Modern Social Media Based Film Data Analyse: A Comprehensive Case Study Related With NLP and DL

ACM Reference format:
. 2016. Modern Social Media Based Film Data Analyse: A Comprehensive Case Study Related With NLP and DL. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 6 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Big film data stands for big capacity and big value where capacity is the sum of total information bought by one movie online and value is the instructive finding of latent target interests or tendency behind data representation. From investment to profit, movie data analyse play an important role in maximizing interest both investors and audiences. x.x.x.

In recent years, with the heat of the movie market to increase, a large number of capital influx movie industry. However, the film industry in China is still in its development stage, and the immature market has made movie investment characterized by high risk and high returns. In order to pursue high incomes, the usual way is to use a large number of "celebrities" to promote the box office through the fan effect. At that time, only a few of the works were successful. But it has caused the soaring star worth, the consequent increase in production costs. The reason, Star Bowl although there is a huge fan base, but the star match with the degree of work, star fan characteristics match with the extent of the work is the most important factor affecting the box office. Actors, scripts, publicity, release schedule and other aspects of the combined effect of the results to measure the level of contribution to the box office record can not be obtained directly from the available data, the existing method is the lack of an effective quantitative method to assess the creative record of the box office Contribution value, thus helping investors to assess the investment star's rationality.

1.1 Challenges and Solutions

Considering improvements and new methods developed moving with time, natural language process (NLP) and deep learning (DL) is sharp solution in both data mining especially text mining and machine learning area. Based on modern technique, we proposed most valuable end-to-end analyse system Film Knowledge Analyse Platform (FKAP) which cover movie data of films. The main contribution of FKAP is finding state-of-art approach to traditional movie big data mining task. We summarized the key problem that evolved system FKAP focused on.

DIFFICULTY 1.1.

DIFFICULTY 1.2.

DIFFICULTY 1.3.

1.2 Roadmap

2 PLATFORM OVERVIEW

3 HETEROGENEOUS DATA INTEGRATION AND ANALYSIS

3.1 Data standardization

An obvious phenomenon is that the box office of the movie has significant difference with the year and type. In order to eliminate these effects, we define the box office score to measure the box office quality of a movie. Given the box office of a movie bo and its movie type set $T = T_1, T_2, ..., T_k$ and release Year y, the box score is $boxscore = \frac{1}{|T|}\sum_{t\in T}\frac{log_2 bo}{maxbox_{y,t}}$,where $maxbox$ is defined as $max(log_2(boset)_{year=y,type=t})$, which $boset$ is a set of boxoffice. Similarly, due to the difference of movie types, the original movie ratings can not accurately describe the word of mouth of the movie. We evaluate the word of mouth score according to the min-max method. Given the word of mouth a movie wm and its movie type set $T = T_1, T_2, ..., T_k$, the word of mouth score is $wms = \frac{1}{|T|}\sum_{t\in T}\frac{wm-min(wms_t)}{max(wms_t)-min(wms_t)}$, where wmt the set of the word of mouth of movies where type is t.

3.2 The Analysis of Actors and Directors

$bos = \frac{1}{N}\sum_{i\in N} bo_i$
$bov = \frac{1}{N-1}\sum_{i\in N}(bo_i - bos)^2$
$wms = \frac{1}{N}\sum_{i\in N} wm_i$
$wmv = \frac{1}{N-1}\sum_{i\in N}(wm_i - wms)^2$

3.3 Name ambiguity elimination

Name ambiguity is a special case of identity uncertainty where one person can be referenced by multiple name variations in different situations or even share the same name with other people.

