Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Overlapping community detection combining content and link

Zhou-zhou HE[1], Zhong-fei (Mark) ZHANG[‡1], Philip S. YU[2]

(*[1]Zhejiang Provincial Key Laboratory of Information Network Technology,
Department of Information Science and Electronic Engineering,
Zhejiang University, Hangzhou 310027, China*)

(*[2]Department of Computer Science, University of Illinois at Chicago, IL 60607, USA*)

E-mail: {zju_hzz, zhongfei}@zju.edu.cn; psyu@uic.edu

Received Mar. 5, 2012; Revision accepted July 9, 2012; Crosschecked Oct. 12, 2012

Abstract: In classic community detection, it is assumed that communities are exclusive, in the sense of either soft clustering or hard clustering. It has come to attention in the recent literature that many real-world problems violate this assumption, and thus overlapping community detection has become a hot research topic. The existing work on this topic uses either content or link information, but not both of them. In this paper, we deal with the issue of overlapping community detection by combining content and link information. We develop an effective solution called subgraph overlapping clustering (SOC) and evaluate this new approach in comparison with several peer methods in the literature that use either content or link information. The evaluations demonstrate the effectiveness and promise of SOC in dealing with large scale real datasets.

Key words: Overlapping, Content, Link, Community detection

doi:10.1631/jzus.C1200049 **Document code:** A **CLC number:** TP391

1 Introduction

Recently, community detection has been a hot research topic in data mining with many important applications such as collaborative filtering (Chen *et al.*, 2008), dynamic recommendation (Tantipathananandh *et al.*, 2007), and social network analysis (Lin *et al.*, 2010). The majority of the existing literature on community detection assumes that the identified communities are exclusive in the sense that every member in data collection can be in only one community at the same time, no matter whether the communities are detected through soft or hard clustering.

While in certain situations this assumption is valid, there are many other situations in which members may belong to more than one community. In other words, there are overlaps among the identified communities. For example, if we are interested in detecting the social communities of a college student population based on their email and phone communications, we may find that some people belong to different groups (e.g., a hobby group and a study group). In the recent literature (Fu and Banerjee, 2008; Lancichinetti *et al.*, 2009; Wang *et al.*, 2010; Gregory, 2010), the topic of overlapping community detection has attracted attention in the data mining area.

In the existing literature, researchers use either content or link to discover the overlapping communities. To the best of our knowledge, there has been no attempt in data mining or machine learning that uses both content and link information. In this paper, we develop a subgraph matching technique called subgraph overlapping clustering (SOC) using both types of information to identify the overlapping communi-

‡ Corresponding author

ties in a data collection. SOC is a clustering technique that does not require any training data. We demonstrate the effectiveness and promise of the proposed solution through extensive evaluations using real datasets in comparison with the state-of-the-art methods.

2 Related works

We review three most related areas, i.e., overlapping community detection based on link, content, and both link and content.

2.1 Overlapping community detection based on link

In recent years, many approaches to overlapping community detection or clustering based on link have been proposed. Palla *et al.* (2005) used k-clique to represent the structure of a network and then find the overlapping clusters, since there are always nodes belonging to multiple cliques. Baumes *et al.* (2005) sought the overlapping subgraphs in two steps. They used LA to initialize the seed clusters and IS^2 to iteratively improve these clusters. Zhang *et al.* (2007) combined fuzzy c-means clustering based on a spectral feature selection method and a new modular function to achieve overlapping clustering.

More recently, Gregory (2010) uncovered overlapping clusters using the label propagation approach. Lancichinetti *et al.* (2009) discovered both overlapping and hierarchical communities by taking advantage of local optimization of a fitness function. Lee *et al.* (2010) discovered overlapping clusters based on the expansion of clique-seeds and optimization of a local fitness function. Kovacs *et al.* (2009) introduced ModuLandan which is an integrative network module determination approach. Ahn *et al.* (2010) proposed a method based on the fact that the cluster is defined as a set of edges. Airoldi *et al.* (2008) and Yan *et al.* (2011) used statistical models to discover the communities. Fortunato and Castellano (2009) gave a systematic summary.

