A Constrained Multi-view Clustering Approach to Influence Role Detection

Chengyao Chen[1], Dehong Gao[2], Wenjie Li[1], Yuexian Hou[3]
[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[2]1688cn, Alibaba.INC(Hangzhou), China
[3]School of Computer Science and Technology, Tianjin University, China
cscchen@comp.polyu.edu.hk, dehong.gdh@alibaba-inc.com,
cswjli@comp.polyu.edu.hk, yxhou@tju.edu.cn

Abstract

Twitter has provided people with an effective way to communicate and interact with each other. It is an undisputable fact that people's influence plays an important role in disseminating information over the Twitter social network. Although a number of research work on finding influential users have been reported in the literature, they never really seek to distinguish and analyze different influence roles, which are of great value for various marketing purposes. In this paper, we move a step forward to further detect five recognized influence roles of Twitter users with regard to a particular topic. By exploring three views of features related to topic, sentiment and popularity respectively, we propose a novel constrained multi-view influence role clustering approach to group potential influential Twitter users into five categories. Experimental results demonstrate the effectiveness of the proposed approach.

1 Introduction

Nowadays, Twitter has become one of the most popular social media platforms for people to share information and communicate with each other. It creates more and more new business opportunities with a variety of online marketing activities [Anagnostopoulos *et al*., 2008]. Recent years have witnessed that an increasing number of enterprises have started to attach importance to locating favorable influential users and manipulating their opinions to attract potential customers or improve sales. Understanding social influence over large-scale networks is crucial to business marketing management.

Although all influential users perform influence on others, [Brown and Hayes, 2008]has verified that the way people use to influence others varies and produces different effect. Someone always strongly praises a product and persuades others to buy. Someone changes others' opinions on a product with professional analysis. Someone timely informs others the latest news of a product. And someone promotes the product by popularity. It is quite clear that different influential users play different influence roles. Meanwhile, a company may have different objectives in different promotion stages and needs users with different influence roles to conform to [Brown and Hayes, 2008]. For example, a company which targets to improve product brand awareness may want to choose the users with high popularity to help with. However, for a company whose product quality is questioned by customers, it may be a better choice to invite domain experts who have professional knowledge to explain and convince. Selecting influential users with appropriate influence roles in accordance with specific marketing objectives is more effective than just seeking for the most influential ones in general.

Despite the importance of influence role, previous work mostly emphasizes on measuring the general influence power of a user on others through the information of the netwrok structure[Cha *et al*., 2010; Weng *et al*., 2010], or maximizing the influence propagation which assists companies to find the proper set of people to promote products [Kempe *et al*., 2003; Chen *et al*., 2009]. Without any exception, they all take the influence as the same type. The lack of considering the effects of different influence roles on different marketing objectives will inevitably hinder the companies from proposing more suitable marketing strategies. This motivates us to further analyze and detect different influence roles of users, which could be used to further extend the previous work in achieving different marketing goals. [Chen *et al*., 2014] proposed the idea to distinguish different types of influential users, but lacked compelte study on how to detect them.

Table 1. Five categories of influence role.

Role Category	Influence Manner	Marketing Effect
Enthusiast	Support and defend products	Improve sales
Information Disseminator	Publish product information	Enhance brand memorability
Expert	Gather facts and professional opinions	Improve reputation
Celebrity	Popular among people	Improve awareness
Others	Show no obvious influence	None

Copyright © 2015 for the individual papers by the papers' authors. Copying permitted for private and academic purposes. This volume is published and copyrighted by its editors

To better characterize influence roles, we define five distinct categories with reference to the definition in the WOMMA's influencer guidebook (www.womma.org/influencers). They are enthusiast, information disseminator, expert, celebrity and others. The brief descriptions of them are summarized in Table 1. We can clearly see that one's influence role is largely determined by his/her behaviors and personal characteristics, but not totally dependent on how much influence he/she has. Different from previous work that measures users' influence mainly based on social connections, we summarize three aspects that help to distinguish influence roles, including the interest to a topic (e.g., enthusiast, information disseminator and expert pay more attention than the other two), the attitude to the topic (e.g., enthusiast always praises, expert sometimes praises and sometimes not) and the popularity over the social network (e.g., celebrity has more followers). Accordingly we extract three views of features, i.e., the topic view, the sentiment view and the popularity view from users' posts and profiles for influence role detection.