4 SENTIMENT ANALYSE

The main purpose of sentiment analyse is to discover views and opinions diversity of different targets in films which reflect what attract audience watching the movie. The sample dataset in this area are review corpus $R = \{r_1, r_2, \ldots, r_n\}$ and each r_i has film name f_i it belonging to, time point $time_i$, scores towards this film s_i and user u_i who scores. These attributes of reviews are discrete (has domain specific value) and many sentences in it contribute to the

Conference'17, Washington, DC, USA

DOI: 10.1145/nnnnnnn.nnnnnnn

whole sentiment of this single review. Since reviews are spread in social media platform such as twitter or micro-blogs, the sentences are limited and review of movie are usually enough short. Thus, many phenomenon occurred on short texts or natural language add difficulties when analysing the sentiments of targets (e.g. Rhetoric, metaphor, proverb and nicknames). Unlike many works focus on proceed in sentence-level document-level and time-period level step by step, we try to figure out the sentiment to a specific actors or director(leading creator of this film). A useful prior knowledge in reviews together with reviewers' rating score or stars made by user is evaluation score of specific film. We make the best of this knowledge to enlarge our sentiment words database because user usually behave consistent sentiment to same target.

4.1 Named Entities Recognition

If we want to know sentiments behind human expression, targets that people are taking about should be identified firstly. Here we concentrate on comment target about leader creators in films. The main challenging is the nicknames for actors and directors.

Sequence to sequence approach. Given a word sequence $X = \{x_1, x_2, \ldots, x_n\}$ we observed and the 4-tag label set $Tags = \{O, B, E, M\}$, the objective task is to find correct corresponding labels $Y = \{y_1, y_2, \ldots, y_n | y_i \in Tags\}$. By optimizing the loss function with parameter θ

$$J(\theta) = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} loss(f(x_i; \theta), y_i) \tag{1}$$

recurrent neutral networks such as *LSTM*, *BidirectionalLSTM*, *BidirectionalLSTM+CRF* have been proved to be state-of-art structures. They consider both long time and short terms information and find tags with maximum possibility for sequence labeling. However, previous machine learning approaches need huge domain training data, while it is hardly to label these social media corpus accurately which are full of informality and flexibility. Especially actors get constantly evolving new roles in various films as time goes by, we hope find a simple and efficient way that recognize most of the comment targets in films reviews.

Fast and adaptive approach. In order to mining entities in linear time, we apply key word matching for leader creators in films. Online knowledge database which store film meta information can be strong support providing comprehensive and dependable analysis. Focused crawling technology based on web linkage contribute to construction of prior knowledge of **actors**, **roles** and **directors**.

Assume average noun morphemes length is m for all n reviews and t targets for all F films. Algorithm 1 need compare $c \times nmt$ times. Since word comparing consume limited space and time and $m \ll n, t \ll n$, we extract potential comment targets in $O(n)$ time complexity which greatly less than deep learning approach. We can easily adjust designed parameter and improve efficiency with parallel framework when processing TB data.

In order to decrease the storage space of data, we replace film name and user name by mapping them using film id and user id. $f_i \in F = \{f_1, f_2, \ldots, f_m\}$ and $u_i \in U = \{u_1, u_2, \ldots, u_U\}$. Let score for each review be $s_i \in [0, 10]$ and $time_i$ be the timestamps. The features we finally retrieved can be summarized in table 1.

Algorithm 1 Framework of nickname mining for our system

Input:
The name set of **actors** and **directors** in each film
The name set of **roles** for each actors used to played
Threshold $\boldsymbol{\alpha}$ for minim similarity of misspell and variation
Threshold $\boldsymbol{\beta}$ for max tolerance of misspell and variation

Output:
Potential comment targets mapping dictionary A, P

1: Aggregating reviews group by corresponding film name along with preparing actors list A_i, directors list P_i in ith film according given **actors** and **directors**;
2: Constructing actors mapping A for every actor in a_i, so dose P for every director in p_i;
3: Quality similarity between every noun morphemes w and t in **roles**, **actors** and **directors** by *Jaccard* char distance in all comments for one film;
4: Recognizing potential nickname by the similarity and threshold $\boldsymbol{\alpha}$ we set, if $Jaccard(w, t) > \alpha$ add it to mapping dictionary A or P;
5: Checking potential nickname in mapping dictionary, if $Levenshtein(w, t) > \beta$, delete it from mapping dictionary A or P;
6: **return** A, P

4.2 Target-Dependent Sentiment Classification

Sentences tend to be more complex when targets we want to analyse increase and sentiments detection being a hard work when analysing more objects. We take leader creator of film as major targets in this paper. Examples containing multi targets are shown as follows.

