CSCI 5980 Homework 2

Trend

GROUP 1

Brief summary of the narrative and presentation
* Group 1 summarized chapter 10 of the textbook, pertaining to time as it relates to spatial computing. They broke the chapter down into three main categories of temporal representations: snapshots, lifetimes, and events/actions/processes.
* From the Encyclopedia of GIS, they discussed the article on Moving Object Languages. They describe two perspectives, the location management perspective and the spatio-temporal perspective data perspective. They also include a useful and helpful table of some example operations for a Moving Object Language.

New information learned from this work relative to the information available in corresponding textbook-chapter
* I learned about how time is modeled in databases at a high level, as well as a few concrete examples of methodologies and operations on spatial temporal data.

Constructive suggestions to improve the work
* 1. Slide 2 has a misspelling: temporally (instead of temporaly).
* 2. The last few slides are relatively cluttered compared to the first few. If possible, you may want to balance this out a bit more.
* 3. Expanding upon point 2 above, depending upon how much time you are given to present, you may want to reduce either the breadth or depth of topics covered towards the end in order to avoid going over time. I'm sure that simply rehearsing the presentation will help you gauge your time.

GROUP 6

Brief summary of the narrative and presentation
* Group 6 summarized chapter 6 of the a spatial databases textbook, specifically with spatial networks.
* In addition, they reviewed 3 articles from Professor Shekhar's research group which incorporated a time component into spatial networks.

New information learned from this work relative to the information available in corresponding textbook-chapter
* I learned details relevant to spatial networks, and more specifically, novel methods and data representations for the incorporation of a temporal aspect to spatial networks.

Constructive suggestions to improve the work
* 1. The slides discuss many topics, and end with the representation. Are there algorithms in the literature you read which use this representation? A brief overview of such algorithms might be good to add to the end of the slides.
* 2. This is a very minor point: in the applications slide, both road networks and air networks are given as applications. These seem similar. Are there other applications that could be stated here? If not, this is absolutely fine as-is.
* 3. In slide 5, what are “RA” languages? This term is stated but not introduced or defined.
Project
Codified and Tacit Sources
After consultation with Prof Shekhar we realized that our goal for this project is to and geocode non geocoded tweets. Only around 1% of tweets are geo tagged making it difficult to know where tweets originate. This is an important question because many local businesses try to use social media like twitter to reach local customers. If a business could focus on users or twitter that are close to their business, they could have more successful social media marketing. Using information in the tweets, and comparing that information with outside information it will be possible to approximate where a tweet came from (or should have come from). Doing this will require 4 main steps: accessing the data from twitter, storing the data, comparing the information in the tweet to information that has geographic information, and visualizing the results. A final step would be to verify the accuracy of these predictions by predicting the location of tweets that are already geo-tagged and comparing the prediction to the actual data. The next part of this write up will look at each step and look for codified and tacit sources that can help us with this project. I hope this is okay to combine parts a and b of the homework in one dialogue (sometimes monologue).

Accessing Twitter Data
To access twitter data we will be using the work that has been done by twitter to make their data very accessible, and a client in a scripting language. Current experimentation has taken place with tweepy, a package for python to easily work with twitter streams, users, and user timelines.

Codified
Expert knowledge or sources are not needed in accessing the twitter data itself, because the data itself is very simple and well presented through the api. The work that will require knowledge of twitter will be filtering important results (assuming that we can not store all data) and storage of the data. By focusing on the location of tweets it is very clear we will be filtering for tweets that have location attached. According to twitter api tweets can have location in two different ways, location preset by user, and location where the tweet was sent from. This is the main piece of codified knowledge that is needed for accessing tweets. The api also documentation also goes through all of the returned data in a data stream. The possible data is: tweet, user, entity, and place. Entity and place can be thought of as metadata, with the bulk of the data being associated with user and tweet. An entity stores hashtags and mentions of a tweet, and where those hashtags and mentions occur (index). A twitter place is a simplified geojson representation of a place. For example there is a place entry for the united states, that includes its minimum bounding box and generalized outline. The api is very clear, and using a scripting language client along with understanding of the api should make implementation fast and easy.

