Community Detection of Screenplay Characters

Christos Makris, Pantelis Vikatos

To cite this version:

Christos Makris, Pantelis Vikatos. Community Detection of Screenplay Characters. 12th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI), Sep 2016, Thessaloniki, Greece. pp.463-470, 10.1007/978-3-319-44944-9_40 . hal-01557593

Community Detection of Screenplay Characters

Christos Makris[1] and Pantelis Vikatos[1]

1. Computer Engineering and Informatics Department, University of Patras, Greece,
{makri, vikatos}@ceid.upatras.gr

Abstract. In this paper, we present a model for automatic community detection of the characters by parsing movie screenplays. In our procedure it is proposed a classification model to predict the casting by categorize each line of the script in a character or not and the co-occurrence of the characters in the same scene constitutes the link between two characters in the social network. We use an existing modularity based community detection algorithm for cutting the created graph in communities. The innovation of our methodology is contained in the extraction of the casting of the screenplay from Wikipedia pages in order to train and build an efficient classifier to identify the characters in a screenplay. The proposed methodology for extracting automatically the social network and communities of screenplay characters can be probably used for enhancing movie recommendations.

Keywords: Community detection, Machine Learning, Social Network Analytics

1 Introduction

The automatic extraction of social network from sources such as unstructured text has gained the interest of researchers as is presented in several studies [1–6]. In some studies such as in [2, 5] the extraction of social networks is by parsing literature text which the characters of the book are the nodes and the dialogue between a pair of character constitutes a link. This procedure is expanded also to movie screenplays such as in [3, 4, 6] which can be seen as unstructured literary works which contain interactions between characters that could be presented as social network. In this paper we introduce an innovative scheme to extract communities of the characters involving in the same scene by only parsing the script of a film using an efficient classification model to predict the characters in a screenplay.

The main challenges is this research is the lack of pre-annotated dataset which can be used as a training and test set of a model. Also most of the screenplays are not well-structured so as to declare a universal rule of regular expression for automatic detection of significant information.

An important aspect of our work is exploiting by envisaging a similarity metric between movie screenplays that should be based on the structure of their social network derived from the co-occurrence of characters in scenes that could be useful for movie recommendations.

2 Related Work - Motivation

In [1] is presented a first approach in this field mapping out texts according to geography, social connections and other variables.

A study to extract social networks from nineteenth-century British novels and serials is presented in [2] which the networks have been constructed by dialogue interactions.

A similar study is presented in [5] related to social event detection and social network extraction from a literary text and particularly to the book Alice in Wonderland.

An expansion also to movies' screenplay as a source of as unstructured literary works is presented in [3, 4, 6, 5]. In [3] is proposed the extraction of social network by parsing screenplay in order to investigate communities, hidden semantic information and innovations to automation of story segmentation. Another study in [4] is focused in character interaction and networks between characters from plays and movies.

In [6] it is presented a formalization of the task of parsing movies' screenplays as well as a extraction of social network of all characters having a dialogue with each other in a scene with links.

3 Methodology-Model Overview

In the following methodology the automatic detection of communities using movie's screenplay is described. In Figure 1 the whole system architecture and the separated modules is depicted. In the following sections the detailed description of the each part of proposed model is included and in the Section 5 the experimental procedure and the results are presented.

3.1 Crawling IMSDB webisite

We collect a set of movie screenplays via crawling the Internet Movie Script Database (IMSDB) website [1] which contains a list of movies and the sceenplays for most of them. In studies [6, 8] it is mentioned that scripts are constituted by 5 elements:

1. the scene boundary which describes if the scene is to take place inside or outside using the tags INT. and EXT. respectively and the name of the location.
2. Scene description which is below the scene boundary and declares detailed information about the scene.
3. Character name which is the role that is involved in the scene.
4. Dialogue that the active role will act.
5. Meta-data for special information in the script.