(1) Live here now! The point of life is looking for the point. —*A dog's purpose*
(2) It has been a year since I were familiar with Alexander Sandro Gonzalez Inarritu whose films have a cruel irony and full bitterness.—*Bird man*
(3) Mia had dreamed of becoming an actress known by more audience, while Sebastian want to won a place in his loving jazz. —*LaLa land*

Sentence(1) is a non-target review, while Sentence(2) shows one-target sample and Sentence(3) stands for multi-target sentences. Considering targets after NER process in section 4.1 consist of actors map list $A = \{a_1, a_2, \ldots, a_A\}$ and directors map list $P = \{p_1, p_2, \ldots, p_P\}$, we cluster sentences into comments set with two polarity(positive +1 and negative −1). Binary sentiment classifier play an important role in sentiment classification. Sentiment polarity $SP = \{+1, -1\}$ labeled by classifier for each sentences in corpus st_i^j of ith film with corresponding target j, $j \in A \cup P$ and targets capacity $T = ||A|| + ||P||$. Let T_i be number of targets in film i. We try to find out following segments.

$$Neg = \bigcup_{1}^{F} ng_i \mid polarity\ of\ ng_i = -1,\ ng_i \in \bigcup_{j=1}^{T_i} st_i^j \tag{2}$$

Review No.	Contents	Referred	Film No.	time point
1	Yang Yang's hard temperament is enough to hold up the role	Yang Yang	Ten great II of peach blossom	2016-07-25T13:36:12.000+0800
2	Yang Yang's acting is really embarrassing, Crystal Liu is OK	Yang Yang, Crystal Liu	Ten great II of peach blossom	2017-08-07T20:18:55.000+0800
3	The story of embarrassment, feeling Yehua did not love Baiqian	Yang Yang, Crystal Liu	Ten great II of peach blossom	2017-08-07T21:09:07.000+0800
r_i	s	s	f_i	$time_i$

Table 1: review targets after named entities recognition

$$Pos = \bigcup_{1}^{F} ps_i \mid polarity\ of\ ps_i = +1,\ ps_i \in \bigcup_{j=1}^{T_i} st_i^j \tag{3}$$

Since we have to retrieve *Neg* and *Pos* shown in Equation (2) and (3), we take training data R_p with scores larger than 9 for positive sentiment and R_n which scores lower than 3 for negative one. We exclude reviews R_{ambi} which scores range between 4 and 8 that might be ambiguous over sentiment in training phrase. Instead, R_p and R_n show more concentrated sentiments in words. We train a classifier on R_p and R_n, then we split R with context window according referred target in each review. Reviews segments cover target's contextual information which construct sentiment corpus Neg and *Pos* towards specific targets are objectives for sentiment classifier. We mainly apply it on targets segments split from R_{ambi}. Finally, we get comprehensive sentiment polarity on targets segments in R.

Lexicon based sentiment analysis is constrained by sentence structure, latent word meaning and confused word features. We introduce bidirectional long short term memory (*Bi-LSTM*) neural network for end-to-end sentiment classification and automated feature learning as figure 1 shows.

