SPATIALLY AWARE RECOMMENDATION ENGINES

Animesh

Ayushi Dalmia

Tapati Middya

SPATIALLY AWARE RECOMMENDATION ENGINES

A PROJECT REPORT

Submitted By

Animesh (095181)
Ayushi Dalmia (095173)
Tapati Middya (095101)

Under the Supervision
Of

Dr. Prosenjit Gupta
Professor, Department of Computer Science & Engineering

Dr. Subhashis Majumder
Professor and H.O.D., Department of Computer Science and Engineering

in partial fulfilment for the award of the degree
of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING
HERITAGE INSTITUTE OF TECHNOLOGY, KOLKATA

November, 2012

DECLARATION

We certify that

a. The work contained in this report is original and has been done by us under the guidance of our supervisors.
b. The work has not been submitted to any other Institute for any degree or diploma.
c. We have followed the guidelines provided by the Institute in preparing the report.
d. We have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
e. Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due credit to them by citing them in the text of the report and giving their details in the references. Further, we have taken permission from the copyright owners of the sources, whenever necessary.

Signature of the Student

(ANIMESH)

(AYUSHI DALMIA)

(TAPATI MIDDYA)

CERTIFICATE

This is to certify that the dissertation entitled, "**Spatially Aware Recommendation Engines**" submitted by **Mr Animesh**, **Ms.Ayushi Dalmia, Ms Tapati Middya** to Heritage Institute of Technology, Kolkata, is a record of bonafide project work carried out by them, under our supervision and guidance and is worthy of consideration for the partial fulfilment of the award of the degree of Bachelor of Technology in Computer Science and Engineering of the Institute.

Prof. (Dr.) Prosenjit Gupta	Prof.(Dr.) Subhashis Majumder
Heritage Institute of Technology	Heritage Institute of Technology
Date:	Date :

ACKNOWLEDGEMENT

We would take this opportunity to thank Prof. B.B. Paira, Director, Heritage Institute of Technology for providing us with all the facilities to make our project work and of worth.

We will be thankful to Prof. (Dr.) Prosenjit Gupta and Prof.(Dr.) Subhashis Majumder for constantly supporting us and guiding us with their valuable insights.

We thank our teachers, faculty members and laboratory assisstants at the Heritage Institute of Technolgy for playing a pivotal and decisiveroleduring the development of the project. Last but not the least we thank all our friends for their cooperation and encouragement that they have bestowed upon us.

ABSTRACT

In the recent years the Web has undergone a tremendous growth in terms of both content and number of users. This has led to a serious problem of information overloading in which it is becoming difficult for the users to locate the authentic information in the given time.

Recommender Engines have been developed to address this problem,by guiding the users through the information and helping them find the right information. Traditional Recommender Engine sought to predict the 'rating' or 'preference' that a user would give to an item or social element they had not yet considered, using a model built from the characteristics of an item or the user's social environment.Spatially Aware Recommender Engine on the other hand produces a location-aware recommender system that uses location-based ratings to produce recommendations. The Recommendation Engine exploits the user based ratings through *user partitioning*, a technique that influences recommendations with ratings spatially close to querying users in a manner that maximizes system scalability while not sacrificing recommendation quality.

This project will present the design, implementation, testing and evaluation of a recommender system within the books domain. Our results show that the functionality of the recommender system is satisfactory.

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Background

The World Wide Web contains an enormous amount of information. According to June 30,2012 Internet World Stats the web contains over 8.67 billion indexed web pages and 2,405,510,175 internet users worldwide. This represented about 34.3% of the population worldwide and a 566.4% growth compared to 2000.

As the world moves into a new era of globalization, an increasing number of people are connecting to the Internet to conduct their own research and are given the ability to produce as well as consume the data accessed on an increasing number ofwebsites. This is leading to the overabundance of data in which users are increasingly facing problem in finding the correct data at the right time. This problem of information overloading has already caught attention of the researchers. They have tried to develop systems to provide personalized information which is suitable to the needs of the users. One such subclass of information filtering systems is known as Recommendation Engines.

Recommender engines have proved to help achieving this goal by using the opinions of a community of users to help individuals in the community more effectively identify content of interest from a potentially overwhelming set of choices. Two recommendation strategies that have come to dominate are content-based and collaborative filtering. Content-based filtering rely on rich content descriptions of the items that are being recommended, while collaborative filtering recommendations are motivated by the observation that we often look to our friends for recommendations

Recommender systems have also been deployed within commercial domains, for examplein e-commerce applications. A well-known example is Amazon, where a recommender system is used to help people find items they would like to purchase. Many online communities within the movie domain use recommender systems to gather user opinionson movies, and then produce recommendations based on these opinions.