Tweepy documentation: Tweepy is being discussed because it is an easy implementation of the twitter api for python. The documentation provides in depth information how to authorize and account, and how to create a basic search or stream. The code itself has almost no documentation outside of this leaving the user to dig through all of the scripts to make a small change. This lack of documentation may lead to us looking at the other api clients that have better documentation.
Streaming vs. Search api: Twitter has multiple different ways to access the data. The two main approaches are the streaming and the search api. The streaming api allows for retrieval of real time events, and places no limit on number of returned data, while the searching api allows for retrieval of up to 1500 tweets to a query and allows access to historical tweets.

Tacit
tweepy, twitter api, python
Although twitter can be accessed with a number of programs, we have knowledge of the syntax of python, which will allow us to spend less time on learning how a program works to access tweets. Information in the twitter api will allow us to know what data is stored where and how to access the information of interest. There are very few question to ask about python or the twitter api, because both are very powerful, and we don't have many options. The only real question that should be answered about the twitter api is the limitations of the search api and the streaming api. we could also look at other scripting languages that people use to see if they have better performance than python using twitter.

For tweepy there are more questions.
Is this the best twitter client for python? Is this good enough?
What are the limitations? in terms of speed and number of requests?
What causes it to fail?
How does the speed of tweepy compare to other twitter clients?
Are there other twitter clients that have better documentation?
Answering these questions will require experimentation with tweepy and other twitter clients

Storing Twitter Data
To store the twitter data we will need to think about the important features of a tweet, and practices for storing data of this type. We also need to consider what parts of the tweet will be used in later steps so access to those parts are as quick and easy as possible. Constructing the data store will require a massive amount of tacit and codified information.

Codified
Tweets or users that are geotagged will be spatiotemporal, so considerations for spatiotemporal database design must be considered (insert sources here). We will first look at broad spatial databases and spatio-temporal data before looking at specific social media database implementations.
When storing spatial data in a database it is important to understand how that database stores and works with spatial data. Through the labs we have learned a little bit about the ability of Oracle spatial, and readings on Oracle Spatial along with the Oracle Spatial help will help us understand how to optimize the storage of spatial data

Twitter discusses on its engineering blog where and how it stores data internally. They use flockdb to store data, along with Cassandra and Hadoop. Hadoop is a key-value store that allows for distributed computing. Twitter stores its data and Hadoop and uses Pig to write queries on top of hadoop. Although we will probably not implement a solution like this, it is important to read about how twitter works, and how we could scale up in the future if need be. Twitter has taken many different approaches to modeling their data as their size has increased. Using a graph database, even if it is on top of a relational database, makes sense when dealing with social media data, but we may not need to follow all of the things that they do because much of what they do is based off of the size of their data.
**FlockDb**: Is a graph database used by twitter. Most graph databases are worried about graph traversal. FlockDB is only worried about relationships. If twitter were concerned about distance between users (in network space) then they would also have to worry about graph traversal. Because they do not need to worry about traversal they are allowed to make an optimized graph database. Some people claim that this makes FlockDB not actually a graph database. The limitations of FlockDB are the fact that it is an unknown to us, and apparently is not very well tested or used outside of twitter.

**Relational Model For Twitter**: This paper looks at a simple way to create a relational database using twitter data. Although the structure could be changed, this would more or less be the structure that would be used for creating a relational database of the data

Many of the sources above talk about the structure needed at twitter to store all of the data in real time. This forces twitter to use distributed computing and think creatively about how data is stored and accessed across many servers. Although it is good to read about how twitter operates, and how twitter conceptualizes its data, much of this information is not needed for a smaller scale database storing only a small percentage of tweets, and with no real time obligations.

**Tacit**

Prof Shekhar is an expert in spatial databases and and spatial data mining so the knowledge that he and his team have will be a very necessary tacit source for how all steps of our process.

**Questions**

How have people stored data for similar projects?

What are the limitations to using NoSql databases, which seem to be a common database for social media?

If we are to use a relational database is Oracle and Oracle spatial the best way to go? Should others like PostGres/ PostGIS or MySQL be considered?

How can we use spatial statistics inside of a spatial database?

What are the advantages of something like flockDb, or other graph databases? Can any of these work with complex or simple spatial data?

Project Cascade uses MongoDB. what are the advantages to using a document oriented database like this? What are the disadvantages?

We both have experience working with spatial and aspatial data in a relational database, which will speed up the time to create this database. We also have experience with data modeling (from comp sci 5707) which will allow us to create a well formed efficient database to store our data in. On this topic we will need to read documentation on databases, after direction from Prof Shekhar.