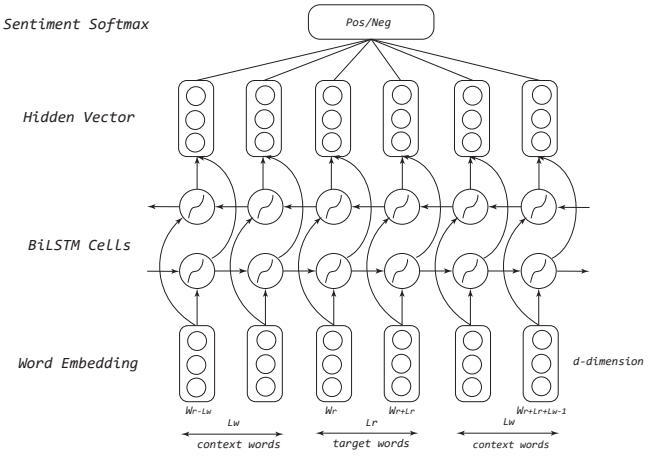


Figure 1: Structure of Bi-LSTM Sentiment Classifier, where L_r for target words length and L_w for context length we extract from original sentence. First layer in Bi-LSTM stands for forward hidden layer and the second layer are the backward one.

Specially, word sequence are represented as a low dimensional , continuous and valued vector. It is called word embedding and we use pre-trained vectors so as to make better use of semantic and grammatical associations. Assume each d-dimension vector $W_i = \mathbb{R}^{d\times|V|}$ for ith word in whole vocabulary V. Targets words with L_r length cover $tw = \{W_r, W_{r+1}, \ldots, W_{r+L_r-1}\}$ and context words $cw = \{W_{r-L_w}, \ldots, W_{r-1}, W_{r+L_r}, \ldots, W_{r+L_r+L_w-1}\}$ with window length L_w surround corresponding targets words tw. *BiLSTM* maps word vectors to fix-length sentence vector by recursively transformation above vectors of previous time step h_{t-1}. Cells for *BiLSTM* in Figure 1 contains neural gates: input gates, forget gates, and output gates which adaptively remember input vector, forget history, and generate output vector. We add softmax layer to output sentiment of sentence vector which cover hidden representation of vectors for specific target segment as Figure 1 shows. Core calculation in LSTM cells are listed in bellow equations. $\odot$ is element-wise multiplication and σ is sigmoid activation, $W_i, W_f, W_o, b_i, b_f, b_o$ are parameters for input, forget and output gates.

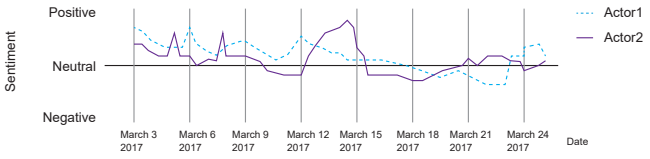
$$i_t = \sigma(W_i \bullet [h_{t-1}; w_t] + b_i) \tag{4a}$$
$$f_t = \sigma(W_f \bullet [h_{t-1}; w_t] + b_f) \tag{4b}$$
$$o_t = \sigma(W_o \bullet [h_{t-1}; w_t] + b_o) \tag{4c}$$
$$g_t = \tanh(W_r \bullet [h_{t-1}; w_t] + b_r) \tag{4d}$$
$$c_t = i_t \odot g_t + f_t \odot c_{t-1} \tag{4e}$$
$$h_t = o_t \odot \tanh(c_t) \tag{4f}$$

4.3 Film-level Sentiment Trend

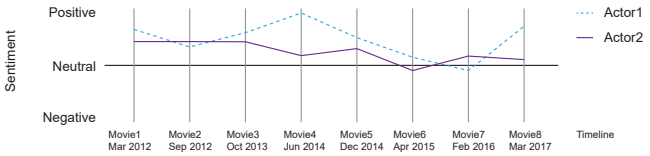
In this section, we mainly care about how to predict latent sentiment transform. Obviously, we quantify explicit sentiment towards special actors or directors. Although all information came from overall corpus, we can not deny that different sentiment exist in different time period during movie life. Audience's sentiments have trends and always behave dynamically. Previous work concentrate on sentiment polarity and we extend it by capture dynamic time-period sentiment. Denote a time period range from $time_r = [time_i, time_j]$, we capture review statistical sum R_{f_i} of each movie f_i at time period $time_r$ from *Pos* and *Neg*. Then we get different target j's positive or negative sentiment transformation at different $time_r$.