We also note that each view can only partially reflect the influence role from its own perspective. However, when they complement with each other, the three views together provide more complete information for influence roles. Based on the three-view user representations, we propose a novel Constrained Multi-view Influence Role Clustering (CMIRC) approach upon an optimization framework to partition influential users into five recognized categories. Unlike other existing multi-view clustering approaches, CMIRC allows the cluster numbers in the different views to be different and so provides more flexibility for integrating data from multi-views. It connects the local clustering information from each individual view and the global multi-view clustering results with a local-global mapping mechanism.

Another advantage of CMIRC is its capability to incorporate the prior knowledge based upon the semi-supervised learning framework. Actually, it is very common that the influence roles are known to a small number of users who are easily identified by a company. Then people can use such information as the prior knowledge to find out many others for their needs. To incorporate the prior knowledge to guide clustering, we apply two kinds of group-level constraints, the same-cluster constraints and the different-cluster constraints, to define which groups of users must be or must not be in the same cluster. The experimental results demonstrate the effectiveness of CMIRC when compared with other single-view and multi-view clustering approaches

2 Influence Role Detection

2.1 Three-View User Representation

Topic-view Representation

The motivation of using topic view is the intuition that different roles may have different degrees and different focuses of attention to the topic. To start with, a word like "*iPhone*" is selected as the topical word. Then, measured by the mutual information, the *K* most relevant words that co-occur with the topical word within a window of size two are extracted as keywords to form the topic profile collectively. These *K* words provide a more complete picture of the topic than the topical word itself. For all the tweets of a given user, a topical vector weighed by tf-idf is built to capture his/her word distribution over the extracted keywords.

Sentiment-view Representation

The sentiment view reveals the preferred attitudes when a user expresses his/her opinions and tends to differentiate among the enthusiast who often posts tweets with positive sentiments, the disseminator whose tweets is mainly neutral ones and the expert whose opinions may be either positive or negative. To measure the sentiment of users, the lexicon AFINN (http://www2.compute.dtu.dk/~faan/data/) is used, where each word is attached with an integer value between negative five and positive five, denoting its sentiment polarity and strength. Based on this lexicon, the positive/negative sentiment scores of a tweet are calculated by aggregating the sentiment strengths of all the positive/negative words it contains. The sentiment view representation of a user is then defined as the average positive-sentiment score and average negative-sentiment score of all his/her tweets

Popularity-view Representation

Apart from the interests and attitudes to a topic, the popularity (or to say the authority) of a user can also imply the influence role in some extent. Three features are selected including the number of followers, the number of followees and a binary value indicating whether a user account is verified or not. The popularity view tends to distinguish the people with different levels of popularity like celebrities and enthusiasts.

2.2 Constrained Multi-view Influence Role Clustering

To better use the data collected from multiple sources, multi-view clustering approaches partition data into clusters by integrating features from multiple views. They have been successfully applied to image recognition and text mining, etc. [Bickel *et al*., 2004; Cai *et al*., 2013; Liu *et al*., 2013]. These approaches share a common assumption, i.e., the features from each single view are complete for clustering, yet better clustering performance can be expected by exploring the rich information among multiple views. Naturally, the cluster numbers of different views are often supposed equal to the final multi-view cluster number. From the previous analysis, however, we believe that it is more reasonable and practical to allow the cluster numbers of different views to be different for influence role detection. As a result the clustering results in each view will be also different from the ultimate clustering results. To this end, we develop a Constrained Multi-view Influence Role Clustering (CMIRC) approach to group data into different numbers of clusters in individual views (i.e., local clusters) and utilize the mapping matrix to bridge the gap between the single-view clusters and the multi-

view clusters (i.e., global clusters). The introduction of the mapping matrix is one of the main contributions of this work.