[1] IMSDB website: http://www.imsdb.com/

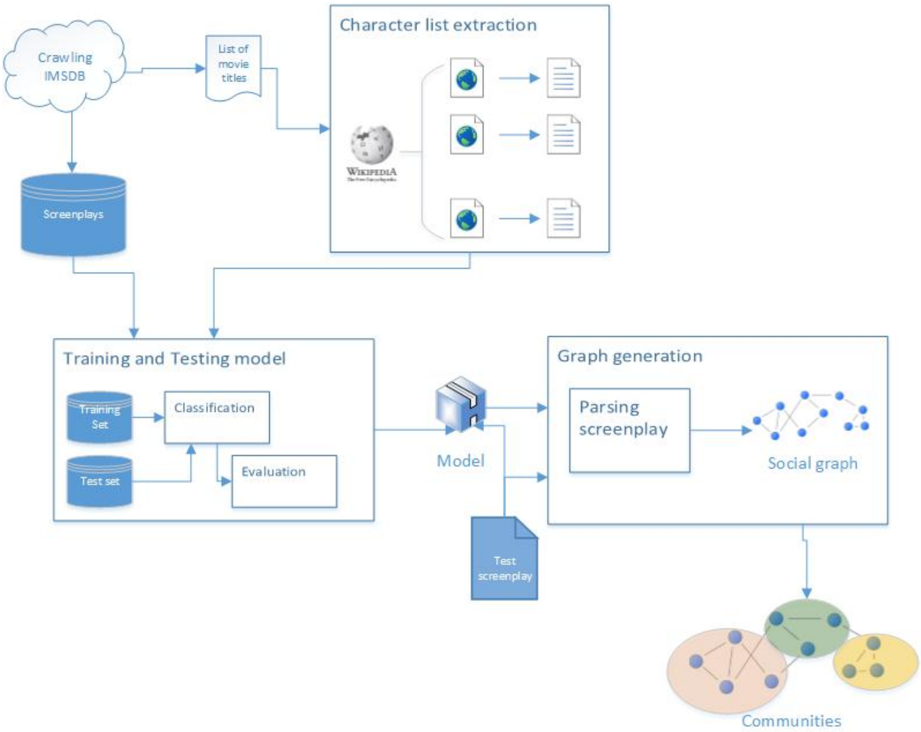


Fig. 1: System architecture

In our case it is necessary to recognize the scene boundary element in the script and split the scenes in order to identify the characters between them. Also an initial test about $EXT./INT.$ tags in the text should be occurred in order to check remove unstructured screenplays from our data collection.

3.2 Scrapping Wikipedia information

In study [6] it is mentioned that regular expressions is not the appropriate way to identify and annotate the character names from the scripts. We propose a procedure using Wikipedia website [2] as a source to extract the roles in the movie. Wikipedia site uses a well-structured information about the cast in a certain movie. We created a log file with the specific urls for all movie titles and we scraped the html code. The information is between the following html tags:

- $<h2><span\ class = "mw - headline" id = "Cast" >$
- $<h2><span\ class = "mw - headline" id = "Production" >$

The characters form each Wikipedia page are extracted in order to be matched with the ones in the screenplay.

[2] Wikipedia website:https://www.wikipedia.org/

3.3 Parsing and Annotation of Screenplays

We annotate each line of the script with the tag "C" for the line which only contains the character name and the rest lines with the tag "O". It is noted that in the well-structured screenplays the name of each character is between scene boundaries, after the scene description and before dialogue in direct or indirect speech. Also character names are capitalized, with an optional (V.O.) or (O.S.) information for "Voice Over" or "Off-screen." respectively. Examining the correctness of the annotation we check if all character names are within two scene boundaries. Parsing all the scripts from our collection all lines where gathered in a super-text with the proposed annotation C/O for each line. The tag is the category that will be predicted by the classification model as it is shown in Figure 1.

3.4 Using Linguistics to create Feature Vector

In this section we aggregate all the scripts to one super-text and we argue each line of the text will be transformed to a vector with linguistic and emotional features. We used LIWC [7] as a tool to extract linguistic characteristics. As a result, we created a vector of 80 characteristics. The linguistic and emotional analysis of each line will provide the appropriate type of features in order to train efficiently the classification model.

3.5 Training the Classification Model

As it is mentioned our scope is to identify the characters in the screenplay. A classification model has been used and trained for this purpose. The first step is the separation of our dataset in train and test set as it using K-Fold Cross-Validation. The main advantage is that all instances in the dataset are eventually used for both training and testing. In Section 5 the procedure for the separation of the dataset in training and test set as well as the performance of the classifier is described in detail.

4 Graph Generation and Community detection

The social graph represents the co-occurrence of characters in the same scene. More specifically the nodes of the graph constitute the characters and the link the co-occurrence of the pair of nodes in the same scene. Each link contains an attribute weight related to the frequency of the co-occurrence in the whole screenplay. We parse the screenplay in order to recognize the scenes. According to the structure of the screenplay the scene are between the $INT.$ and $EXT.$ tags. The communities in the social graph are detected by a well-used community detection approach [9] using modularity optimization as algorithmic progresses. The density of edges inside communities and outside communities is related to a modularity in a scale value between -1 and 1. Heuristic algorithms are used in order to eliminate checking all possible iterations of the nodes into groups while optimizing the modularity value. Using the Louvain Method [9, 10] of community detection, first small communities are detected by optimizing modularity locally and then each small community is clustered into one node and the procedure is repeated.

