topic analysis is the space of topics for which a local expert system would make sense.

Top topics for some of the locations around the world are shown in table 5.2.1:

Location	**Topics**
Chicago	Tickets, Food, Travel, Moving in, Sports, Location Discussion
Denver	Football, Sports, Tickets, Travel(help), Location Discussion
Houston	Tickets, Football, US sports, Travel(help), Local Crime
Iran	Politics, Local Crime, Travel, Sports, Location Discussion
India	Politics, Sports, Travel, Food, Local Crime
London	Tickets, Travel, Moving in, Sports, Weather
Los Angeles	Moving in, Food, Tickets, Fashion, Sports
New York	Tickets, Moving in, Food, Football, Weather
Singapore	Tickets, Moving in, Food, Travel
Washington	Moving in, Sports, Politics, Football, College

Table 5.5: Top topics by Location

A look at the top topics of the various locations shows interesting associations. Sports being a topic of discussion across locations, it shows that homes of football teams have that association but outside of United States "football" is not so famous

and other sports take its place. The observations also hold true in real life. Politics is a major theme of discussions on Twitter from India, Iran and Washington and moving-in in metropolitan cities like New York, Singapore and Chicago is a big deal in real life too. If we go one level deep (with topics within the sports topic), finding a "cricket" expert in United States would be a challenging issue whereas the same in India would be easier and it would be vice-versa when looking for a "baseball" expert in United States versus in India. This analysis helps know the strengths of the local associations of topics and also gives an idea about local queries where the expert system would be more effective.

The above analysis gives us valuable insights into the problem and the observations strengthen the intuition that location and a topic when viewed together mean different than when they are viewed independently. The solution to find local experts should be sensitive to these observations. The LDA based topic model can used in future versions of OLE, to infer higher-level topics from query and topics of expertise of a user using their user-location mention tweets.

6. OLE - ONLINE LOCAL EXPERTS

As part of this thesis, I developed a local expert finding system, OLE (for "Online Local Experts") based on the model proposed in section 4. This chapter describes the implementation of OLE in detail. OLE is an Apache Lucene[1] based search engine where Twitter users are indexed and scored using the LSTS approach. The ULM dataset was used to construct a file based "User Location Mention map" where each line represented a user document (to be indexed), consisting of endorsing tweets by location for a user. OLE has been implemented in a modular fashion where the indexer module indexes the user location mention map and any subsequent updates as they are available (online indexing). The search module which handles lucene search and ranking of results, has been implemented as a web service using the Bottle web framework[2]. The following subsection explains the implementation in detail.

6.1 Implementation Details

Figure 6.1 shows the architecture of OLE. Tweets from the Twitter public streaming API for a fixed interval of time are directed to the filter module which filters location mention tweets and forwards to the mapper modules which converts them to file based user location mention map. The lucene indexer module looks for new user location mention maps and indexes the users using the user-location mention tweets by topics (from keywords from the tweet content) and by location.

When a query is received, OLE needs to perform the following actions:

1. Find potential relevant experts from the query box based on LSTS.

[1]http://lucene.apache.org/
[2]http://bottlepy.org/docs/dev/

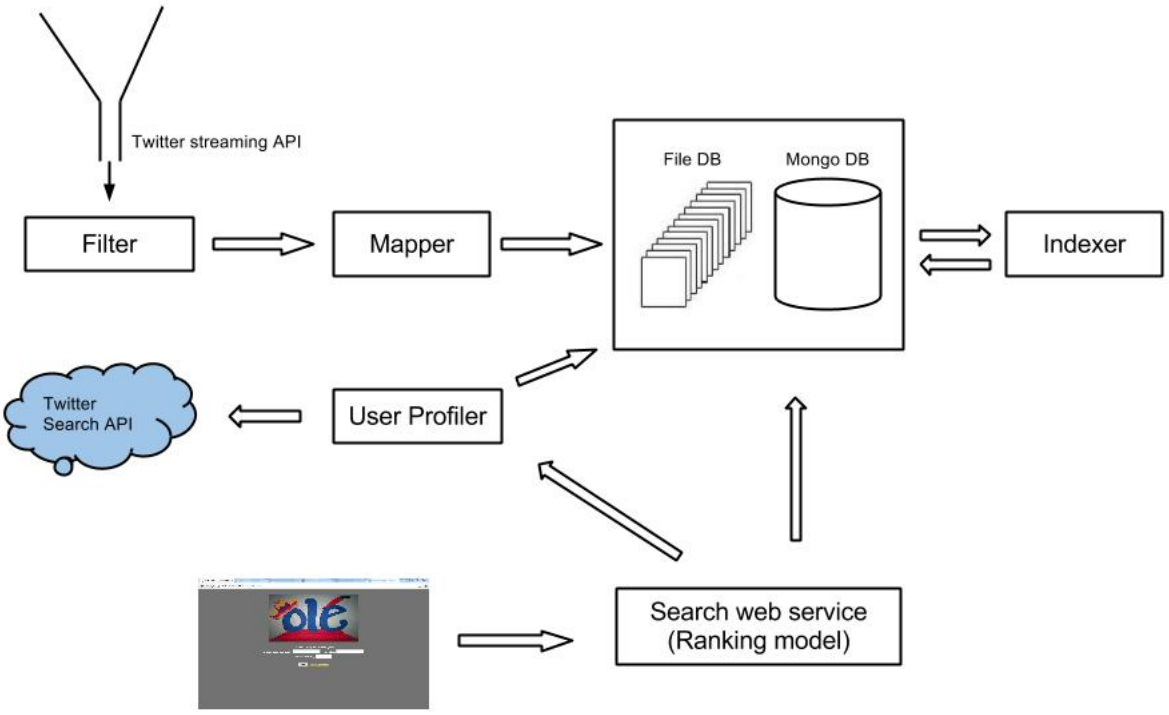

Figure 6.1: OLE - Architecture Diagram

2. Rank experts using proposed ranking model based on LSTS and profile based features.

Given a query, potential relevant experts are obtained by querying the OLE index. The user profiles are fetched from the profiles database and in case some profile is not available in the database it is fetched from the Twitter search API in real time. The potential experts returned are ranked as per the tf-idf based scoring over the user location mention tweets. These experts are then re-ranked using a custom ranking model described in 6.2.