**Predicting Location**

**Codified**

Each tweet object in a stream contains information about the user that sent the tweet, and information about the tweet itself. we will parse through both pieces of information looking at the content of many attributes to approximate location. There are many pieces of codified information that will be important in predicting location.

Time Zone Map: Each user has a home time zone, so using a time zone map will allow us to figure out which timezone the user is in
A gazetteer can be used to reference geographic names mentioned in tweets to geographic locations in the gazetteer tweet data:

**Linguistics**

Working with linguistics inside of GIS is a growing field. It seems as if the main source for this data is Ethnologue. Ethnologue has an online database that goes through where different languages, and different dialects of those languages are spoken. This is a tool that costs money so we might not be able to use it, but we may be able to find some way to work around this cost. It is tough for us to evaluate the shortcomings of Ethnologue currently because much of the information about it is locked, but interviewing a Prof would probably allow us to learn more about available software

**Linguistics inside ArcGIS**: A simple little article about how linguistics and GIS has evolved over time, and goes through shortcomings. Interesting but does not provide very much detail in terms of how to use GIS and linguistics together.

**Modern Language Association**: Provides data from the census about the language spoken in an easier to access format. Could be useful if looking at more languages than just English. This data is limited to the United States, and does not have further breakdowns within language, but could still prove to be a useful data set.

We should also dig through the readings, and other sources to find codified sources on the statistics required to allow us to compare tweets to location.

**Tacit**

Being a team with a computer science major and a GIS student we have a lot of the tacit knowledge within our team to conquer many of the main issues of the project.

Outside data: From the GIS end, the knowledge of where and how major gis data is stored will allow us to work with outside data, like the gazetteer, with ease. As an employee of the spatial core of the population center I have over a year experience working with geographic census data, putting at our disposal a wealth of knowledge of how to work with census data, and other demographic data. Although this is not a demographic project at heart, the wealth of information that the census releases will allow us to use demographic and geographic information to improve our predictions.

Accessing data: Both of us have extensive knowledge, partially from this course, in how to store and query spatial and aspatial data from a relational database. Being able to access data from our well formed database will allow us to find how twitter data is related to geography, and with statistics, possibly predict the location of that tweet.

Statistics: Being able to predict a location or a tweet depends on correlating information in and about a tweet to other information that has geographic information. All steps From the GIS perspective spatial statistics require different statistics because of spatial autocorrelation. Tweepy. Prof Steve Mansen in the Geography department is an expert in spatio-temporal statistics and spatial modelling. With twitter being so time sensitive being able to run appropriate temporal and spatio-temporal statistics is necessary to creating accurate results.

Questions on statistics

- What sort of similar linguistic statistic projects have you seen?
- How did they handle the linguistics?
- How do people manage the spatio-temporal statistics in such projects?
- What sort of statistical tests or models would be appropriate?
- Are there books or other readings that you would recommend
Exploratory Data Analysis: We are operating with minimal knowledge of what the relationship is between a non geotagged tweet and its location. This means we have no real hypothesis that we are testing to start with. Our knowledge of the spatial exploratory data tools in GEODA will allow us to find trends in the data without bias, and use these trends to feed our statistical models with better information.

We could ask a Geography Prof or Computer Science Prof about the best ways to find relationships in data through Exploratory Data Analysis.

Questions
What methods do you use for exploring spatio-temporal data?
Are self organizing maps (SOM) applicable in our case?
If we want to visualize data straight from a database is there a better process then exporting shapefiles and using a GIS package?
What methods do you use to find

linguistics
Our lack of knowledge about linguistics, and computing with linguistics, will be our biggest hurdle. Linguistics is a very complicated area of study, mainly because language is so complex. Having good sources, both tacit and codified, to help us implement an appropriate linguistic system is necessary. We will need to do research and find people from both the Computer Science realm, and the pure linguistics realm to provide us with information about how to move forward with processing linguistics.

Sample Questions
What are parts of our written language that vary from region to region?
Is there data on how this varies in a unified source?
How do people parse sentences in linguistics?
Are summary statistics like length of sentence, number of verbs, or misspells commonly used?
Can research on variation in spoken language be applied to written language?
How do people work with words that negate other words?
How do people work with word order?
Does punctuation change from area to area?

Presentation
To present the information that we find we will need to create useful visualizations to make all of the information that we create visually understandable. This can be achieved with static or dynamic mapping, or other visualization tools

codified
Previous attempts visualize tweets geographic location will be an important reference point to see what other people are currently doing, and how they are tackling the problem of showing social media information on a map.