5 Implementation and Results

We implemented the crawler of the IMSDB website in python 2.7. The total number of files that we gathered was 1112 screenplays. In Section 3.1 it is noted that a well-structured screenplay contains 5 elements and thus we checked each file to overview it. Firstly it was checked the existence of the element $EXT./INT.$ for the scene boundary and 972 screenplays had this element in their text. Also we checked the existence of characters' name inside the scene boundaries and the number of files that passed this test was 501 which means that 45% of the initial screenplays were well-structured and appropriated to be introduced in our methodology as it is shown in Table 1 .

Table 1: Instances after preprocessing

Preprocessing Step	**number of screenplays**	**% of dataset)**
$EXT./INT.$ occurrence	972	87%
Existence of characters between $EXT./INT.$	501	45%

Based on the films' name a list was created with the urls that link films with Wikipedia pages. The scrapping of html code for discovering characters' name is described in detail in Section 3.2 using regular expression to identify the appropriate tags. We annotated each line of the screenplays with labels C/O for character and other respectively and we gather all lines for the screenplays in a a super-text with total size 220 MB. A vector with 80 linguistic and emotional characteristics for each line was created with LIWC [7] software forming a dataset with the 86% and 14% of instances in label C and O respectively as it is presented in Table 2.

Table 2: Distribution of Labels

Label	**Number of instances**	**% in dataset)**
Other Line	2859281	86%
Character	465464	14%

We developed a classifier to predict each label using Weka [3] environment. We used the J48 decision tree classifier in order to predict the label in the test

[3] Weka toolkit: http://www.cs.waikato.ac.nz/ml/weka/

instances. The classifier from Weka is used with the default settings. We separated the dataset to training and test set, using K-Fold Cross-Validation (K=10 Fold). We evaluated the classifier based on F-measure which is the harmonic mean of precision and accuracy of the classification and our classification model achieved 99.3%.

We worked in two case studies for "X-Men" and "The Lord of the Rings: The Fellowship of the Ring" (LOR) to present the produced social networks and communities in our methodology. In Section 4 it is described the procedure for social graph generation and the modularity optimization algorithm for community detection. According to the predicted character list which was derived from our classifier the pattern matching in the screenplays led to the information of Table 3.

Table 3: Parsing Screenplay and Attributes Extraction

Screenplay Attributes	**X-men**	**LOR**
# Characters	28	21
# Scenes	185	111
# Characters' co-occurrence	192	134
Mean Characters per Scene	2	3

In Figure 2 is depicted the social networks of co-occurrence characters in screenplay scenes. Line thickness express the weight in the link between two characters. In the social networks we calculate node centralities as it is shown in Table 4.

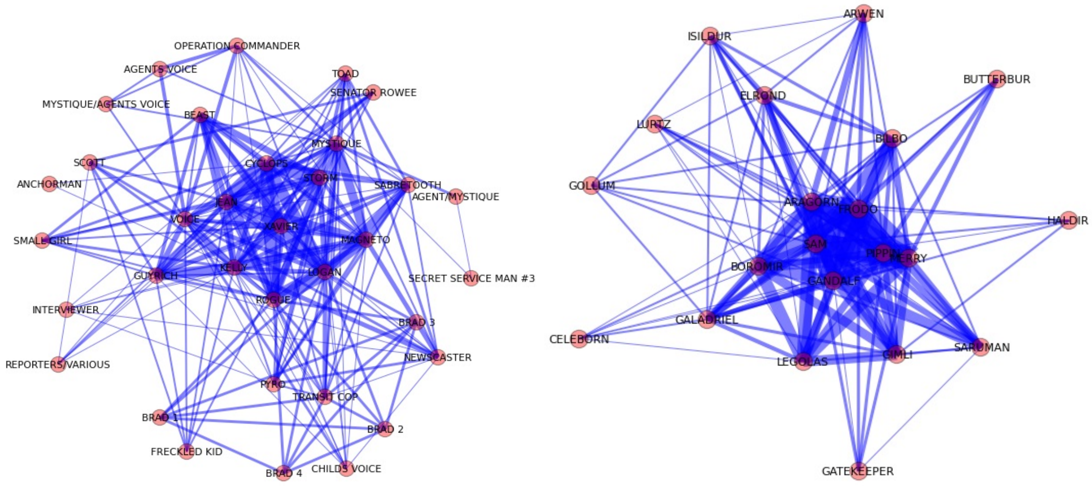
Table 4: Social Network Centralities

Social Network Centralities	**X-men**	**LOR**
Degree Centrality	0.342	0.638
Betweenness Centrality	0.024	0.019
Closeness Centrality	0.582	0.755

6 Conclusions and Future Work

In this paper there is the detailed description of the methodology for extracting social network and communities of the characters involving in the same scene by only parsing the movie screenplay. The proposed methodology is based on the performance of an efficient classifier for character recognition.