2.2 Overlapping community detection based on content

Several approaches to overlapping community detection based on content have been proposed in the literature. Banerjee *et al.* (2005) proposed a model-based overlapping clustering (MOC), which is an extension of Segal *et al.* (2003)'s model and can assign one object to multiple clusters for the general model. Fu and Banerjee (2008) used a probability model named multiplicative mixture models (MMMs) to find overlapping clusters. The obvious difference from traditional models is that the latent variables they used are boolean data, decided by whether one object belongs to a corresponding cluster. Every object is assumed to be generated from a part of the product of the component distributions. Wang *et al.* (2010) sought overlapping communities in two steps. First, they viewed every edge (behavior) as an item and represented it using different measures. Second, they used EdgeCluster (Tang and Liu, 2009) to cluster these edges into k-clusters and then assigned objects to the corresponding multiple communities.

2.3 Community detection based on both link and content

In the field of non-overlapping community detection, the research based on both content and link information has generated a large body of literature. Zhu *et al.* (2007) proposed a matrix factorization method by dealing with the content matrix and the link matrix at the same time. Zhou *et al.* (2009) generated the attribute vertices to form an attribute augmented graph, which is a new graph combining content and link information. Then, they used an efficient random walk technique to iteratively identify communities. Zhang *et al.* (2008) proposed a relaxation labeling based clustering method to measure the heterogeneous links. This method can effectively improve clustering based on both content and link. Yu *et al.* (2009) performed a spectral clustering based on kernel fusion combining both types of information.

Using a probability model based on both content and link information is also a main research direction. Cohn and Hofmann (2000) combined PHITS-PLSA to achieve community detection. Nallapati *et al.* (2008) proposed Pairwise-Link-LDA, which combines LDA and the mixed membership block stochastic model. They also proposed Link-PLSA-LDA, which combines PLSA and LDA into a single model. Yang *et al.* (2009) used a discriminative model based on both content and link analyses to detect communities, which differentiates itself from traditional generative models.

3 Problem definition

We first give the notations to be used in the following. $U = \{u_1, u_2, \ldots, u_N\}$ is a set of objects, $A = \{a_1, a_2, \ldots, a_M\}$ is a set of attributes, and $u_i \to a_j$ indicates that a_j is an attribute of u_i. The attribute matrix E establishes the content information between the objects and the attributes, and in this study the content means binary attributes. The adjacency matrix W establishes the link information among the objects. E is an $N \times M$ attribute matrix in a network with all binary entries $e_{ij} \in \{0, 1\}$, where $e_{ij} = 1$ indicates that the ith object has the jth attribute. W is an $N \times N$ symmetric adjacency matrix in a network with its entries $w_{ij} \geq 0$ indicating the weight of the link between the ith and jth objects.

According to the above definition, we first construct an object-attribute graph from the original dataset. The proposed method is based on the data structure of this object-attribute graph.

Definition 1 (Object-attribute graph) An object-attribute graph is denoted as G=(U, A, W, E), where U is the set of the object nodes on one side, A is the set of the attribute nodes on another side, W is the set of links among the object nodes, and E is the set of edges connecting U and A. If $e = (u_i, a_j) \in E$, $u_i \in U$ and $a_j \in A$.

Fig. 1 is a simple example of an object-attribute graph. It is clear that the edges in the graph are the main source of the observations to the object behavior. For instance, node u_4 with both red and blue attributes belongs to cluster 1 ($u_1 - u_4$) and cluster 2 ($u_4 - u_6$). Traditional relational clustering methods, either hard or soft clustering, fail to identify the fact that u_4 belongs to both cluster 1 and cluster 2, as through clustering the edges it is clear that two types of edges can be found and u_4 belongs to both types.

We then define the candidate subgraph as follows.

Definition 2 (Candidate subgraph) Given a dataset represented as an object-attribute graph, a set of candidate subgraphs is denoted as $S_i = (U_i, A_i, W_i, E_i), 1 \leq i \leq L$, where every candidate subgraph is a subgraph of the object-attribute graph. In addition, $\bigcup_{i=1}^{l} U_i = U$ and $U_i \cap U_j = \varnothing$ for any $i \neq j$. A_i is the set of attributes that U_i has, W_i is the set of links among U_i, and E_i is the set of edges between U_i and A_i. Moreover, we assume that the objects are closely connected in the same candidate subgraph and that there exist a set of dominant attributes in one candidate subgraph.