The ranking model in section 6.2 is based on the set of features shown in table 6.1. An initial version of the model had every feature equally weighted. The final ranking model is based on random forest based point-wise learning-to-rank technique to learn the ranks using the following set of features.

Feature	**Description**
lsts	Location Specific Topic Score
ent	location entropy of user for the topic (4.3)
ls	lucene score
h	1 if home location is one of the query locations
des	1 is the user's description contains the query topic

Table 6.1: Features used for ranking

6.2 Learning to Rank

The dataset for learning to rank was prepared by automatically executing 750 local queries (30 topics across 25 locations) to find top 20 experts from a basic implementation of OLE based on a linear ranking model with equal weighting for all features and not considering the user's topic-location entropy. Out of the 750 queries, the results of 400 queries were manually labeled with three levels of expertise; 0 for "Not an Expert", 1 for "In-comprehensive Expert" and 2 for "Comprehensive Expert". In-comprehensive experts have only partial knowledge about the query whereas comprehensive experts are expected to have extensive knowledge about the query. This dataset is referred to as "labeled" dataset in the thesis. In order to find the best ranking model for OLE using the features mentioned in table 6.1, a random forest based point-wise learning-to-rank technique was used.

70% of the labeled dataset was randomly selected for training and remaining 30% for testing. It was found that the ranked model comprising of 5 trees in the random

forest produced a Root Mean Square Error (RMSE) of 0.7503 and the NDCG score of 0.879 for the test data. Table 6.2 shows the features in the order of the information gains. "ls", the lucene score feature which is based on the tf-idf and cosine similarity based measure of a user's ULM dataset tweets provides the maximum information gain, followed by entropy, and "lsts" scores. The feature gain due to expert topic-

Feature	**Feature Gain**
ls	35250.213
ent	19828.216
$lsts$	13670.955
des	8896.598
h	3345.577

Table 6.2: Feature gains in the random forest learning-to-rank model

location-entropy can be attributed to the characteristic entropy distributions for topics which were sensed by the random trees based model and incorporated in the model at the time of training.

Figures 6.2, 6.3 show that the learned ranking model performs better than the equally-weighted features based linear model. The mean ndcg score for linear ranking model is observed as 0.848 and mean ndcg for learned ranking model is observed as 0.908.

6.3 Local Expert Perimeter

The provision of ϵ radius in the query for the problem of finding local experts was to give the user the ability to select the expertise focal window. If a user of OLE was looking for a local "tourist places" expert in College Station, and OLE identifies a user from Bryan, who is much more relevant to the query, the system must consider that user as a potential local expert. In such scenarios, the system needs to rank

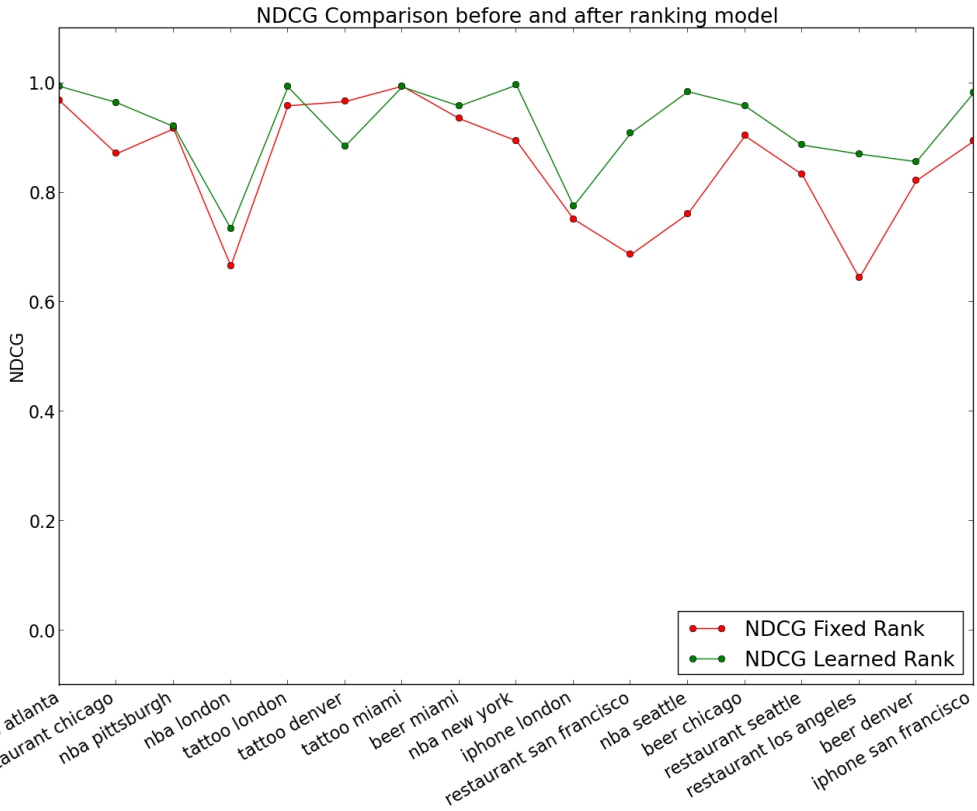

Figure 6.2: Comparison of NDCG scores, Linear ranking model vs Learned ranking model

experts for a given query including the nearby locations using an optimal epsilon for the location if the OLE user strictly advised the application to only look at results comprising of college station (by specifying $\epsilon = 0$) to avoid missing out on relevant results.