1.2 Problem definition

This project shall focus on development and evaluation of a recommender system withinthe books domain. We propose a novel location-aware recommender system built specifically to produce high-quality location-based recommendations in an efficient manner. Evaluation will be done by assessing the system functionality and comparing the recommender precision.

1.3 Interpretation

Throughout the last years, a lot of applications have been developed that give location based ratings that include user based ratings. Such ratings motivate a new paradigm of spatially aware recommendation engines whereby the system takes into account the location of the users when producing recommendations. Existing recommendation engines which are triplets of users, items and ratings are not well designed to produce spatially aware recommendation engines. Thus we claim that the standard strategies are not always sufficient to reflect a person's preference, where preference often is location dependent. By integrating a mechanism for incorporating the spatial aspect in the recommender system, it may be possible to give recommendations that bettersuits a person's often varying preference.

1.4 Outline

Chapter 2 –Basic Knowledge and Theoretical Backgroundwhich primarily focuses on the basic ideas required to understand the problem.

Chapter 3 –Literature Overview which highlights the previous work done related to the problem discussed

Chapter 4 – Design proposes the design of our recommender system and its components.

Chapter 5 - Conclusion draws the conclusion of this thesis and recommends future work

Chapter 6-Bibliography

Chapter 2

Basic Knowledge and Theoretical Background

This chapter will introduce the problem that recommender systems are trying to solve, and different approaches for solving this problem. Present techniques used to improverecommender systems are also explained.

2.1 The World Wide Web

The World Wide Web (WWW or web) emerged in the early nineties. As the World Wide Web continues to grow at an exponential rate, the size and complexity of web pages grow along with it. Different techniques have been applied to develop systems that help users find the information they seek. These techniques belong to the fields in software technology called *information retrieval* and *information filtering*.

2.2 Recommender systems

We have 6.2 million customers; we should have 6.2 million stores. There should be the optimum store for each and every customer."-Jeff Bezos, CEO of Amazon.com.

In everyday life, when presented with a number of unfamiliar alternatives, people normally tend to ask friends for guidance, or to seek expert help by reading reviews inmagazines and newspapers. In the recent years, online *recommender systems* have begunto provide a technological proxy for this social recommendation process, in whichthey are used to either predict whether a particular user will like a particular item *(prediction)*, or to identify a set of *N* items that will be of interest to a certain user *(top-Nrecommendation)*.Recommender systems (RS) are used in a variety of applications. Examples areweb stores, online communities, and music players. Currently, people mostly tend toassociate recommender systems with e-commerce sites, where recommender systems areextensively used to suggest products to the customers and to provide customers withinformation to help them decide which products to purchase. Products can be basedon the top overall sellers on a site, on the demographics of the consumers, or on an analysis of the past buying behaviour of the consumers as a prediction for future buying behaviour.

2.3 Collaborative Filtering

One of the most commonly used techniques for developing recommendation engines is Collaborative Filtering. It has been used for years by the researchers for implementing recommender systems. Collaborative Filtering, also known as social information filtering is based on the principle of finding a subset of users who have similar taste and preferences to that of the active user, and offering recommendations based on that subset of users. The idea is that given an active user, u, compute her n similar users $\{u_1, u_2, \dots u_n\}$ and predict u's preference based on the preferences of $\{u_1, u_2, \dots u_n\}$. Similar users mean users who share the same kind of tastes and preferences over items. The basic idea behind collaborative filtering is that users who agreed on the past tend to agree on the future also. Collaborative Filtering works based on the following assumptions:

- Users with similar interest have common preferences and vice versa.
- Sufficiently huge number of user preferences is available.

Applications of Collaborative Filtering typically involve very large data sets. It has been successfully applied to diverse data sets including monitoring and sensing data, financial data and e-commerce and web application data. If we have a sufficiently large number of customer preferences and users with similar interest share common preferences, collaborative filtering can accurately predict user preferences. In many commercial applications, getting access to large set of user preference data is infeasible and therefore collaborative filtering based applications suffer from sparsity issues.

2.4 Quadtree

A **quadtree** is a tree data structure in which each internal node has exactly four children. Quadtrees are most often used to partition a two dimensional space by recursively subdividing it into four quadrants or regions. The regions may be square or rectangular, or may have arbitrary shapes. This data structure was named a quadtree by Raphael Finkel and J.L. Bentley in 1974. All forms of Quadtrees share some common features:

- They decompose space into adaptable cells

- Each cell (or bucket) has a maximum capacity. When maximum capacity is reached, the bucket splits
- The tree directory follows the spatial decomposition of the Quadtree.