$$Sentiment_i^j(time_r) = \frac{|pos_i^j - neg_i^j\ in\ time_r|}{R_{f_i}} \tag{5}$$

Dynamic sentiment on leader creator shows how people transform their attention from different targets in movies. We pick out most interesting part (actor) from them and label it the main factor which mainly contribute to the box-office of specific film. The trend analyse is shown in Figure 2.



(a) Period Sentiment Trend: Change of sentiment in a specific film



(b) Dynamic Sentiment on Leader Creator : General sentiment transform of an actor in past five years

Figure 2: Film-level sentiment trend analyse of different time period. a) shows the change of sentiment in one film life period which indicates what attract people. b) shows the average sentiment of a specific actor which indicates highest moments and lowest moment of an actor career

5 DYNAMIC IMPACT OF ACTORS AND DIRECTORS

The goal of this module is to find who is the most bankable person in a movie and to capture the dynamic trends about actors and directors during the period of a movie released.

5.1 Contribution of Actors

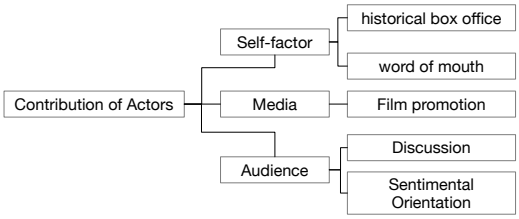


Figure 3: Composition of creative influence

The actor's own box office appeal. Actor's own box office appeal includes historical box office results and historical word of mouth performance, in which the film's historical movie box office performance can objectively reflect the creative ability of gold absorption, while the historical reputation score can objectively reflect the creative performance by Audience recognition, so this article at the same time using the history of the box office and the history of reputation to measure the creative box office call ability.

We define the The box office quality boq to measure the box office generated by the creator itself based on the historical film data. As for directors, we define $boq = wms \cdot bos$.Considering different role importance, as for actors, we define $boq_k = c_k \dot{w}ms\dot{b}os$, where k is the kth actor and c_k is the influence factor. The larger k is , the smaller c_k gets.

Media exposure. The movie's promotional period is always accompanied by a wide range of media coverage, the media's title content reflects the current public focus, if an actor or director appear in the title a lot, you can explain the public Familiarity and attention more.

Feedback from the audience. Liu mentioned the impact of word-of-mouth on the box office is very large (Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. Journal of marketing, 70 (3)). With the rapid development of the Internet, people can express their emotions on social media at any time. Thanks to the development of social media networks, the promotion of the movie has also gradually increased the proportion of online publicity. At the same time, the content of the discussion of the movie has also formed a considerable amount of data. The movie review often includes the concept of " If you can get the audience's idea of fittfittan actor or master from the commentary, you can see whether the source of the audience's emotional inclination toward the movie comes from the influence of the movie's main actor (including the protagonist).

5.2 Dynamic Impact of Leader Creator

In the short life of the movie, the contribution of the genre to the box office is obviously not constant. With the development of the Internet media, real-time feedback of the movie users on the movie will affect the watching desire of the non-movie users, and the positive contribution Refers to the potential box office or the desire to watch the increase; the negative is the box office to bring the negative image, to dispel the wishes of watching. Therefore, in the film's life, the star effect on the box office's contribution is dynamic, the project will be divided into the life of the movie before the release, the first week of release, the second week of release, the third week of release, released the fourth Friday A period, respectively, to explore the five periods of major box office contributions.