Project Cascade: This is a project showing how to store and visualize twitter data in relation to the New York Times. They store their data in MongoDB, and visualize it using processing. It is very clear that they have a very good structure for looking at how tweets are related to each other, but do not bring in any information about the location of the user or of the tweet. This is a very clean and easy way to think about and store the data, and raises questions about how we should proceed in storage and visualization of our data. One thing that we like about this project is the simplicity of the backend, and how it uses all free tools for storage and visualization.
Tacit

The knowledge of cartography and GIS in our team will allow us to make high end, visually appealing maps of the information that we gather in this project. Skills, like creating appealing maps using ArcGIS, or Adobe products will be important in communicating our data. These are skills that will need to be adopted to the data, but there will be many questions that should be asked of expert Cartographers. Mark Lindberg runs the Cartography Lab, and would be a good tacit source for presentation of data and cartography.

Cartography Questions
What are the best ways to present maps online?
Would a google map, mapserver or ArcGIS online map have advantages or disadvantages?
How do people present linguistics on maps? What are the best ways to do this?
How do people present text on maps?

Lab
(spool file is also attached)

BEGIN;
INSERT INTO USER_SDO_GEOM_METADATA (TABLE_NAME, COLUMN_NAME, DIMINFO, SRID)
VALUES ('ACTIVITIES', 'GEOM',
MDSYS.SDO_DIM_ARRAY
(MDSYS.SDO_DIM_ELEMENT('X', -122.327217000, 73.283330000, 0.000000050),
MDSYS.SDO_DIM_ELEMENT('Y', -72.311490000, 45.533021000, 0.000000050))
),
2001);
INSERT INTO USER_SDO_GEOM_METADATA (TABLE_NAME, COLUMN_NAME, DIMINFO, SRID)
VALUES ('ROADS_MAJOR', 'GEOM',
MDSYS.SDO_DIM_ARRAY
(MDSYS.SDO_DIM_ELEMENT('X', -74.464878000, -68.327047700, 0.000000050),
MDSYS.SDO_DIM_ELEMENT('Y', 17.798452400, 19.950185200, 0.000000050))
),
2002);
INSERT INTO USER_SDO_GEOM_METADATA (TABLE_NAME, COLUMN_NAME, DIMINFO, SRID)
VALUES ('NATURAL', 'GEOM',
MDSYS.SDO_DIM_ARRAY
(MDSYS.SDO_DIM_ELEMENT('X', -74.443720500, -68.349748900, 0.000000050),
MDSYS.SDO_DIM_ELEMENT('Y', 17.684553100, 19.951935100, 0.000000050))
),
2003);
COMMIT;

/* b.1 */
SELECT SDO_GEOM.SDO_AREA(n.geom, m.diminfo) as area
FROM F12C5980G20.NATURAL n, user_sdo_geom_metadata m
WHERE m.table_name = 'NATURAL'
AND m.column_name = 'GEOM'
AND n.type = 'forest'
order by area;

/* b.2 */
SELECT n.id, SDO_GEOM.SDO_AREA(n.geom, m.diminfo) as area
FROM F12C5980G20.NATURAL n, user_sdo_geom_metadata m
WHERE m.table_name = 'NATURAL'
AND m.column_name = 'GEOM'
AND n.type = 'forest'
AND rownum <= 1
order by area desc;

/* part b.3 */
SELECT SDO_GEOM.SDO_MBR(r.geom) as mbr
FROM F12C5980G20.ROADS_MAJOR r
WHERE r.name = 'Rue Saint Vincent';

/* part b.4 */
SELECT SDO_GEOM.SDO_CENTROID(n.geom, m.diminfo) as centroid
FROM F12C5980G20.NATURAL n, user_sdo_geom_metadata m
WHERE m.table_name = 'NATURAL'
AND m.column_name = 'GEOM'
AND n.name = 'Parque Central';

/* part b.5 */
SELECT r.type, sum(SDO_GEOM.SDO_LENGTH(r.geom, m.diminfo))
FROM F12C5980G20.ROADS_MAJOR r, user_sdo_geom_metadata m
WHERE m.table_name = 'ROADS_MAJOR'
AND m.column_name = 'GEOM'
GROUP BY r.type;