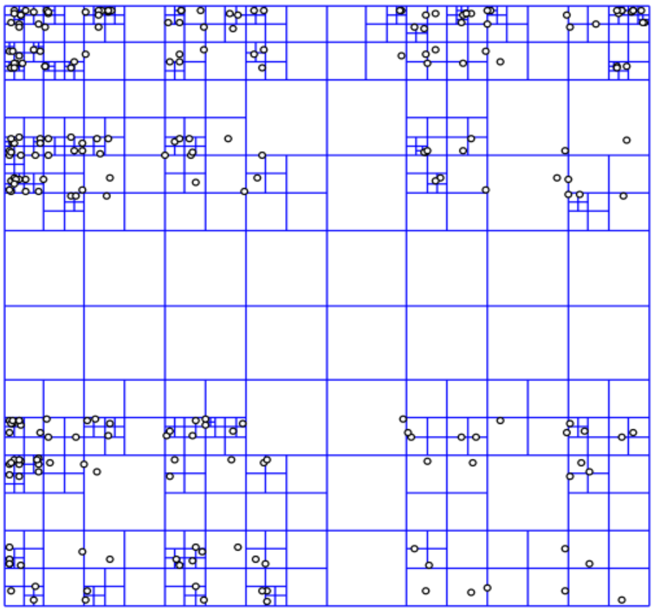

Figure 12.1:A region quadtree with point data

2.5 Spatial Autocorrelation

The first law of geography according to Waldo Tobler (1970) states that "Everything is related to everything else, but near things are more related than distant things." Analysis of spatial data shows that assumption about the independence of samples is generally false. In fact spatial data tends to be highly self-correlated. We often see that people with similar characteristics, occupations, and backgrounds tend to cluster together in the same neighbourhoods. The economies of a region tend to be similar. Changes in natural resources, wildlife and temperature vary gradually over space. In spatial statistics, the analysis of spatial data is called *Spatial Autocorrelation*. Spatial Autocorrelation is a statistical measure which examines the spatial ordering of geographic data. It deals with both attributes and locations of spatial features. Spatial autocorrelation determines whether adjacent or neighbouring values in the geographic data vary together, and if so, how. Spatial autocorrelation statistics allows us to measure interdependence in spatial distribution and to use statistical methods to test hypotheses about spatial interdependence. Spatial autocorrelation is affected by the scale of spatial pattern and thus the same spatial pattern at different scale may produce different spatial autocorrelation results. Two most commonly used measures of spatial autocorrelation are Geary's index (c) and Moran's index (I).

Geary's Index: The index measures the similarity of *i's* and *j's* attributes, c_{ij}, which can be calculated as follows:

$$c_{ij} = (z_i - z_j)^2$$

where z_i and z_j are the values of the attribute of interest for object i and j.

A locational similarity w_{ij} was used by Geary, and $w_{ij} = 1$ if i and j shared a common boundary and $w_{ij} = 0$ if not. Geary's index is expressed as follows

$$c = \frac{\sum_i \sum_j w_{ij} c_{ij}}{2 \sum_i \sum_j w_{ij} \sigma^2}$$

where σ^2 is the variance of the attribute z values, or

$$\sigma^2 = \frac{\sum_i (z_i - \bar{z})^2}{(n-1)} \quad \text{where} \quad \bar{z} = \frac{\sum_i^n z_i}{n}$$

If the value of c =1, the attributes are distributed independently of location. If c < 1, similar attributes coincide with similar locations and if c > 1, attributes and locations are dissimilar.

Moran's Index : Moran's index (I) gives us more logical result, with positive value implying that nearby areas tend to be similar in attributes, negative values implying they are dissimilar; a zero value indicates uncorrelated, independent and random arrangement of attribute values.

Moran's index (I) uses c_{ij} and w_{ij} as explained in Geary's index and is defined as follows:

$$c_{ij} = (z_i - \bar{z})(z_j - \bar{z})$$

and

$$I = \frac{\sum_i \sum_j w_{ij} c_{ij}}{s^2 \sum_i j \sum w_{ij}}$$

where s^2 denotes the sample variance,

$$s = \frac{\sum_i (z_i - \bar{z})^2}{n}$$

Both Geary's and Moran's indices were developed for area objects, but they may be applied to points, lines, and raster objects. For point objects, one can compute distances between pairs of points and use the inverse distance weighing to compute similarity.

Chapter 3

Literature Overview

In this section we bring into limelight the previous work done in the field of recommender system.