Another advantage of this approach is its semi-supervised framework that allows us to incorporate the prior knowledge easily. Say, we can take a small number of users whose influence roles are manually labeled as the prior knowledge to guide the clustering of others. To incorporate the prior knowledge into CMIRC, we employ two kinds of group-level constraints [Law *et al*., 2004] to define which group of users must be or must not be in the same cluster. Specifically, the same-cluster (SC) constraints include several groups of users and the users in each group must belong to the same cluster, either local or global cluster. The different-cluster (DC) constraints contain several group pairs and the users in the two groups of a pair cannot be in the same cluster.

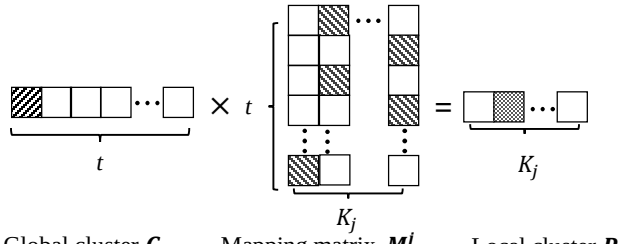
To better describe our approach, let's start with a variant K-means clustering algorithm which utilizes data from multiple sources [Cai *et al.*, 2013]. Let $U = \{u_1, u_2, \dots, u_n\}$ represents *n* Twitter users. Each user u_i is represented by m views of features, $X_i = \{X_i^1, X_i^2, \dots, X_i^m\}$, where the *j*-th element X_i^j represents the features of view *j*, and it is a row vector containing d_j elements. Then a typical multi-view clustering task can be formulated as the following optimization problem.

$$\min_{P,C} \sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_j \left\| X_i^j - P_{ij} C^{j^T} \right\|_2$$

$$s.t. \sum_{k=1}^{K^j} P_{ijk} = 1, P_{ijk} \in \{0,1\}, \forall i = 1,2,\dots,n, \sum_{j=1}^{m} \alpha_j = 1$$

Similar to K-means, $P_{ij} \in \boldsymbol{R}^{1 \times K^j}$ here describes the cluster indicator for user u_i in view *j*. It also represents the local clustering results. K^j and $C^j \in \boldsymbol{R}^{d^j \times K^j}$ denote the cluster number and cluster centers in view *j*. α_j is a factor to balance the weight of view *j*. If the cluster numbers in all *m* views are the same (i.e., K^j =*t*, where *t* represents global cluster number), P_{ij} for all the views should be consistent. This implies that the local clustering results in every view are equal to the global clustering results. However, with our assumption, the cluster number in each view is different, so we cannot derive the global clustering results directly from P_{ij}. In order to connect local clustering and global clustering together, we transform the local clustering results P_{ij} in view *j* into the combination of global cluster assignment $G_i \in \boldsymbol{R}^{1 \times t}$ and a mapping matrix $M^j \in \boldsymbol{R}^{t \times K^j}$.

Figure 1. Illustration of global and local cluster mapping



Global cluster $\boldsymbol{G_i}$ Mapping matrix $\boldsymbol{M^j}$ Local cluster $\boldsymbol{P_{ij}}$

Figure 1 explains how u_i 's local cluster in view *j* corresponds his global cluster. Assume that u_i belongs to the first global cluster as presented in G_i and the mapping matrix describes the first global cluster is mapped to the second local cluster. Then u_i should be found in the second local cluster.