"Local Expert Perimeter" of a location is defined as the estimate of the area around a location, which when considered for finding local experts in that location, provide optimal results for query topics. In the following experiment, 6 topics per location were considered from the labeled dataset and their Mean Average Precision

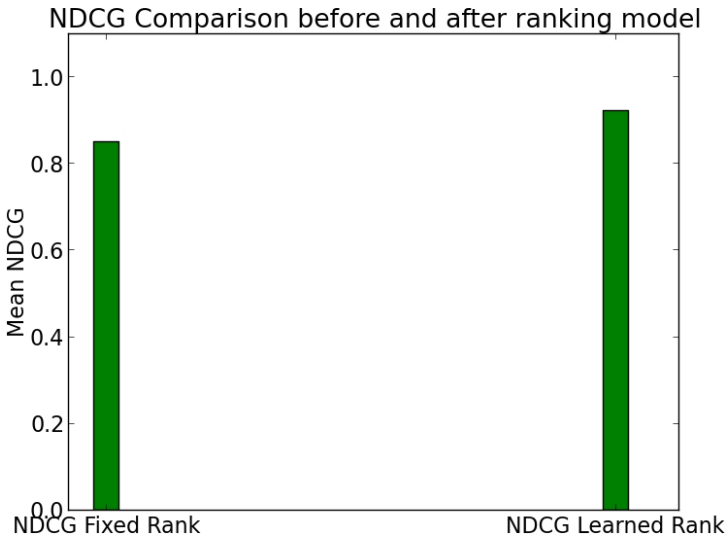

Figure 6.3: Comparison of Mean NDCG scores, Linear ranking model vs Learned ranking model

(MAP) scores were plotted for various epsilon values for the 25 locations. The best radius around a location was found where the MAP scores across topics were maximized.

Figure 6.4 illustrates mean average precision scores varying with distance ϵ from the query location for few locations. An interesting observation was that the best radius around a location for finding experts was found to coincide with the geographical limits of the city including the nearby important locations if any. In the final ranking model of OLE, the optimal ϵ radius was used as the default radius.