A lot of work has been done in the area of recommendation engines in the past. Several systems (using the movie domain) have been developed which attempted to recommend movies according to user preferences. Examples of online recommendation systems based on movie domain includes Jinni , MovieCritic , MovieLens, Netflix. They used collaborative filtering approach to recommend movies. Brunato, Battiti, Villani, and Delai combined user's current location along with the preferences for providing recommendation services. They developed a location-dependent recommender system which is based on a standard web browser. They determined in an automated way the relevance of a given URL in a given region through the collaboration of users of the system. The system recommends a specific URL to a user in a given location that considers where and how often it was accessed by the previous users. Brunato, Battiti used user's position as a relevant piece of information while selecting and ranking links of interest to the user. Their work *PILGRIM : A location broker and mobility-aware recommendation system*, designed a middleware layer, the location broker, that maintains a historic database of location and corresponding links used in the past. They calculated a preference metric when the current user asks for resources of interest. Google developed Hotpot, a recommendation engine for places, to make local recommendations more personal and relevant, by recommending places based on your ratings and ratings of your friend. It allows you to rate places and invite friends to share those ratings and as you rate sites via Hotpot, the service will recommend other similar places that you might also like. Li, Mi, Zhang, and Wu, integrated GPS into recommender system to create a location-aware recommender system. They explored the increasing demand of mobile commerce and developed a recommender system for tourism mobile commerce. The system can recommend attractions to the customer with the customer's rating of attractions and customer's sensitivity to location. Yang, Cheng, and Dia proposed a location-aware recommender system for mobile shopping. The system identified the customer's shopping needs and suggested vendors WebPages which includes offers and promotions depending on the location of the customer. Kuo, Chen, and Liang integrated recommendation technologies with location-based service (LBS) and proposed a location-based service recommendation model. Their work used a concept of preference adjustment (long-term and short-term) to improve the quality of recommendation. Espinoza, Persson, Sandin, Nystrom, Cacciatore, and Bylund studied social and navigational aspects of location based information system and proposed a system that allows users to participate as content providers to achieve a dynamic information space. Information filtering techniques were used to prevent information overload as users provide information on a mass-scale. Chen proposed a context-aware collaborative filtering system that uses the current context of a user to predict preferences. The system used collaborative filtering technique to find what other like-minded users would have done in the similar context to predict the preferences for the active user. It can predict user's behaviour in different context without the user actively defining it.

Chapter 4

Proposed Algorithm

4.1 Outline of the algorithm

In this chapter we will describe the problem definition in depth followed by the proposed algorithm which we have opted for performing our experiments.

In this work, we propose a spatially-aware recommender system using Quadtree. Although the decomposition algorithm and the recommendation engine are general in nature, we have tested our ideas on the books dataset. The proposed work will use this dataset to recommend books to users that are likely to be preferred by them. While recommending our primary focus will be on the location of the user. We use latitude and longitude to identify the user's location. We+ try to explore the concept of Spatial Autocorrelation. Quadtree is used to dissect the user's space with respect to location. Spatial autocorrelation index (Geary's index) is applied to measure correlation among the users in the regions resulting from Quadtree decomposition. Space decomposition partitions the entire user's space into smaller regions and we will provide recommendation to users in those regions. Our work tries to find the presence of correlation in the regions and then recommend books with a view that the suggested books will be liked by the user. Collaborative filtering algorithms will be used for recommendation. The system will offer recommendations to a new user according to his choices and the preferences of other users who share the same location with the current user. Similarity metrics such as Pearson's correlation coefficient is used to compute the similarly among the users.

4.2 Decomposition Algorithm

In this work, we use a Quadtree based approach for space decomposition. Space partitioning is done on the basis of latitudes and longitudes. User data file in the dataset has information about the user and their locations. Our work tries to decompose this user space. Location is represented by latitudes and longitudes. We construct the quadtree datastructure using these latitudes and longitudes. We try computing the spatial autocorrelation indices in the entire data space. Geary's index is used to measure the correlation. We consider the movie ratings as the parameter for computing the attribute similarity (c_{ij}) among users. Locational similarity (w_{ij}), is measured by computing the distance between the pairs of user location and use the inverse of the distance to compute similarity. One book is considered at a time and its correlation is measured within the entire region. If the index value calculated lie within a predefined range, then we can infer that spatial autocorrelation exists in the data and the users share similar taste and preferences for items and our recommendation will be better otherwise we split the region based on the latitude and longitude into four regions. The scheme is detailed in its algorithmic form as follows:

Quadtree Decomposition Algorithm(The preprocessing step)

Step 1: Represent location of each user as coordinates (longitude-latitude).

