

Since its inception in the mid-90s, social networks have provided for a way for users to interact, reflecting of social networks or social relations among people, e.g., who share interests and/or activities. At the forefront of emerging trends in social networking sites is the concept of "real time" and "location based". So what makes location based social media services so important?

Privacy and Safety: Posting updates on location-based social networking websites and publishing your current location to the user, and can result in problems like personalized attacks by spammers, threat to your safety.

Trustworthiness of User Location: In certain scenarios, such as the political scenario of the Iran elections of 2009, it becomes important for organizations monitoring the data to be able to verify the location of a user.

Advertising and Marketing: Social networks connect people at low cost and can be beneficial for entrepreneurs and small businesses looking to expand their contact bases. These networks often act as a customer relationship management tool for companies selling products and services.

Having highlighted the importance of location of the user in social media, it is important to understand that it is not provided explicitly by the users for a number of reasons. Some of the users are concerned about their privacy and security; others do not find any incentive in sharing the location. Apart from this class of users who do not disclose their location, there are others who provide locations which are either incorrect or not machine readable or reveal just the state/country. The unstructured and free form of the text consisting of internet slang and incomplete sentences makes use of traditional Natural Language Processing and gazetteer based Data Mining approaches produce inaccurate results.

Research Background

I started out my research by focusing on location extraction from unstructured text, in particular *Craigslist* ads, for identification of places all the way up to the micro level. I developed a heuristic based ranking algorithm using Natural Language Processing Techniques and Gazetteer based Data Mining technique to identify the street level location of apartments listed on *Craigslist*. We developed an application that automatically generates geo-parses the *Craigslist* ads and shows them on a map. The integration of Point of Interest (POI) dataset and crime blotter makes apartment searching simpler and faster, helping the user to make a better decision.

After this we focused our research on location mining in Online Social Networks (OSNs). We began with the home city identification of a user on an OSN. Apart from the traditional content based location extraction techniques used in blogs, web pages, etc., in social networks we can exploit the social graph of a user to determine his location (and other attributes such as age, ethnicity, language, etc.). Our first published work in this area, *Tweethood*, used a k -nearest neighbor approach with variable depth to predict the location of the user purely from his social graph. We later employed label propagation, a semi supervised approach, to reduce the complexity of the algorithm.

Applications Developed

In the present world scenario, where the search engines wars are becoming fiercer than ever, it becomes necessary for each search engine to realize the intent of the user query to be able to provide him with more relevant search results. Amongst the various categories of search queries, a major portion is constituted by those having news intent. Seeing the tremendous growth of social media users, the spatial-temporal nature of the media can prove to be a very useful tool to improve the search quality. In our work, we examine the development of such a tool, called TWinner, that combines social media in improving the quality of web search and predicting whether the user is looking for news or not. We go one step beyond the previous research by mining Twitter messages, assigning weights to them and determining keywords that can be added to the search query to act as pointers to the existing search engine algorithms suggesting to it that the user is looking for news.

Current Research

Till this point we hadn't considered time as a factor for determining location. And as depicted from the U.S. Census Bureau's annual findings migration, especially in the youth, is a significant phenomenon and we know from surveys that almost a third of users on Twitter are aged between 25 and 34 years. So how do we know we know if the predicted location is the *current* location of the user or not? My current research aims to answer the question by performing graph clustering for identifying social cliques allowing us to implicitly consider time as a factor for prediction of user's most current location. A single user may be part of several social cliques. For example, John may have a group of friends he made while at high school, a group of his undergraduate friends and a clique of his office colleagues. Our work focuses on, first, determining these social groups, and then, doing purity based voting to determine the most current location of the user.

In addition to that, to increase the accuracy of the algorithm, we extend our work so that the algorithm now includes location extraction from the content of the messages posted by the user on the social networking site. I designed a technique that makes use of our prior work MapIt, and uses a combination of NLP and gazetteer based data mining technique for ranking location concepts. But, many social media mechanisms limit the size of a communication, both reducing the amount of information available, and increasing the amount of linguistic innovation employed by authors to express the maximum amount of information. Traditional text processing techniques, which rely on the basic assumption that the sentences are grammatically correct and complete, do not perform well in these cases. We thus make use of several crowd-sourced gazetteers and databases to understand a blend of abbreviations, slang and context specific terms; lacking in sufficient context and regularities and delivered with an indifferent approach to grammar and spelling.

The third important contribution of our current work is the identification of micro level locations from the content of a user's messages. By micro level locations, we mean specific Point of Interests (POIs) such as coffee shops, universities, places of worship, etc. Location applications are not just the next big thing in technology world, but also represent our changing culture. My work focuses on identifying and disambiguating any such POI that a user might have mentioned in this messages. There are several

problems that need to be resolved in order to do so. First, we need to disambiguate geo POIs from non-geo concepts. Second, even if we correctly pick out a POI concept, there may be several places by that name and we need to identify the correct POI the user was referring to. And finally, there are several POIs having the same name and of the same type but different locations (E.g. Starbucks) and we need to figure out the correct outlet/store. We shall be employing a series of context based spatio-temporal disambiguating methodologies for the identifying the correct POI concept.

Research Agenda

I plan to continue my research on location mining in online social networks for identifying multiple locations associated with a user. Currently, the algorithm returns a single location, the one with the highest score. But, we often see that there are quite a lot of people who divide their time between two cities (E.g. Consultants with a travelling job often spend Monday through Thursday at the client location). For identification of multiple locations, we need to perform an in-depth spatio-temporal analysis of his messages and we need to extend our current approaches for city level as well as micro location identification.

In addition to that, I will be developing the following geo-social tools to highlight the applicability of my research:

Location based Sentiment Mining

There has been some prior work done for mining the user sentiment from the language used in the messages. But, there are very few or no tools that perform a location based sentiment analysis. The presence of such a tool holds great potential and can be used by corporations for target marketing and/or by government agencies for security analysis and threat detection.

Effect of Location on Psychological Behavior

Another application that I plan to build would help us analyze the effect of location on the psychological behavior of people over time. E.g. If there is an earthquake in Los Angeles on a particular day, what effect does it have on the behavior of people living in first, the Los Angeles area, and second the remaining cities of US.

I believe the above research agenda gives me the opportunity to combine my knowledge and interest in the fields of data mining, geography, linguistics, and social psychology to understand behavior patterns of users on OSNs. The understanding of these patterns, would in turn, lead to the development of applications that would help me in learning, contributing, giving shape and making an impact in this upcoming field.