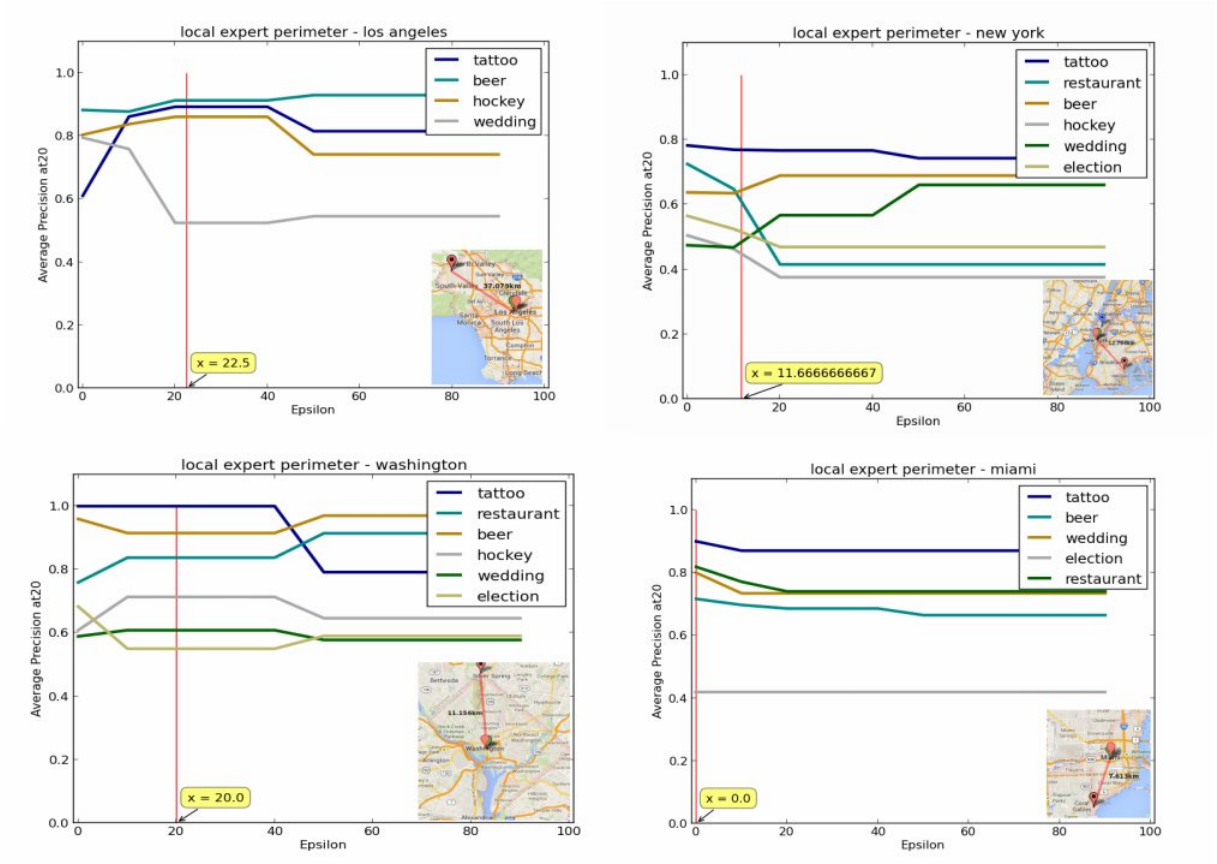

Figure 6.4: Local Expert Perimeter

7. EVALUATION

Most of the current efforts for finding experts in Twitter are centered around topic experts. In order to evaluate OLE, I performed the following set of experiments,

1. Compare OLE with some baseline methods implemented using existing topic expert systems
2. Evaluate performance of OLE
3. Compare OLE with an adaptation of an existing topic expert system to find local experts.

The following metrics were used for evaluating the results:

Metrics

- Mean Average Precision (MAP) score:

$$MAP = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{Q_j}\sum_{i=1}^{Q_j}P(doc_i), \tag{7.1}$$

  where Q_j is number of relevant documents for query j, N is number of queries, $P(doc_i)$ is precision at i^{th} relevant document.

- The Normalized Discounted Cumulative Gain (NDCG) [1]:

$$NDCG@n = Z_n\sum_{j=1}^{n}\frac{2^{c(j)}-1}{log(1+j)}, \tag{7.2}$$

  where Z_n is a normalization factor, n is the position for calculating NDCG, starting from 1 to number of results for the query, $c(j)$ is assigned rank level

[1] We consider NDCG@10, wherever an NDCG score is mentioned throughout the report

(scale of 1-3 in this case, 1 being irrelevant and 3 being excellent), to compare the results.

The following sections provide the details of the above mentioned experiments.

7.1 Comparison of OLE with Existing Methods for Finding Experts

The following experiment demonstrates why it makes sense to consider the problem of local experts differently from problem of finding topic experts in social media. I did some basic implementations for current methods for finding topic experts and ran 18 queries from the test set of section 6.2's data was used to compare OLE with baseline methods.

- In the first method experts are found using a topic expert system, built using @mentions, which is one of the features in paper by Pal el. al. [2]. The results obtained from this method are denominated as "topic".
- In the second method, topic experts were found using a topic expert system, called "Cognos", a user curated Twitter lists based topic expert system, exposed as a web-service by authors of [3]. The results obtained from this method are denominated as "topic_lists".‘
- In the third set, experts are found considering only the query location, i.e. experts who have a good know-how of the location. The @mentions method with topic being the location instead of the query topic was used to find experts. The results obtained from this method are denominated as "local".
- The results from ole are denominated as "ole".

Besides the MAP and NDCG scores, I also plotted the average fraction of comprehensive experts found by the methods. The results in figure 7.1 clearly show that

there is very little chance of a mere "topic" or "location know-how" expert, being a local expert. Even if the list-based topic expert method, is able to find local experts which seem to have an impact regarding the topic across many locations, they are not comprehensive experts. A good fraction of the experts found using the "local" method are comprehensive but as its mean average precision is small, the actual number of comprehensive experts would also be a small number.

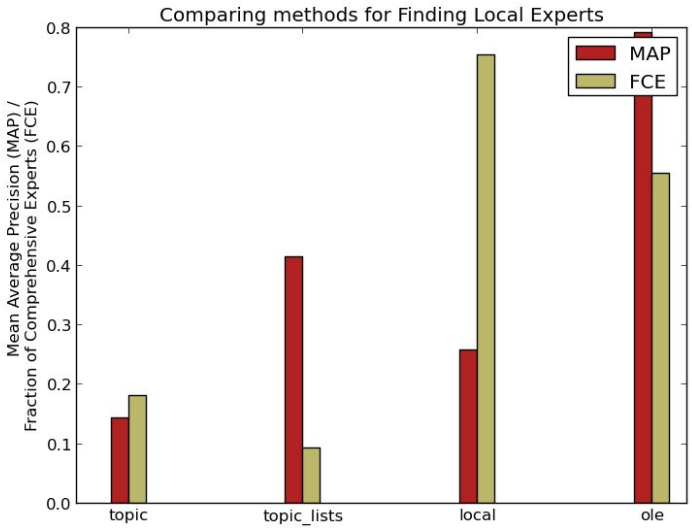

Figure 7.1: Comparison of OLE with topic expert methods

Metric	**topic**	**topic_lists**	**local**	**ole**
MAP	0.14	0.41	0.26	0.79
FCE	0.18	0.09	0.75	0.55

Table 7.1: Comparison of OLE with existing methods for finding experts

The baseline methods were based on the features of the existing topic expert systems. It was observed that they were not able to perform well for queries with geographical constraints and support the claim that a local expert finding system is important for local queries.

7.2 Experiment to Evaluate Performance of OLE

In order to evaluate the performance of OLE, I used Amazon Mechanical Turk, a system where workers work on "Human Intelligence Tasks" (HITS) by the requester and are paid by the terms of the requester. There were 90 HITs submitted, corresponding to 90 unique local queries with 10 results per query and 5 workers evaluating each result.

