703804: Programming Lab: Innovative Interaktion, Visualisierung und Analyse Project Proposal

David Kofler David Westreicher Matej Stanic

May 13, 2015

1 The Project

The following project proposal builds up upon the nine *Heilmeier Questions*. Each subsection gives an answer to one of the nine questions.

1.1 What are you trying to do? Articulate your objectives using absolutely no jargon. What is the problem? Why is it hard?

Basically, the goal of the project is to visualize data concerning football transfers including twitter data, which is interesting for football fans. The single components of the project are the following:

- **Visualization of the football transfer graph.** Using data from `www.soccerbase.com` the goal is to construct a graph of player transfers between clubs. Intuitively, the graph is a directed graph, where the nodes are the teams and the edges have the players and the transfer fees as attributes. Our visualization idea builds up upon the graph from `www.transferwindow.info`. Initially, the graph is visualized in a geographical world map. The user then can click on a league/team/player to show only transfers of the league/team/player. Also, clicking on one of those three would provide additional statistics. Furthermore, an extended filter can be applied, e.g to show only transfers of players which are under 21, or players which are from Spain.
- **Analysis and visualization of team/player statistics from the football transfer graph.** The transfer graph can be used to analyze certain behaviors of players and teams using graph analysis algorithms. For example, depth search can be used to find players which left a team and then came back after some years. Or, using the *PageRank* algorithm[1],

[1] `http://en.wikipedia.org/wiki/PageRank`

teams can be found which are of high importance when it comes to transfers. Brath and Jonker (2015) give an overview of graph algorithms, that are also interesting for this project.

- **Calculation of popularity of teams using twitter.** Each football team has an official twitter account where they post news, events, etc. When clicking on a team in the football transfer graph, statistics of their twitter account would be presented, including the number of followers. Unfortunately, the number of followers doesn't reflect the true number of fans. Ramasamy et al. (2013) propose a method to identify user interests from tweet times, which are compared to the tweet times of events concerning the team. This algorithm can therefore identify true fans.

- **Inclusion and analysis of transfer rumors from different sources.** Using twitter and/or news sites we will include transfer rumors. Kunneman and Van den Bosch (2012) have used different machine learning methods to find tweets that refer to scheduled or unscheduled events like transfer rumors. We can directly use their approach in our project, but for unscheduled events the performance is poor.

- **Classification and visualization of fan opinions from twitter.** The project will also include user opinions about transfers. Shrivatava et al. (2014) have proposed a model for classifying user tweets as positive, negative or neutral using Support Vector Machines. This model can also be applied to transfer opinions.

Obtaining the data for the transfer graph requires scraping, which shouldn't be a big problem. When the needed data is obtained, visualizing and analyzing the transfer graph is straightforward. On the other hand, twitter data is bound to a language. So, we will at first only take a look at English clubs. Also, machine learning methods for classification and prediction cannot be perfect, so that mistakes are expected.

1.2 How is it done today, and what are the limits of current practice?

There exists a transfer visualization, but its visualization can be made more intuitive. Also, the analyses have not been performed yet on soccer data (to our knowledge).

1.3 What's new in your approach and why do you think it will be successful?

There are many pages that offer transfer data and rumors, but there is no service that combines visualizing such data and taking twitter data into account.

1.4 Who cares?

Football fans will have a whole new level of investigating football transfers, including transfer rumors and other interesting statistics coming from twitter.

1.5 If you're successful, what difference will it make? What impact will success have? How will it be measured?

The payoff would be interesting insights into transfer behavior of football players and teams and Twitter usage of fans.

1.6 What are the risks and the payoffs?

There are no risks. The payoff will be a visualization where interested fans and researchers can visualize teams and player transfer patterns, as well as how fans behave.

1.7 How much will it cost?

Nothing in the first year since it can be hosted by using AWS (Amazon Web Services) Free Tier.