6 BOX-OFFICE PREDICTING

6.1 Premiere week box-office predicting

6.2 Phased Box-office Predicting

In the past prediction of box office, we usually only forecast the overall box office, but often ignored the movie's influence on movie box office due to the audience's attitude. Such as "wolf 2", the movie box office to be among the world's top 100, because in the release process, because the audience warmly, attracted the audience is not the original film (film series the fans, fans, like the kind of audience) to watch a movie, which leads to the box office continued to rise however, the traditional model is unable to capture this phenomenon. Therefore, from the pre launch to the 1 months after the release, we set up the box office prediction model with the

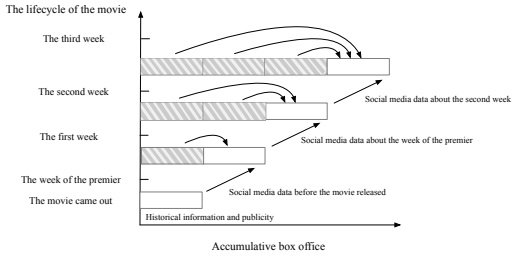


Figure 4: Multi-stage box office prediction model

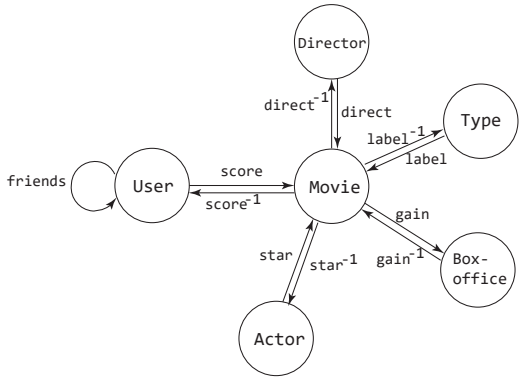
weekly variation of the weekly release to predict the box office for the first week, the box office for second weeks, the third week box office and the box office at the fourth week.

Symbol	Description
sc	The mean bos of well-known actors and directors
gsc	The mean wms of well-known actors and directors
std	The variance bos of well-known actors and directors
gstd	The variance wms of well-known actors and directors
$weibo_i$	The ratio of creative comments and all comments until the ith week
$week_i$	The box office in the ith week

Table 2: Features

7 MINING MOVIE HETEROGENEOUS NETWORK

(prupose…) A movie heterogeneous information network (*MHIN*) is a domain information network with multiple types of entities and relations. We construct our *MHIN* with multiple objects such as movies (M), directors (D), users (U), Actors (A), Types (T) and box-offices (B). Figure 5 shows typical *MHIN* schema defined by us. Links exist between users and movies denoting score and score-by relations, between movies and actors(directors) denoting star(direct) and star-by(direct-by) relations, between movies and types denoting label and (labeled) relations, between movies and box-offices denoting gain and gain by relations. Also, we can add other attributes into the movie heterogeneous network (e.g years (Y), producers (P)). Meta-path is a connect relation between two types of objects in *MHIN*. Denote network schema for our *MHIN* is $\mathcal{S} = (\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ represent object types and $\mathcal{R}$ indicate different types of relationships. A meta-path $\mathcal{P}$ is defined in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_k} A_{k+1}$. Length of $\mathcal{P}$ is defined by number of relations it contains. Semantic composite relations are implied in meta-paths. For example, we can evaluate the similarity of movies or actors by length-2 meta-path MDM (movie-director-movie), MAM (movie-actor-movie) or AMA (actor-movie-actor), AMTMA(actor-movie-type-movie-actor).



(a) Structure of movie heterogeneous network

Figure 5: Social network schema of MHIN.

7.1 Rank Based Cluster For Actors

In this section, we introduce a way to make distinguish between A-list and B-list actors. We prefer applying cluster based classifier on actors popularity recognition. Strong available mutually reinforcing relations between cluster and rank are helpful to deal with different levels of actors. We defined recursion formula by following empirical rules.