The HITs required mechanical turk workers to adjudge the level of expertise of an expert i.e. whether the user had "Extensive Local Expertise", "Some Local Expertise", "No Local Expertise" or there was "No evidence". Human evaluated responses often tend to be very subjective and hence I considered the majority rating and the average rating of the 5 responses per result.

The results of calculating NDCG and MAP scores on the two types of ratings are shown in figure 7.2 and table 7.2.

Metric	**Average Rating**	**Majority Rating**
MAP	0.856	0.786
NDCG	0.878	0.873

Table 7.2: Performance of OLE

Also to see how much agreement was there in the responses of the mechanical

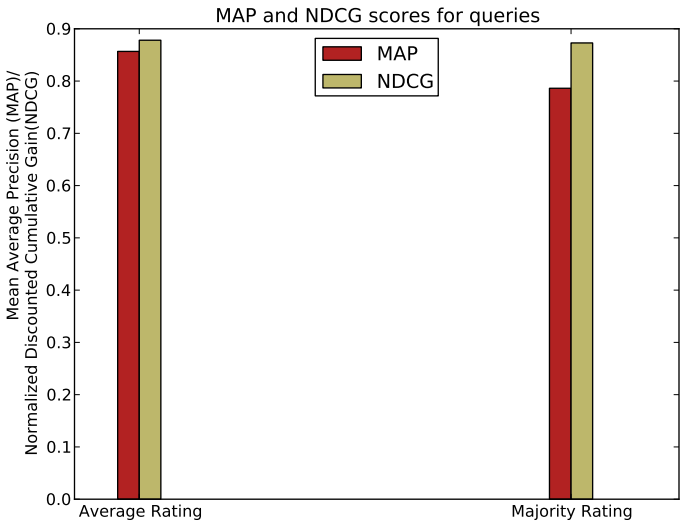

Figure 7.2: Performance of OLE

turk workers, I computed the Fleiss' Kappa statistic given by,

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{7.3}$$

where $1 - \bar{P}_e$ gives the degree of agreement attainable above chance, and $\bar{P} - \bar{P}_e$ gives the degree of agreement actually achieved above chance.

Table 7.3 shows the distribution of kappa values for the queries on mturk,

The κ values suggest that the results on mturk especially in terms of binary relevance judgments are in fair agreement for majority of the queries and hence believable.

From the results its evident that OLE perform reasonably well in finding local experts for a query in-spite of the simplicity of the approach used.

κ **agreement levels**	**% of queries (4 classes)**	**% of queries (2 classes)**
¡ 0 (Poor)	10.7%	4.6%
0.01 - 0.20 (Slight)	47.7%	23%
0.21 - 0.40 (Fair)	32.3%	43%
0.41 - 0.60 (Moderate)	9.23%	21.5%
0.61 - 0.80 (Substantial)	0%	7.7%
0.81 - 1.00 (Perfect)	0%	1.6%

Table 7.3: Distribution of κ values for the queries

7.3 Experiment to Compare LSTS Approach with Adapted List-Based Approach to Find Local Experts

For evaluating the proposed model for finding local experts a user study was conducted wherein OLE was compared to an existing topic expert finder on Twitter [3] adapted to obtain local experts. The app was launched as an internal website, and 30 users (graduate students from computer science department) signed up for the study. After a week, 85 queries had been executed on the system and labeled as relevant or otherwise by the users. We refer to this dataset as the "user_study" dataset.

In order to adapt the user-curated Twitter lists based approach for finding topic experts [3], the Cognos web service was queried with queries modified to combine both the topic and location as a single topic. The experts obtained by doing this comprised of experts who seem to have lists about the topic as well as the particular location associated with them. The result has been denominated as "cognos". The results obtained from OLE are denominated as "ole" and "ole_lr", where "ole" represents the equal feature weights based linear ranking model and "ole_lr" represents the final random forest based ranking model.

Figures 7.3, 7.4 show the precision @ k curves across queries for the three result

sets as mentioned above. The results show that "ole_lr" outperforms the other two methods most of the times. Infact in cases where OLE doesn't have enough no. of good experts for a particular query when compared to "cognos", "ole_lr" is optimal in the top results and has a higher fraction of comprehensive experts as shown in table 7.3 for the results shown in figure 7.4 for all the queries.

Query	**FCE - ole_lr**	**FCE - cognos**
beer - new york	4/5	5/9
fitness - houston	5/5	5/8
museum - california	2/6	0/8
street food - los angeles	4/6	0/9

Table 7.4: Fraction of Comprehensive Experts for Queries in Figure 7.4

The ranking from "cognos" cannot be optimized for the problem of local experts as it orders experts based on list counts of the lists which are combined to give the local experts from "cognos" in an unknown way.

The MAP and NDCG scores for the "user_study" dataset queries for the above mentioned methods are shown in 7.5. The results show that "ole_lr" performs better than "ole" and "cognos" for the problem of local expert finding.

Metric	**cognos**	**ole**	**ole_lr**
MAP	0.562	0.722	0.774
NDCG	0.754	0.852	0.881

Table 7.5: Comparison of methods for finding Local Experts

In above experiments I compare the methods on the basis of mean MAP and NDCG scores. I conducted a statistical test considering the null hypothesis that

the "cognos" and "ole_lr" mean scores were similar. Figure 7.6 shows the results of the Wilcoxon Rank Sums test (ndcg scores for the two methods had some outliers) and the p-value of 0.0094* supports our claim that "ole_lr" performs better than "cognos" at 95% confidence level.