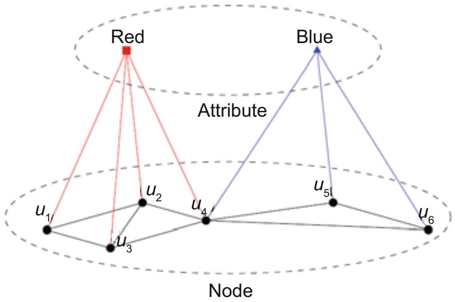


Fig. 1 The object-attribute graph. The upper circle (red and blue) represents the dimensions of the attributes and the lower circle ($u_1 - u_6$) represents the dimensions of the objects

Thus, the problem of overlapping community detection based on both content and link information is described as follows:

Input: an object-attribute graph with both the attribute matrix E (content) and the adjacency matrix W (link).

Intermediate: a stable set of candidate subgraphs.

Output: k overlapping communities.

4 Candidate subgraph evaluation

In this section, we first assume the availability of a candidate subgraph, and then define the methods to measure the relevance between an object node and a set of candidate subgraphs and that between an attribute and a set of candidate subgraphs. These methods will be used to evaluate the relevance as part of the community discovery process. The generation of candidate subgraphs is a part of the iterative community discovery process and will be described later in this paper.

4.1 Subgraph attribute modularity measure

Given a set of candidate subgraphs $\{S_i\}_{i=1}^{L}$ that come from the object-attribute graph, we first define a measure of relevance between attribute a_i and candidate subgraph S_l as Eq. (1), which we call subgraph attribute modularity measure (SAMM):

$$r(a_i|S_l) = \frac{1}{2m_l} \sum_{\substack{u_t,u_g \in U_l \\ u_t,u_g \rightarrow a_i}} \left(w_{tg} - \frac{d_t d_g}{2m_l} \right), \quad (1)$$

where m_l denotes the sum of the weights of the links in W_l, and d_t and d_g denote the degrees of objects u_t and u_g, respectively.

The definition of this metric is motivated from the modularity measure (Newman and Girvan, 2004). We call an attribute 'dominant' if the community that consists of all the objects having this attribute in a candidate subgraph has a high value of modularity. For a clear statement, we can use another description of modularity to denote the above relevance:

$$r(a_i|S_l) = \frac{\mathrm{Cut}(U_l^{a_i}, U_l^{a_i})}{2m_l} - \left(\frac{\mathrm{Cut}(U_l^{a_i}, U_l)}{2m_l} \right)^2, \quad (2)$$

where $U_l^{a_i}$ denotes the set of objects $u \in U_l$, $u \rightarrow a_i$ and $\mathrm{Cut}(U_i, U_j) = \sum_{p \in U_i, q \in U_j} w_{pq}$.

Due to the typical sparsity of the graph in a real world problem, we use a threshold for SAMM and therefore only those dominant attributes are selected for further computation.

In summary, we define the computation of SAMM as follows:

$$p(a_i|S_l) = \begin{cases} \dfrac{\exp\left[\lambda_a (r(a_i|S_l) - t_l)\right]}{H_i}, & r(a_i|S_l) \geq t_l, \\ \dfrac{p_a}{H_i}, & \text{otherwise}, \end{cases} \quad (3)$$

where $t_l \triangleq c/M \times \sum_{i=1}^{M} r(a_i|S_l)$.

In this definition, t_l is the threshold for the dominant attributes, obtained as the average value of all the relevances between an attribute and the lth candidate subgraph, and can be controlled by c. H_i is a normalization constant, λ_a is a controllable parameter, and p_a is a positive constant. Specifically, if the value of the relevance of attribute a_i is larger than t_l, the probability of a_i belonging to S_l is exponential with this λ_a; otherwise, this probability is equal to a small value p_a.