We deal with the main challenge of the lack of pre-annotated dataset which can be used as a training and test set of the classifier adopting the non use of



(a) X-Men

(b) The Lord of the Rings

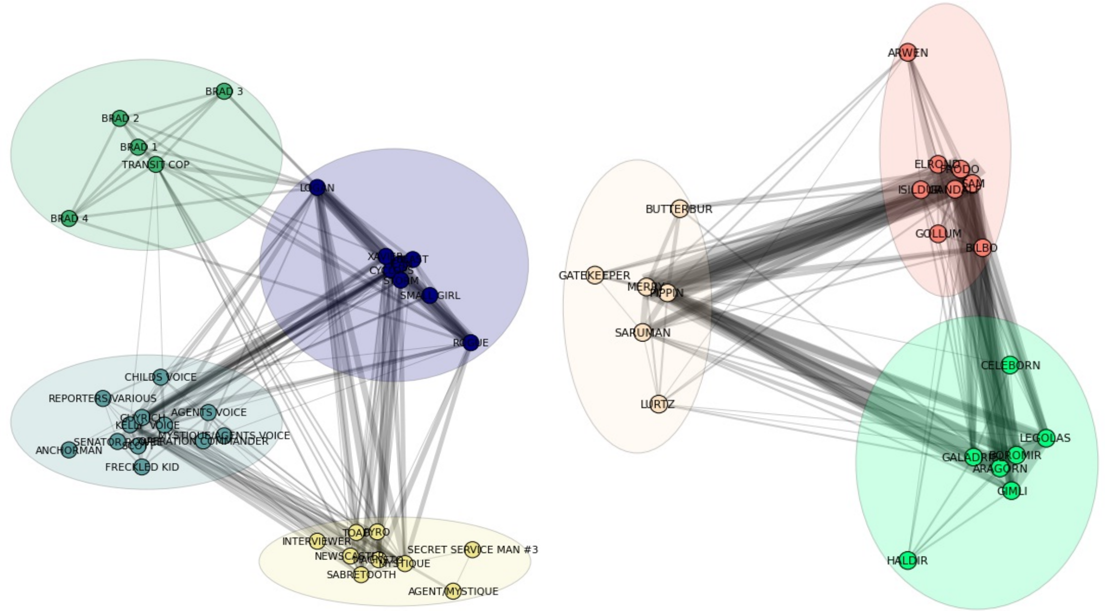
Fig. 2: Social Networks of Co-occurrence Characters in Screenplay Scenes

universal regular expressions for character recognition which is not the appropriate way based on previous studies. On the other hand we introduce as an external source Wikipedia webpages in order to extract by html scraping the appropriate information for screenplays.

As future work, we are interested in examining the temporal information of the scenes and discovering the fluctuation of the sentiment between the characters' dialogues. Also this model could be used as a subsidiary factor for the recommendation of similar movies based on the structure and attributes analysis of their social network.

References

1. Franco Moretti. 2005. Graphs, Maps, Trees: Abstract Models for a Literary History. Verso, London.
2. David K. Elson, Nicholas Dames, and Kathleen R.McKeown. 2010. Extracting social networks from literary fiction. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 138147.
3. Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. 2009. Rolenet: Movie analysis from the perspective of social networks. Multimedia, IEEE Transactions on, 11(2):256271.
4. Sebastian Gil, Laney Kuenzel, and Suen Caroline. 2011. Extraction and analysis of character interaction networks from plays and movies. Technical report, Stanford University.
5. Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. (2013a). Automatic extraction of social networks from literary text: A case study on alice in wonderland. In



(a) X-Men

(b) The Lord of the Rings

Fig. 3: Characters Communities in Screenpaly Scenes

the Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013).

6. Apoorv Agarwal, Sriramkumar Balasubramanian, Jiehan Zheng, and Sarthak Dash. 2014b. Parsing screenplays for extracting social networks from movies. EACLCLFL 2014, pages 5058
7. J. W. Pennebaker, M. E. Francis and R. J. Booth, *Linguistic Inquiry and Word Count (LIWC): LIWC2001*, New Jersey: Lawrence Erlbaum Associates, 2001.
8. Robert Turetsky and Nevenka Dimitrova. 2004. Screenplay alignment for closed-system speaker identification and analysis of feature films. In Multimedia and Expo, 2004. ICME04. 2004 IEEE International Conference on, volume 3, pages 1659 1662. IEEE.
9. V. D. Blondel, J. - L. Guillaume, R. Lambiotte and E. Lefebvre, *Fast Unfolding of Community Hierarchies in Large Networks*, Journal of Statistical Mechanics: Theory and Experiment, P10008, 2008.
10. S. Fortunato, *Community Detection in Graphs*, Physics Reports 486, pp. 75-174, 2010.