RULE 7.1. *high rank actors star more high rank movies*

RULE 7.2. *high rank movies attract more high rank actors*

RULE 7.3. *actors get high rank with high rank co-starers*

Suppose we want K clusters of actors. According the star relationship between actors and movies, we define matrix $M_{MA}(i,j) = c_{ij}$ representing actor j's contribution of movie i (see Section 3.1) and similarly, $M_{AA}(i,j) = m_{ij}$ stands for number of movies that actor i and actor j co-stared. Note that $W_{AM} = W_{MA}^T$, and $i = \{1, 2, \dots, m\}, j = \{1, 2, \dots, n\}$. According the 3 rules we proposed, we have

$$r_A(i) = \alpha \sum_{i=1}^{m} W_{AM}(j,i) r_M(i) + (1-\alpha) \sum_{j=1}^{n} W_{AA}(i,j) r_A(j) \quad (6a)$$

$$r_A(j) \leftarrow \frac{r_A(j)}{\sum_{j'=1}^{n} r_A(j')} \quad (6b)$$

$$r_M(i) = \sum_{j=1}^{n} W_{MA}(i,j) r_A(j) \quad (6c)$$

$$r_M(i) \leftarrow \frac{r_M(i)}{\sum_{i'=1}^{m} r_M(i')} \quad (6d)$$

Note that $\alpha \in [0,1]$ is a believe factor of weighed component of rule 3, $r_A(j)$ and $r_M(i)$ are normalized rank score vector. We

finally get r_A which should be primary eigenvector of $\alpha W_{AM} W_{MA} + (1-\alpha) W_{AA}$. Further, we capture posterior probability $\pi_{i,k}$ that a_i from cluster k, Once an actor acts a movie, he is more likely to star high ranked film and for movie, its success are more likely contributed by high rank actor. Thus we have K dimensional vector $s_{a_i} = \{\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,K}\}$ where $\pi_{i,k}$ denotes a_i's coefficient for component k.

The cluster center and the distance between each actor and each cluster can be defined as

$$S_{A_k} = \frac{\sum_{a \in A} s(a)}{|A_k|} \tag{7a}$$

$$Distance(a, A_k) = 1 - \cos(s_{a_i}, S_{A_k}) \tag{7b}$$

Algorithm 2 RankClus for Actors

Input:

Our movie information network $MHIN = (M, A; W)$

Cluster Number K

Output:

K clusters of actors A_i and rank of actor in each cluster

1: iter = 0;
2: Init partitions for A, get $PA^{iter} = \{A_i^{iter}\}_1^K$;
3: Repeat following until $PA^{iter} - PA^{iter-1} < \epsilon$ or iterations reach limitation
4: For all cluster calculate r_A followed by rank function
5: Evaluate Θ for mixture model and get component efficient estimations s_{a_i} for each actor a_i
6: Update centers $S_{A_k}^{iter}$ of each cluster A_k
7: Reassign each actor a_i according distance between a_i and each cluster center A_k^{iter}

7.2 Alternative Actors

In this section, we discuss method choosing alternative actors for casting agents. The basic idea is measure similarity between actors. Since movie information network has been built, we then introduce meta path-based similarity framework for alternative actors.

Type	Path instance	Meta-path
I(AMA)	Andy-M_1-Sara	Actor-Movie-Actor
II($AMTMA$)	John-M_2-Comedy-M_3-Sara Ben-M_4-War-M_3-Sara Drew-M_5-Dracula-M_2-John	Actor-Movie-Type-Movie-Actor
III($AMDMA$)	Diana-M_6-Spielberg-M_1-Andy John-M_2-Luc Besson-M_5-Drew Sara-M_3-Tom Tykwer-M_5-Drew	Actor-Movie-Director-Movie-Actor

Table 3: Different Types of meta-path in MHIN

7.3 Co-star Relationship Prediction

8 RELATED WORK

9 DISCUSSION

10 CONCLUSION