4.2 Variation of the Markov random field

We use a variation of the Markov random field to measure the relevance between an object node and a candidate subgraph. The probability that an object node belongs to a candidate subgraph is defined as follows:

$$p(u_i|S_l) = \begin{cases} \dfrac{\log\left(\lambda_n \sum_{\substack{i,j \in S_l \\ j \in N_{(i)}}} w_{ij} \right)}{H_i}, & N_{(i)} \neq \varnothing, \\ \dfrac{p_n}{H_i}, & \text{otherwise}, \end{cases} \quad (4)$$

where H_i is a normalization constant, $N_{(i)}$ is the set of nodes that are neighbors of u_i, λ_n is a parameter, and p_u is a tiny positive constant. In this definition, we change the traditional model of exponent (Gao *et al.*, 2010) to the model of logarithm to keep the object degrees roughly balanced. If an object has no neighbor, the probability is set to a small positive value p_u.

After establishing the measures between an attribute and a candidate subgraph and those between an object and a candidate subgraph, we further define the relevance between an edge in the object-attribute graph and an attribute as follows:

$$p(u_i \rightarrow a_j|S_l) = p(e_{ij}|S_l) \propto p(u_i|S_l) p(a_j|S_l). \quad (5)$$

5 Overlapping community detection

In this section, we describe first how to obtain a series of stable candidate subgraphs and then the method for identifying overlapping communities.

5.1 Stable candidate subgraphs

We assume that both the objects and the attributes are the observed variables. The candidate subgraphs are the latent variables. We represent these latent variables as $s = \{s_l\}_{l=1}^{L}$. An edge can be in multiple object-attribute graphs. This fact is represented as a mixture distribution in the form

$$p(u_i \rightarrow a_j) = p(e_{ij}) = \sum_{l=1}^{L} \pi_l p(e_{ij}|s_l = 1), \quad (6)$$

where π_l is the marginal distribution over s, $p(s_l = 1) = \pi_l$, $\sum_{l=1}^{L} \pi_l = 1$.

Motivated by Sun *et al.* (2009a), we use an expectation-maximization (EM) algorithm (Bishop, 2006) to maximize the likelihood function $p(E|\pi)$ with respect to π. We first give the likelihood for the complete data set $\{E, S\}$, which takes the form

$$\begin{aligned} p(E,S|\pi) &= \prod_{i=1}^{N}\prod_{j=1}^{M}\prod_{l=1}^{L} p(e_{ij}, s_l)^{w_{ij}} \\ &= \prod_{i=1}^{N}\prod_{j=1}^{M}\prod_{l=1}^{L} \left(\pi_l^{s_{ijl}} p(e_{ij}|s_l=1)^{s_{ijl}}\right)^{w_{ij}}, \end{aligned} \tag{7}$$

where s_{ijl} denotes the lth component of s_{ij}. We give different weights for different edges in this equation. The equation is then mapped into a form of logarithm:

$$\begin{aligned} \ln p(E,S|\pi) =& \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{L} w_{ij} s_{ijl} \\ & \cdot [\ln \pi_l + \ln p(e_{ij}|s_l=1)]. \end{aligned} \tag{8}$$

In the E-step, we evaluate the $p(S|E, \pi^{\text{old}})$:

The posterior distribution of the latent variables can be denoted as follows:

$$p(S|E,\pi) \propto \prod_{i=1}^{N}\prod_{j=1}^{M}\prod_{l=1}^{L} [\pi_l p(e_{ij}|s_l)]^{s_{ijl}}. \tag{9}$$

Next, the expected value of s_{ijl} is computed as

$$\begin{aligned} E[s_{ijl}] &= \frac{\sum_{s_{ijl}} s_{ijl} [\pi_l p(e_{ij}|s_l=1)]^{s_{ijl}}}{\sum_{s_{ijl}} [\pi_l p(e_{ij}|s_l=1)]^{s_{ijl}}} \\ &= \frac{\pi_l p(e_{ij}|s_l=1)}{\sum_{l=1}^{L} \pi_l p(e_{ij}|s_l=1)} = \gamma(s_{ijl}), \end{aligned} \tag{10}$$

where $p(e_{ij}|s_l=1)$ is computed in Eq. (5) and $\gamma(s_{ijl})$ is the responsibility of component l for e_{ij}.