As part of the user study I also asked users to mark which system performed better in terms of the quality of results. From the 78 valid queries executed on the system, according to users "ole_lr" did well in 45 queries, "cognos" did well in 21 queries and both did same (good or bad) in 11 queries.

The results show that the LSTS-based approach is much more effective than the adaptation of list-based topic expert finding method to find local experts and a unified approach is the way to go to solve the problem of finding local experts.

7.4 Qualitative Results - Sample Outputs

In this section, the top 5 results obtained for queries in OLE as well as the list-based adapted system for two sample queries are shown.

Twitter Handle	**Description**
hiddenboston	*"Founder of Boston's Hidden Restaurants, a restaurant site..."*
RestoWeekBoston	*"http://BostonChefs.com's Insider's Guide to Boston..."*
BostonMagazine	*"The best of Boston every day. Tweets by @kaitkylejohn..."*
BostonTweet	*"BostonTweet is all about life in Boston and things to do..."*
BostonEmpire	*"A 14,000 sq. ft. Asian Restaurant & Lounge located at Fan..."*

Table 7.6: Top 5 results for query: "restaurant" in "boston" (ole_lr)

Twitter Handle	Description
BeerAdvocate	*"Beer tweets by @JasonAlstrom @ToddAlstrom, BeerAdvocate..."*
BenJerrysTruck	*"We'll be giving out #OMGFreeBenJerrys in NYC from 7/2..."*
jbchang	*"pastry chef, bakery/restaurant owner, runner, besotted wife..."*
Mortons	*"Welcome to the official Morton's The Steakhouse Twitter..."*
formaggio	*"Specialty food store offering cheese, wine, charcuterie, olive..."*

Table 7.7: Top 5 results for query: "restaurant" in "boston" (cognos)

Twitter Handle	Description
jenniferconley	*"Working with North Texas startup companies at @gravity_..."*
Connectivehub	*"Plug into a Collaborative Business Community! We offer..."*
CoHabitat	*"The startup hub & coworking space in Uptown Dallas..."*
meyerdunlap	*"Recognizing the helpful, creative, and sometimes absurd..."*
launchDFW	*"Dallas - Fort Worth's Startup Community"*

Table 7.8: Top 5 results for query: "startup" in "dallas" (ole_lr)

Twitter Handle	Description
amuse	*"Co-Founder of Haul. Shopping with Glass. I start things..."*
RPMurphy	*"From a little place called Texas, know every Taylor Swift..."*
techwildcatters	*"Tech Wildcatters is a Forbes Top 10 early-stage tech startup..."*
launchDFW	*"Launch DFW Dallas - Fort Worth's Startup Community"*
alessiamosca	*"Deputato PD, XVI legislatura Candidata alla Camera dei..."*

Table 7.9: Top 5 results for query: "startup" in "dallas" (cognos)