Apart from the use of mapping matrix, two types of constraints are also integrated into CMIRC. They are defined through user groups $ug_i = \{u_1, u_2, \dots, u_{n(ug_i)}\}$, where $n(ug_i)$ is the number of users in this user group ug_i. Same-Cluster constraints are a set of user groups, i.e., $SC = \{ug_1, ug_2, \dots, ug_l\}$. The users in each SC group must be assigned to the same cluster. Different-Cluster constraints are a set of user group pairs, i.e., $DC = \{p_1, p_2, \dots, p_r\}$ and $p_k = \langle ug_i, ug_j \rangle$. The users in two different groups of a pair in DC must belong to different clusters. All users in SC and DC compose U_{con}. Compared with the pair-wise constraints, during cluster assignment, we could assign a cluster to the whole group without the need to assign users to clusters one by one. Such a strategy avoids computational complexity in the optimization procedures introduced later.

Finally, CMIRC that partitions the users U into t clusters with *m*-view features constrained by SC and DC can be formulated by the following optimization problem

$$\min_{G,M,C} \sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_j \left\| X_i^j - G_i M^j C^{j^T} \right\|_2 \quad (1)$$

$$s.t. \sum_{k=1}^{t} G_{ik} = 1, G_{ik} \in \{0,1\}, \sum_{j=1}^{m} \alpha_j = 1,$$

$$\sum_{k=1}^{K_j} M_{ik}^j \geq 1, \ \forall i = 1, \dots, t, \sum_{i=1}^{t} M_{ik}^j = 1, \forall k = 1, \dots, K_j,$$

$$M_{ik}^j \in \{0,1\}, \forall u_i, u_j \in ug_k \wedge u_i \neq u_j, G_i = G_j,$$

$$\forall \langle ug_q, ug_p \rangle \in DC \wedge \forall u_i \in ug_q \wedge \ \forall u_j \in ug_p, G_i \neq G_j$$

where G_i represents the global cluster assignment for user u_i which satisfies 1-of-K coding scheme. C^j is the local cluster center in the *j*-th view and M^j is the mapping matrix. M^j satisfies the constraints that every local cluster must be mapped to at least one global cluster and every global cluster must be mapped to one and only one local cluster.

In order to solve this optimization problem, we rewrite the objective function in Equation (1) as Equation (2), and apply the following iterative updating process to solve it.

$$O = \min_{G,M,C} \sum_{j=1}^{m} \alpha_j H^j, \quad (2)$$

where $H^j = Tr\{(X^{j^T} - C^j M^{j^T} G^T) D^j (X^{j^T} - C^j M^{j^T} G^T)^T$, and D^j is the degree matrix derived from E^j.

$$d_{ii}^j = \frac{1}{2\|e^{ji}\|}, \forall i = 1,2, \dots n, \text{and } E^j = X^j - G M^j C^{j^T} \quad (3)$$

- **Fix G, M^j, D^j and update local cluster center C^j**

As stated before, the combination of G_i and M^j represent local cluster results. In this step, the local cluster centers are

updated by minimizing the distances from users to their corresponding clusters. It is solved by differentiating the objective function in Equation (2) for each view with respect to C^j. The optimal solution of C^j is obtained by setting the derivation to zero, which gives us

$$C^j = \alpha_j X^{j^T} D^j G^j M^j (\alpha_j M^{j^T} G^T D^j G M^j)^{-1} \quad (4)$$

- **Fix M^j, C^j, D^j and update global cluster assignment G**

We update G through each row of its, G_i in the following order. First we update G_i for the users who are not constrained separately, and then update G_i for the users who are in U_{con} together. In particular, if the user u_i is not constrained, we locate the local cluster for each user through the mapping matrix. Then, what we need to do is to find G_i from its limited solutions that minimize the sum of distances between it and the center of its assigned local cluster for each view, as presented in the Equation (5).

$$G_i = \underset{G_i}{\operatorname{argmin}} \sum_{j=1}^{m} \alpha_j \left\| X_i^j - G_i M^j C^{j^T} \right\|_2 \quad (5)$$