1.8 How long will it take?

The project is quite involved. This section contains estimates for the tasks. They will change as we gain experience with the subject and the tools we use.

1.8.1 Visualization of the football transfer graph

Subtasks:

1. Create database model: 2 hours
2. Program crawler: 8 hours
3. Scrape `www.soccerbase.com`: 2 hours
4. Set up hosting environment for visualization: 2 hours
5. Create backend to access the database: 2 hours
6. Get to know the D3[2] framework: 5 hours
7. Create the visualization: 8 hours
8. Test the visualization: 3 hours

[2] `http://d3js.org/`

1.8.2 Analysis and visualization of team/player statistics from the football transfer graph

- **Depth search.** The Depth Search graph algorithm is a well-known graph algorithm and should not pose significant implementation challenges. There is a caveat though: the transfer data is stored in an SQL database, which would make using recursive algorithms inefficient. There a two approaches to solve this problem:
  - *Analyze the data with the help of a graph database:* Graph databases (for example Neo4J[3]) are optimized for executing queries which traverse graphs. After the results are calculated they can be cached in another database.
  - *Limit analysis to few players at a time.* It can be assumed that no player has a transfer history large enough to prohibit loading it into main memory and analyzing it there.

The first approach would consume about five hours to get to know a graph database and about five hours to develop the analysis. The second approach would consume three hours.

Visualizing the results should take only about three hours since at that point we should already know the visualization framework well enough.

1.8.3 Calculation of popularity of teams using twitter

Subtasks:

1. Setup Twitter access: 1 hour
2. Setup analysis database: 2 hours
3. Implement offline algorithm: 5 hours
4. Test offline algorithm: 8 hours
5. Implement online algorithm: 5 hours
6. Test online algorithm: 5 hours
7. Visualize results: 1 hour

This analysis is quite involved: although the method proposed in Ramasamy et al. (2013) does not employ text analysis, the computational effort must not be underestimated. To estimate the number of fans, the tweets of thousands of Twitter accounts have to be analyzed. The initial analysis must be done offline. After that it should be possible to keep the results up-to-date solely by analyzing new tweets.

[3] `http://neo4j.com/`

1.8.4 Inclusion and analysis of transfer rumors from different sources

This analysis is quite involved since it uses supervised learning. It is estimated to take around eight hours of implementing and three hours to train it, although training can be sped up by work sharing.

1.8.5 Classification and visualization of fan opinions from twitter

Subtasks:

- Setup Twitter access: 1 hour
- Implement algorithm: 8 hours
- Test algorithm: 5 hours
- Visualize results: 1 hour

1.9 What are the midterm and final "exams" to check for success? How will progress be measured?

The first milestone will be the visualization of the transfers alone. This milestone should be accomplished after Easter. It should be possible to also implement the depth search analysis by that date. By the midterm presentation (May 3) two further analysis should be implemented. Finally, all of the above analyses should be implemented or have been rejected.

2 Survey

This section gives an more detailed survey of the single sources of the project. Every group member has chosen at least three sources.

2.1 David Kofler

- Cattuto et al. (2013) use Neo4J to store data about social interactions which was collected during a conference by using wearable proximity sensors. They present some example queries and discover that querying over densely-connected nodes makes it difficult to achieve high performance. This is something we have to keep in mind when we create our data model.
- Jarukasemratana et al. (2013) contains an interesting algorithm to find communities of clubs between which players often circulate. Many important real-world networks have the scale-free property (we assume that the football transfer network too), but detecting communities within them is not always easy because subcommunities tend to be classified as communities on their own. This paper presents an algorithm specialized for scale-free networks.

- `D3.js` (`http://d3js.org`) is a JavaScript framework to create animations on webpages. It is similar in usage to other popular JavaScript libraries (like jQuery and Prototype), but contains special features to create animations. We will use it to create our visualizations.