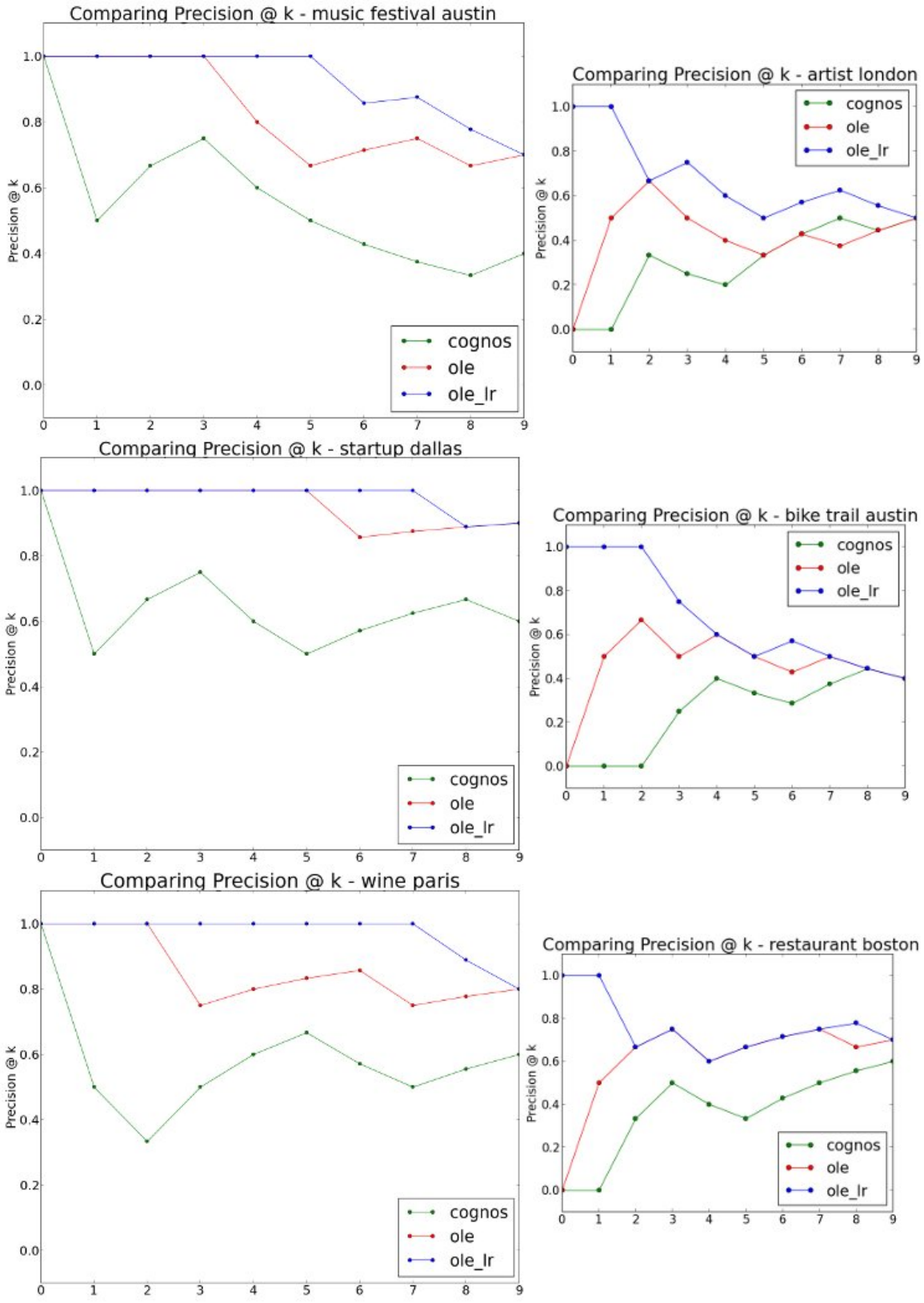

Figure 7.3: Precision @ k: OLE vs Cognos adapted for Local Experts

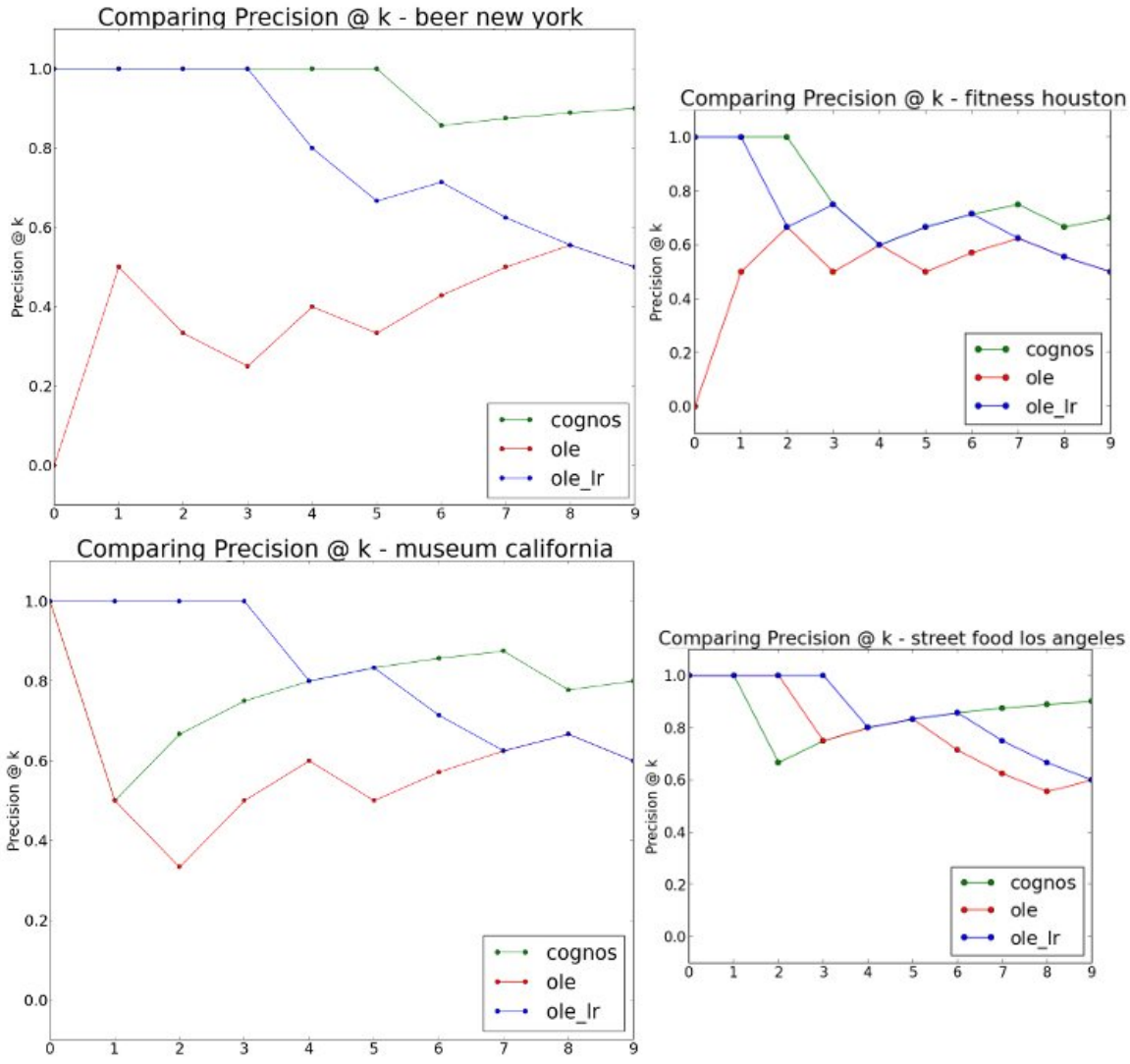

Figure 7.4: Precision @ k: OLE vs Cognos adapted for Local Experts (Interesting Results)