Constrained by SC and DC, we give each user group in U_{con} a global cluster assignment, i.e., $G_{con(ug_i)}$, a row vector that represents the assignment for users in user group ug_i in SC. By concatenating the assignment vectors for each user group, we form a certain number of candidate assignment matrixes that guarantee the DC constraints in column. From all these candidates, the one that minimizes the objective function in Equation (6) is defined as G_{con},

$$G_{con} = \underset{G_{con}}{\operatorname{argmin}} \sum_{j=1}^{m} \sum_{i=1}^{l} \sum_{k=1}^{n(ug_i)} \left\| X_{ug_{ik}}^{j\ T} - G_{con(ug_i)} M^j C^{j^T} \right\|_2 \quad (6)$$

where $X_{ug_{ik}}^j$ are the *j*-view features of user u_k who is in group ug_i. Then the global cluster assignment G_k for user u_k in user group ug_i is regarded as $G_{con(ug_i)}$.

- **Fix G, C^j, D^j a, and update global and local cluster mapping matrix M^j**

M^j is the mapping matrix between global and local clusters. For each view, based on the constraints for M^j, we construct candidate mapping matrixes and possible choices for the local cluster assignment by transforming from the global cluster assignment G. The one that assigns users to the best local clusters to guarantee the overall minimized distance over all the users is selected to be the updated mapping matrix.

$$M^j = \underset{M^j}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| X_i^j - G_i M^j C^{j^T} \right\|_2 \quad (7)$$

- **Fix G, C^j, M^j and update D^j**

D^j is introduced to aid solving the optimization problem in Equation (4) and it can be calculated directly from G, C^j, M^j according to Equations (3).

Of the four steps in CMIRC iterations, three are convex problems related to one variable. It can be proved that each is guaranteed to converge to an optimal solution. Once the global clusters are ready, we select the labels of the users who are constrained by Same-Cluster as the influence roles of these clusters.

3 Experiments and Discussion

Four topics about well-known electronic products, "iPhone", "Samsung Galaxy", "Xbox" and "PlayStation" are selected to construct the experimental datasets. We collect the tweets that contain the topical word like "iPhone" from 3rd to 30th April 2014. Among users who post these tweets, the ones who have more than 500 followers and have been re-tweeted at least once are regarded as influential users. The size of an influential user pool ranges from 4912 (for Samsung Galaxy) to 90906 (for iPhone). To be consistent, 4912 influential users are sampled for each topic. These users together with their tweets and account information are used in the experiments.

Due to the lack of annotated datasets, for each topic we randomly select 200 from 4912 influential users and invite human annotators to label their influence roles for evaluation purpose by providing users' posts and their account information. The numbers of the annotated users across five influence roles are presented in Table 2. We randomly choose 1/5 users of each influence role to build the constraints required by CMIRC, and the rest are used for evaluation.

Table 2. Evaluation data on four topics

Topic \ Role	Enthusiast	Information Disseminator	Expert	Celebrity	Others
iPhone	9	31	13	20	127
Galaxy	21	32	15	19	113
Xbox	20	25	14	15	126
PlayStation	13	29	15	14	129

We compare CMIRC with (1) Baseline K-means clustering (BKC) and Constrained K-means clustering (CKC) that concatenates three views together; (2) two existing multi-view clustering approaches, i.e., Multi-view K-means Clustering (MKC) [Cai *et al*., 2013] and Negative Matrix Factorization (NMF) based Multi-view Clustering (NMFMC) [Liu *et al*., 2013]. To further understand the contribution of each view, we also compare with (3) Constrained Single-View K-means Clustering (CSC_{topic}, $CSC_{sentiment}$ and $CSC_{account}$) and (4) Constrained Two-View K-means Clustering (**C**$MIRC_{ts}$, **C**$MIRC_{sa}$ and **C**$MIRC_{ta}$). In addition, (5) CMIRC without constrains (MIRC) is also compared. Three commonly-used metrics are used to evaluate performances. They are macro-average precision (MP), macro-average recall (MR), and macro-average F-measure (MF).