Step 2: Find the spatial correlation of the entire region.

Step 3: If the correlation lies within a specified range then the users share similar preferences otherwise we split the region into four regions depending upon the latitude and longitude.

Step 4: Repeat steps 2 to 4 for each region.

The output of the Quadtree Decomposition algorithm is that every user is assigned a particular region based on the correlation.

4.4 Recommendation Algorithm

Once the quadtree datastructure has been constructed, regions defined, and spatial autocorrelation measured, we can start our recommendation algorithm. Recommendations can be provided to a naïve user or an existing user. For a new user we identify the location (latitude and longitude) and accordingly map the user to its destined region. Collaborative filtering technique is used to recommend items of interest to the user. Similarity in the user's preferences in a region is measured by Pearson's correlation coefficient. During preprocessing, we calculate Pearson's coefficient value for every pair of users in the region. For a new user, we identify a subset of 10 top rated books to form a book set. We choose 5 books randomly from this set to recommend books to the new user. For an existing user A, we locate the top-5 users who have the highest correlation of ratings with A. We then consider 10 top rated books from each of these 5 users and from this we choose 5 books depending upon the local average rating of these books. The algorithm is briefly described below:

Algorithm Recommend_books(Online step)

Step 1: Select a user for recommendation.

Step 2: Identify the location (latitude and longitude) of the user.

Step 3: Map the user in the quadtree datastructure according to his/her location.

Step 4: Find a subset of users who share the similar preferences for items with the active user.

Step 5: Recommend items based on the preferences of the above subset of users (Collaborative Filtering). For a new user, identify a subset of top rated movies and select randomly.

The output of the recommendation algorithm is a set of books recommended to the user depending upon the algorithm

Chapter 5

Conclusion and Future Work

8.1 ACHIEVEMENTS

This chapter summarizes our dissertation by first concluding whether our achievements have fulfilled the problem definition. Then, future work is presented.

The problem definition from Section 1.2 is stated below:

This project shall focus on development and evaluation of a recommender system withinthe books domain. We propose a novel location-aware recommender system built specifically to produce high-quality location-based recommendations in an efficient manner. Evaluation will be done by assessing the system functionality and comparing the recommender precision.

Our conclusion from the experiment can be summarized as follows:

1. Real-time feedback increases recommender precision. As we have spitted the preprocessing work and the online work it has lead to faster query results without loss in the quality of recommendation.

2. User location has proved to be an important aspect. Our experiment has shown that by integrating a mechanism work for adding location into therecommender system, it has been possible to produce recommendations that better suits the users preferences.

8.2 FUTURE WORK

In this section we present issues that should be explored to improve our recommender system.

1. Binary Trees: Instead of using quadtree the usage of binary tree might help in improving the split. This is because using a quadtree forces us to split into four equal parts while using a binary tree allows us to divide the region depending upon the data points.

2. Comparison with other recommender systems: We will be comparing the quality of recommendations with other recommendation models.

Chapter 6

Bibliography

[1] Bhasker, B. and Srikumar, K., 2010. *Recommender Systems in e-Commerce,* Tata McGraw Hill.

[2] Herlocker, J. L., Konstan, J.A., Riedl, J., 2000. Explaining Collaborative Filtering Recommendations. *CSCW '00 Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Philadelphia, USA, December 02-06, 2000).

[3] Lo, C.P., Yeung, A.K.W., 2007. *Concepts and Techniques of Geographic Information Systems*, Prentice Hall.

[4] Brunato, M. and Battiti, R., 2003. PILGRAM: A Location Broker and Mobility-Aware Recommendation System. *In proceedings of IEEE PerComp2003* (March 23-26, 2003), 265-272.

[5] Kuo, M., Chen, L., Liang C., 2009. Building and Evaluating a Location-Based Service Recommendation System with Preference Adjustment Mechanism.Expert Systems *with Application: An International Journal* (Volume 36, Issue 2, Part 2, March, 2009), 3543-3554.

[6] Espinoza, F., Persson, P., Sandin, A., Nystrom, H., Cacciatore, E. and Bylund, A., 2001. GeoNotes : Social and Navigational Aspects of Location-Based Information Systems. *Lecture Notes in Computer Science* (Volume 2201/2001, 2001), 2-17.

[7] EE PerComp2003 ([16] Chen, A., 2005. Context-Aware Collaborative Filtering System:

Predicting the User's Preference in the Ubiquitous Computing Environment. Lecture

Notes in Computer Science (Volume 3479/2005, 2005), 75-81.

[8] http://www.movable-type.co.uk/scripts/latlong.html.