In the M-step, we evaluate π^{new} given by

$$\pi^{\text{new}} = \arg\max_{\pi} \mathcal{Q}(\pi, \pi^{\text{old}}), \tag{11}$$

where

$$\begin{aligned} & \mathcal{Q}(\pi, \pi^{\text{old}}) \\ &= \sum_{s} p(S|E, \pi^{\text{old}}) \ln p(E,S|\pi) \\ &= E_s[\ln p(E,S|\pi)] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{L} w_{ij}\gamma(s_{ijl}) \left[\ln \pi_l + \ln p(e_{ij}|s_l=1)\right]. \end{aligned} \tag{12}$$

To find the best π, we use a Lagrange multiplier:

$$\begin{aligned} & \frac{\partial}{\partial \pi_l}\left[\mathcal{Q}(\pi, \pi^{\text{old}}) - \lambda\left(\sum_{l=1}^{L}\pi_l - 1\right)\right] = 0 \\ & \Rightarrow \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{w_{ij}\gamma(s_{ijl})}{\pi_l} = \lambda \\ & \Rightarrow \pi_l = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}\gamma(s_{ijl})}{\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}}, \end{aligned} \tag{13}$$

where π_l is the parameter we expect to update. The difference from a typical mixture model is that here we do not directly use $p(S|E)$ to reassign the candidate subgraphs; instead, we use $p(S|U)$ to reclassify U into the L candidate subgraphs.

$$p(s_l = 1|u_i) = \frac{\pi_l p(u_i|s_l=1)}{\sum_{t=1}^{L} \pi_t p(u_i|s_t=1)}, \tag{14}$$

where we use vector $\{p(s_l=1|u_i)\}_{l=1}^{L}$ to denote the object u_i. Next, we use this information to cluster all the objects and obtain all the candidate subgraphs. After all the candidate subgraphs are reassigned, the EM algorithm is performed repeatedly. Finally, we can identify the L stable candidate subgraphs.

5.2 Clustering the edges

After the edges are represented by the object-attribute graphs, we cluster these edges into K clusters and obtain the K overlapping communities. The approach for clustering is the simple k-means method. Specifically, we use cosine similarity to measure the distance between an edge and the corresponding center of a cluster.

The algorithm of overlapping community detection combining content and link is listed in the following. We name this approach 'subgraph overlapping clustering' (SOC). The algorithm can be downloaded from http://www.isee.zju.edu.cn/dsec/publication_ch.html.

Initialization is a critical step for the overlapping community detection algorithm. Instead of randomly building candidate subgraphs at the beginning, we use a global method, i.e., dimensionality reduction, to make a simple assignment for the objects. Some ordinary attribute will be dominant and some noise will appear at first, but the following adjustment of subgraphs will weaken these adverse effects. The value of π_l can be denoted as $|U_l|/N$, where $|U_l|$ is the number of objects in U_l.

Algorithm 1 Subgraph overlapping clustering

```
Input: Object-attribute graph $G = (U, A, W, E)$,
       candidate subgraph number $L$
Output: $K$ overlapping communities
Initialization: build $L$ subgraphs, $\{S_l\}_{l=1}^{L}$ and $\{\pi_l\}_{l=1}^{L}$
loop
  do forward loop, yielding stable subgraphs
  for $S_l \in S$ do
    For each $a_j \in A$, compute $p(a_j|s_l = 1)$
    For each $u_i \in U_l$, compute $p(u_i|s_l = 1)$
  end for
  //E-step:
  for $e_{ij} \in E$ do
    Compute $\{\gamma(s_{ijl})\}_{l=1}^{L}$
  end for
  //M-step:
  For each $l$, compute $\pi_l$
  for $u_i \in U$ do
    Compute $\{p(s_l = 1|u_i)\}_{l=1}^{L}$
  end for
  //Update: Reassign the subgraphs
end loop
Cluster the edges based subgraphs
```

6 Experiments

In this section we report the experiments on one synthetic dataset and two real datasets. The two real datasets are collected from DBLP and Newmovie (Tang *et al.*, 2009), respectively. We compare the proposed method with two state-of-the-art approaches, correlational learning (Wang *et al.*, 2010) and COPRA (Gregory, 2010). The correlational learning approach detects overlapping communities only based on content information and COPRA only based on link information.