For CMIRC, we compare different settings of cluster number for each view from 2 to 5 to find the one with best F-measure. For the topics "iPhone" and "PlayStation", (3, 2, 3) for topic, view, sentiment view and popularity view is the best one, while for the topics "Samsung Galaxy" and "Xbox", (3, 5, 5) is the best one. The cluster number for each view on two-view clustering CMIRC and MIRC are also set the same as CMIRC. And for BKC, CKC and CSC, the cluster number is set the same as the global cluster number 5.

Table 3: Performance evaluation

Approach \ Topic		iPhone			Galaxy			Xbox			PlayStation		
		MP	MR	MF	MP	MR	MF	MP	MR	MF	MP	MR	MF
Combined view	BKC	0.2551	0.3526	0.1551	0.2725	0.2575	0.1443	0.2063	0.2340	0.0901	0.2964	0.3095	0.1467
	CKC	0.3366	0.3518	0.2649	0.3529	0.3314	0.1526	0.2419	0.2152	0.1585	0.2803	0.3407	0.2026
Multi-view	NMF**M**C	0.3627	0.4371	0.3465	0.3497	0.3568	0.2154	0.2874	0.2839	0.2770	0.3892	0.3170	0.2812
	MKC	0.3404	0.2983	0.1979	0.4132	0.3253	0.3155	0.3035	0.2960	0.2436	0.3333	0.3393	0.2328
CMIRC		**0.4670**	**0.5056**	**0.4020**	**0.4914**	**0.3417**	**0.3616**	**0.4338**	**0.3337**	**0.3207**	**0.4031**	**0.3676**	**0.3531**
MIRC		0.4200	0.3731	0.3730	0.4012	0.3126	0.3298	0.3352	0.3074	0.2914	0.3752	0.3477	0.3065
Constrained Single-view	**C**SC$_{topic}$	0.2667	0.4552	0.2546	0.2367	0.3585	0.1542	0.1957	0.1898	0.1530	0.2745	0.2176	0.1809
	CSC$_{sentiment}$	0.2357	0.2527	0.1341	0.2256	0.1087	0.1417	0.2141	0.2036	0.1436	0.2044	0.2064	0.1007
	CSC$_{account}$	0.2628	0.3525	0.2270	0.3108	0.3179	0.2868	0.3236	0.1160	0.1376	0.2520	0.2088	0.1133
Constrained Two-view	**C**MIRC$_{ts}$	0.2812	0.3240	0.1746	0.2977	0.2065	0.1879	0.4183	0.2761	0.2781	0.2971	0.2156	0.1903
	CMIRC$_{sa}$	0.2988	0.4559	0.3050	0.4386	0.3435	0.2917	0.2850	0.2630	0.2439	0.2466	0.2231	0.1993
	CMIRC$_{ta}$	0.3555	0.4230	**0.3220**	0.4066	0.3330	**0.2987**	0.3908	0.2952	**0.2879**	0.3419	0.2578	**0.2474**

For another parameter α_j, it is set to make all single views have balanced contributions to the final clustering results. We compute α_j based on the average ℓ_2-norm distance, i.e., dis_j, between a user and all other users in view j. α_j is negatively related to dis_j. That is,

$$dis_j = \sum_{i=1}^{n}\sum_{k=i+1}^{n}\left\|X_i^j - X_k^j\right\|_2$$

and (8)

$$\alpha_1 dis_1 = \alpha_2 dis_2 = \alpha_3 dis_3, s.t. \sum_{j=1}^{3}\alpha_j = 1$$