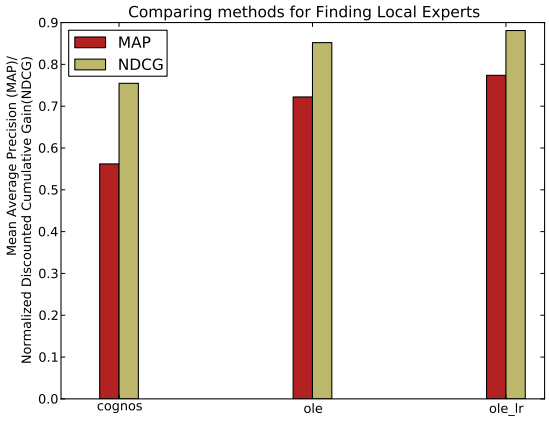

Figure 7.5: Comparison of methods for finding Local Experts

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
cognos	76	5110.50	5814.00	67.2434	-2.596
ole_lr	76	6517.50	5814.00	85.7566	2.596

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
6517.5	2.59592	0.0094*

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
6.7484	1	0.0094*

Nonparametric Comparisons For Each Pair Using Wilcoxon Method

q*	Alpha
1.95996	0.05

Level	- Level	Score Mean Difference	Std Err Dif	Z	p-Value	Hodges-Lehmann	Lower CL	Upper CL
ole_lr	cognos	18.50000	7.126563	2.595922	0.0094*	0.0549159	0.0086756	0.1015187

Figure 7.6: Results of Wilcoxon Rank Sum test for comparing the ndcg scores from "ole_lr" and "cognos"

8. CONCLUSION

The problem of finding local experts is a fairly new research topic. This work showcases an effective method for finding local experts where simple features all derived from the user-location mention tweets turn out to be valuable indicators of local expertise. I developed a system (OLE) based on these features. The simple design of OLE and using software which supports distributed implementation, it can easily scale to larger datasets.

The system currently suffers from limitations such as unable to, distinguish between location types (Houston Street in Manhattan vs city Houston), identify locations with alternative names as same, use methods such as "TwitterRank" or other topic expertise features for enhancing results or evaluating the results, etc. As part of future work I would like to work on overcoming these limitations and build in mechanisms to prevent spam and faulty results such as one time events causing lots of mentions of users in locations they are not connected with, etc. to build a better, more robust solution to the problem.

I would also like to experiment with distance-weighted entropies, tiered or topic-specific indexes and query expansion using LDA topic model to infer topics from queries and user-location mention tweets, to get more diverse results. Several other features such as user's activeness, authoritativeness on a topic and network features, can be used to improve the ranking algorithm for the experts found. I plan to continue working on some of the suggested improvements and launch the web application publicly to get better feedback about the system from users in real-life settings.

REFERENCES

[1] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[2] Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 45–54, New York, NY, USA, 2011. ACM.

[3] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Cognos: Crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 575–590, New York, NY, USA, 2012. ACM.

[4] M. S. Ackerman and T. W. Malone. Answer garden: A tool for growing organizational memory. In *Proceedings of the ACM SIGOIS and IEEE CS TC-OA Conference on Office Information Systems*, COCS '90, pages 31–39, New York, NY, USA, 1990. ACM.

[5] Kevin McNally, Michael P. O'Mahony, Maurice Coyle, Peter Briggs, and Barry Smyth. A case study of collaboration and reputation in social web search. *ACM Trans. Intell. Syst. Technol.*, 3(1):4:1–4:29, October 2011.

[6] C. Honey and S.C. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, pages 1–10, 2009.

[7] Kyle D. Dent and Sharoda A. Paul. Through the twitter glass: Detecting questions in micro-text. In *Analyzing Microtext*, 2011.

[8] Sharoda A. Paul, Lichan Hong, and Ed H. Chi. Is twitter a good place for asking questions? a characterization study. In *ICWSM*, 2011.

[9] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why?: A survey study of status message q&a behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1739–1748, New York, NY, USA, 2010. ACM.

[10] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. A comparison of information seeking using search engines and social networks. In *ICWSM*, 2010.

[11] Jiang Yang, Meredith Ringel Morris, Jaime Teevan, Lada A. Adamic, and Mark S. Ackerman. Culture matters: A survey study of social q&a behavior. In *ICWSM*, 2011.

[12] Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1225–1236, New York, NY, USA, 2013. ACM.

[13] Jeffrey Nichols and Jeon-Hyung Kang. Asking questions of targeted strangers on social networks. In *CSCW*, pages 999–1002, 2012.

[14] Michael Bernstein, Desney Tan, Greg Smith, Mary Czerwinski, and Eric Horvitz. Collabio: A game for annotating people within social networks. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, pages 97–100, New York, NY, USA, 2009. ACM.

[15] Damon Horowitz and Sepandar D. Kamvar. The anatomy of a large-scale social search engine. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 431–440, New York, NY, USA, 2010. ACM.

[16] F. Maxwell Harper, Daniel Moy, and Joseph A. Konstan. Facts or friends?: Distinguishing informational and conversational questions in social q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 759–768, New York, NY, USA, 2009. ACM.

[17] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Evgeniy Gabrilovich, Yoelle Maarek, Dan Pelleg, and Idan Szpektor. Predicting web searcher satisfaction with existing community-based answers. In *SIGIR*, pages 415–424, 2011.

[18] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: Everyone knows something. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 665–674, New York, NY, USA, 2008. ACM.

[19] Sharoda A. Paul, Lichan Hong, and Ed H. Chi. Who is authoritative? understanding reputation mechanisms in quora. *CoRR*, abs/1204.3724, 2012.

[20] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[21] Xuan-Hieu Phan and Cam-Tu Nguyen. Gibbslda++, 2007.

[22] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.