2.2 David Westreicher

- `http://eyeseedata.com/football-player-transfers/` is a blogpost describing the visualization of global football transfers. It contains a video which shows a map of the world in which every transfer is shown as a edge between two countries. The video progresses through the years 1900 to 2013. Then there is also a interactive map where you can click on a country and see the ingoing and outgoing transfers of the selected country. The last interactive part is a statistic of the number of transfers and their fees, grouped by countries. These visualizations give a good idea of how such a map would look, but the difference to our project is that we don't group transfers to countries but to clubs and we will combine the transfers with social data.

- `www.transferwindow.info` is a interactive visualization of football transfers of popular leagues. The main idea here is to represent the transfers by a graph between leagues or teams. On the right side of the site there is also information of the top transfers in the currently selected view. The site is visually appealing and gives a good glimpse of how a beautiful style for graphs is achieved. Our project will however lay out the graph on top of a world map.

- Ramasamy et al. (2013) is a interesting paper about extracting interests from twitter users solely by the time they write their tweets. The idea is to connect the time of the tweets to the time of external events. We could use such an approach to find out if a user is a fan of a certain team. For example if team A currently has a match with team B and the user T tweets while they are playing. Then the user T is likely a fan of A or B.

2.3 Matej Stanic

- `www.soccerbase.com` is a football news site that not only contains news but also many statistics about leagues, teams and players. This includes also a detailed transfer history for every team and player. It's a site where football fans can keep up to date. For our project, detailed transfer information is inevitable in order to build a correct graph of transfers. Soccerbase.com offers exactly that type of transfer data that is needed to build such a graph. By scraping we can directly get the data that we need.

- Shrivatava et al. (2014) have evaluated a method for sentiment analysis for Twitter posts. They implemented a crawler to obtain twitter posts, which then where classified as positive, neutral or negative using a standard

Support Vector Machine (SVM). The SVM was evaluated on a predefined dataset using cross validation. The average accuracy of their method was 70.592%. For our project, the same approach can be used for transfers and/or transfer rumors. After extracting football fan opinions on transfers from Twitter they could be classified as positive, negative or neutral. This would somehow be a measure of acceptance between the fans and the owners and managers of the club. There may be one problem, which is assigning tweets to transfers. This could be solved by only taking into account tweets which are referencing the transfer tweet.

- Kunneman and Van den Bosch (2012) evaluate methods which identify tweets referring to scheduled or unscheduled events. Their case study consists of two scenarios: The first one are football matches which are scheduled. On the other hand they also used an unscheduled scenario, football transfers. They have tested several state-of-the-art machine learning methods for their scenarios. The results are twofold: For the scheduled event scenario, the accuracy for some methods was above 80% while for the unscheduled scenario the accuracy was near-baseline. This means that even though the machine learning methods can directly be used for our project, it would not be productive because of the poor performance.

References

Brath, R. and Jonker, D. (2015). *Graph Analysis and Visualization: Discovering Business Opportunity in Linked Data*. Wiley.

Cattuto, C., Quaggiotto, M., Panisson, A., and Averbuch, A. (2013). Time-varying Social Networks in a Graph Database: A Neo4J Use Case. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 11:1–11:6, New York, NY, USA. ACM.

Jarukasemratana, S., Murata, T., and Liu, X. (2013). Community detection algorithm based on centrality and node distance in scale-free networks. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pages 258–262, New York, NY, USA. ACM.

Kunneman, F. and Van den Bosch, A. (2012). Leveraging unscheduled event prediction through mining scheduled event tweets. In Roos, N., Winands, M., and Uiterwijk, J., editors, *Proceedings of the 24th Benelux Conference on Artficial Intelligence*, Maastricht, The Netherlands.

Ramasamy, D., Venkateswaran, S., and Madhow, U. (2013). Inferring user interests from tweet times. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, pages 235–240, New York, NY, USA. ACM.

Shrivatava, A., Mayor, S., and Pant, B. (2014). Opinion mining of real time twitter tweets. In *International Journal of Computer Applications*.