This gives us (0.177, 0.621, 0.202), (0.086, 0.588, 0.326), (0.093, 0.604, 0.303) and (0.180, 0.618, 0.202) for the topics "iPhone", "Samsung Galaxy, "Xbox" and "PlayStation". The parameters α for CMIRC on two-view clustering are set analogously. The parameters used in MKC and NMFMC to balance the relative weights among different views are also turned for their best performance. The constraints in all the constrained approaches are used in the same way. We assign the labels of constrained users as the roles of the corresponding clusters for all constrained clustering approaches. For BKC, MKC and NMFMC, we choose the assignment that maximizes the MF as the mapping of the clusters to the influence roles. We repeat the experiments for all the approaches 10 times using random initialization and present their average performance in the Table 3. The performance of the proposed CMIRC consistently beat all others in all three metrics.

Besides, we note that all multi-view clustering approaches outperform the baseline BKC, and CMIRC beats the CKC. It demonstrates the power of multi-view clustering approaches and verifies that representing data in different views actually works for influence role detection. However, comparing different multi-view clustering approaches, CMIRC and even MIRC without constraints get more accurate results. It proves the rationality of our assumption that each view can only represent partial information, and by employing the insufficient views together, we infer better global clustering results. Meanwhile, we also see that CMIRC performs better than MIRC that lacks of prior knowledge, which proves that building appropriate constraints to model the different influence role demands from a company is important. Moreover, By comparing constrained clustering approaches with single-view, two-view, and three-view, we observe that the performance gets better when more views are involved. It shows that the three views including topic, sentiment and popularity views are all necessary to identify influence roles. At last, in three single-view constrained K-means clustering approaches, it is difficult to distinguish which view is better. However, when compare three two-view constrained clustering approaches, we find that the combination of the topic view and the popularity view performs the best, followed by the combination of the sentiment view and the popularity view. The importance of user's popularity in identifying influence roles is clear. Meanwhile, topic view and sentiment view are still important and necessary to supplement the popularity view.

To provide a more intuitive understanding of what are the users with each influence role look like, we provide the cluster centers to illustrate the characteristics of each role in each view in Tables 4 to 6. We present the five most representative (popular) words used by the users in each role and the ratio of average positive score and positive score of all the users belong to the same role. The ratio is bigger if in general people are more positive. We also give the average numbers of followers and followees, and the percentage of the verified accounts for reference. The general feelings from the topic view analysis are (1) enthusiasts and celebrities tend to share their own experiences and assessments with the words like "buy" and "love"; (2) experts who care more about specific aspects like to mention the detailed words such as "charger" and "battery"; (3) the general words like "news" and "mobile" are often used by information disseminators who pass the latest news to people. From the sentiment view analysis, we do observe a significant trend that in general enthusiasts express more positively while information disseminators hold more neutral sentiment. We can also see that the popularity of celebrity is pretty high and it alone is able to pick out celebrities easily.

4 Conclusion

In this work, we address the issue of influence role detection.

We propose a Constrained Multi-view Influence Role Clustering (CMIRC) approach to partition Twitter users into five clusters with three views of features (i.e., topic view, sentiment view and popularity view). In CMIRC, different cluster numbers are allowed for different views and the constraints are used to model the prior information. The results indicate the effectiveness of our proposed approach. In the future, we will continue to explore more features to capture their actual marketing effects on their followers.

Table 4: Role characteristics on topic view

	Enthusiast	**Information Disseminator**	**Expert**
iPhone	love, real, gaming, screen, battery	news, apple, charger, battery, selling	news, apple, charger, battery, selling
Galaxy	win, chance, space, buy, s5 chanlle	fingerprint, android, 5s, tech, launch	fingerprint, android, 5s, tech, launch
Xbox	play, game, enter, buy, lol	360, ps4, Microsoft, tv, coming	white, china, flaw, security, sales
PlayStation	Game, play, win, lol, awesome	Xbox, sony,coming, update,release	sales, code, confirm, communiyy, console