/* part c.1 */
SELECT water.osm_id as water_osm_id, park.osm_id as park_osm_id
FROM F12C5980G20.NATURAL water, F12C5980G20.NATURAL park,
user_sdo_geom_metadata m
WHERE m.table_name = 'NATURAL'
AND m.column_name = 'GEOM'
AND water.type = 'water'
AND water.name IS NOT NULL
AND park.type = 'park'
AND park.name IS NOT NULL
AND SDO_GEOM.WITHIN_DISTANCE(water.geom, m.diminfo, 1, park.geom,
m.diminfo) = 'TRUE';

/* part c.2 */
SELECT SDO_GEOM.SDO_UNION(park1.geom, m.diminfo, park2.geom, m.diminfo)
FROM F12C5980G20.NATURAL park1, F12C5980G20.NATURAL park2,
user_sdo_geom_metadata m
WHERE m.table_name = 'NATURAL'
AND m.column_name = 'GEOM'
AND park1.name = 'Parque Central'
AND park2.name = 'Parque Central'
AND park1.id <> park2.id;

/* part c.3 */
SELECT park.osm_id as park_osm_id, activities.id as activities_id
FROM F12C5980G20.NATURAL park, F12C5980G20.ACTIVITIES activities,
user_sdo_geom_metadata m
WHERE m.column_name = 'GEOM'
AND park.name IS NOT NULL
AND activities.verified = 'YES'
AND SDO_GEOM.WITHIN_DISTANCE(park.geom, m.diminfo, 1,activities.geom,
m.diminfo) = 'TRUE'
AND park.osm_id in (SELECT park.osm_id
FROM F12C5980G20.NATURAL park, F12C5980G20.ACTIVITIES activities,
user_sdo_geom_metadata m
WHERE m.column_name = 'GEOM'
AND activities.verified = 'YES'
AND SDO_GEOM.WITHIN_DISTANCE(park.geom, m.diminfo, 1,activities.geom,
m.diminfo) = 'TRUE'
GROUP BY park.osm_id
HAVING count(*) > 10)
ORDER BY park.osm_id, activities.id;

/* part c.4 */
SELECT r.name as longest_road_name
FROM F12C5980G20.ROADS_MAJOR r, user_sdo_geom_metadata m
WHERE m.table_name = 'ROADS_MAJOR'
AND m.column_name = 'GEOM'
AND SDO_GEOM.SDO_LENGTH(r.geom, m.diminfo) =
(SELECT MAX(SDO_GEOM.SDO_LENGTH(r.geom, m.diminfo))
FROM F12C5980G20.ROADS_MAJOR r,
user_sdo_geom_metadata m
);
** topological
6. space-filling curves (fig 6.8, p 234)
** fractal geometry
7. planar configurations (such as polygons)
** graph theory
8. straight-line distance
** euclidean space
9. convex polygon
** euclidean space
** set theory
10. landslide probability (ch 9.4 p 350)
** set theory
*** Probability is based on set theory. For example, the sample space and the event space are both sets.
11. interior
** topological
12. symmetric relationship
** set theory
13. complement
** set theory
14. exterior
** topological

3.4
1. Cylindrical
   Topological (preserves topological properties but distorts others)
2. Azimuthal
   Projective (this is a projection onto a circle)
3. Upside-down
   Similarity, Affine, Projective, Topological

4.2
1. Determine warmest temperature (or year) for each cell.
   ** local
2. Determine warmest cell in each country in year 2000.
   ** zonal
3. Identify country with highest average cell temperature in year 2000.
   ** zonal
4. For each cell, compute spatial-neighborhood average temperature in the year 2000.
   ** focal
5. For each cell, compute heat-island-factor as the difference between its temperature and its spatial-neighborhood average temperature for the year 2000. Assume the results of previous step were available as an input for this step.
   ** local (don't need to look at neighbors if you have the results of the previous step available as input)
6. For each country, list the cell with highest value of heat-island-factor. Assume the results of previous step were available as an input for this step.
   ** zonal
7. Compute average annual temperature of surface of Earth for each year.
** local (not zonal because it is computed point by point, not country by country)

4.4
1. Austria, Europe
   ** Austria inside Europe
2. USA, Canada
   ** USA meets Canada
3. Europe, Asia
   ** Europe meets Asia
4. Lake Superior, USA
   ** Lake Superior overlaps USA
5. Turkey, Europe
   ** Turkey overlaps Europe
   ** Note: it is not the case that Europe covers Turkey, because there are parts of turkey which are in Asia
6. Lesotho, South Africa
   ** Lesotho inside South Africa