Table 5: Role characteristics on sentiment view

View \ Topic		**iPhone**				**Galaxy**			
		Enthusiast	**Information Disseminator**	**Celebrity**	**Expert**	**Enthusiast**	**Information Disseminator**	**Celebrity**	**Expert**
Sentiment	Positive	1.0702	3.34E-05	1.0702	0.0816	2.0	2.33E-08	0.7978	0.9997
	Negative	0.0772	3.74E-05	0.0772	0.1680	3.25E-09	1.58E-08	0.112	0.0003
		Xbox				**PlayStation**			
sentiment	Positive	1.1178	3.78E-07	0.8545	0.0871	3.0	1.1426	0.0002	1.1426
	Negative	0.083	2.45E-07	0.0898	1.110	7.38E-10	0.0852	0.0004	0.0852

Table 6: Role characteristics on popularity view

View \ Topic		**iPhone**				**Galaxy**			
		Enthusiast	**Information Disseminator**	**Celebrity**	**Expert**	**Enthusiast**	**Information Disseminator**	**Celebrity**	**Expert**
Popularity	Follower	1800	1800	129968	1800	3157	2893	63537	2994
	Followee	857	857	1866	857	997	1032	1538	995
	isVerified	0	0	0.9999	0	0	0	0.9999	0
		Xbox				**PlayStation**			
Popularity	Follower	3326	4066	185459	2666	1680	1680	169180	1680
	Followee	905	1077	1448	968	365	365	738	365
	isVerified	0	1.05E-06	1	0	2.24E-07	2.24E-07	1	2.24E-07

Acknowledgments

The work described in this paper was supported by the grants from the Research Grants Council of Hong Kong (PolyU 5202/12E and PolyU 152094/14E) and a grant from the National Natural Science Foundation of China (61272291).

References

[Anagnostopoulos *et al*., 2008] Anagnostopoulos Aris, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *SIGKDD 2008*. pages 7-15. ACM, 2008.

[Bickel *et al*., 2004] Bickel Steffen and Tobias Scheffe. Multi-View Clustering. In *ICDM 2004*, pages 19-26. IEEE, 2004.

[Brown and Hayes, 2008] Brown D and Hayes, N. Brown, Duncan, and Nick Hayes. Influencer Marketing: Who Really Influences Your Customers? Butterworth-Heinemann, Oxford. 2008.

[Cai *et al*., 2013] Cai Xiao, Feiping Nie, and Heng Huang. Multi-view K-means Clustering on Big Data. In *IJCAI 2013*, pages 2598-2604. AAAI Press, 2013.

[Cha *et al*., 2010] Cha Meeyoung, Haddadi Hamed, Benevenuto, Fabrıcio, Gummadi Krishna P. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM 2010*, pages 10-17. AAAI Press, 2010.

[Chen *et al*., 2009] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *SIGKDD 2009*, pages 199-208. ACM, 2009.

[Chen *et al*., 2014] Chengyao Chen, Dehong Gao, Wenjie Li, Yuexian Hou. "Inferring topic-dependent influence roles of Twitter users." In *SIGIR 2014*, pages 1203-1206 ACM, 2014.

[Kempe *et al*., 2003] Kempe David, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *SIGKDD 2003*, pages 137-146. ACM, 2003.

[Law *et al*., 2004] Law Martin HC, Alexander Topchy, and Anil K. Jain. Clustering with Soft and Group Constraints. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 662-670.2004.

[Liu *et al*., 2013] Liu Jialu, Wang Chi, Gao Jing and Han Jiawei. Multi-view Clustering via Joint Nonnegative Matrix Factorization. In *SIAM 2013* , pages 252-260.2013.

[Weng *et al*., 2010] Weng Jianshu, Lim Ee-Peng, Jiang Jing and He Qi. TwitterRank: Finding Topic-sensitive Influential Twitterers. In *ICWSM 2010*, pages 261-270. ACM, 2010.