Setting of the parameters: λ_a and λ_n are the main parameters that should be set in this algorithm. We set $\lambda_a = 4$, $\lambda_u = 0.2$ for DBLP and synthetic data and $\lambda_a = 4$, $\lambda_u = 0.3$ for Newmovies, to balance the weights of content and link information. The other parameters are set as $c = 0.3$, $p_a = 0.1$, and $p_n = 0.2$.

6.1 Synthetic data

To generate the link information, we assume that the distribution of the object degrees follows the power-law, which is consistent with the characteristics of the typical sparse and local density distribution in a real-world network. It is observed that objects in an overlapping community in a network typically have more neighbors than other objects in the network. Thus, in the synthetic dataset these objects have a higher probability to own a higher degree. To generate the content information, we generate different dominant attributes in different communities with overlapping objects having multiple attributes.

The structure of the synthetic dataset is described in Fig. 2. The five overlapping communities $\{C_i\}_{i=1}^{5}$ are the main body of the synthetic dataset. $\{a_i\}_{i=1}^{10}$, which are the dominant attributes, exist in the five communities. O_1 is a set of overlapping objects existing in C_1, C_2, and C_5 and owning the attributes of these three communities. O_2 is a set of overlapping objects existing in C_2 and C_3 and owning the attributes of these two communities; similarly for O_3.

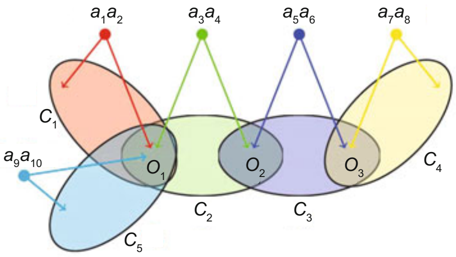


Fig. 2 A simple structure of the synthetic data

Specifically, we generate a synthetic dataset that has 3600 objects for five communities, where two communities have 1000 objects and the other three communities have 800 objects. O_1, O_2, and O_3 all have 200 objects. Next, we generate the degree of every object based on the power-law and set the maximum degree of overlapping objects as twice that of the common objects. A link is supposed to exist only between two objects in the same community. We assign 70% of the objects with one attribute and other objects with two attributes in one community.

6.1.1 Evaluation using an extension of NMI

To evaluate the clustering accuracy, we use the normalized mutual information (NMI) as the performance metric, which is a powerful metric for testing the similarity between the clustering results. Since NMI applies only to the classic clustering scenario where communities detected are exclusive to each other, to accommodate overlapping community detection, we adopt an extension of the classic NMI proposed by Lancichinetti *et al.* (2009):

$$\begin{cases} H(X|Y)_{\text{norm}} = \\ \quad \dfrac{1}{|C_X|}\displaystyle\sum_k \dfrac{\min_l \in \{1,2,\cdots,|C_Y|\}H(X_k|Y_l)}{H(X_k)}, \\ H(Y|X)_{\text{norm}} = \\ \quad \dfrac{1}{|C_Y|}\displaystyle\sum_k \dfrac{\min_l \in \{1,2,\cdots,|C_X|\}H(Y_k|X_l)}{H(Y_k)}, \\ \text{NMI} = 1 - \dfrac{1}{2}\left[H(X|Y)_{\text{norm}} + H(Y|X)_{\text{norm}}\right], \end{cases} \tag{15}$$

where $H(X|Y)$ and $H(Y|X)$ are the conditional entropies, and $|C_X|$ and $|C_Y|$ are the numbers of the related clusters. The definition states as follows. We first fix a cluster in a clustering and seek its most similar cluster in another clustering based on the mutual information. Next, we compute the extension of the NMI based on the average of the information.

In the experiments, we first cluster the synthetic data into five overlapping clusters based on the three approaches mentioned above. Next, we evaluate all the three methods based on different levels of noise; each noise level is generated with 50 iterations. The noise is generated by randomly adding edges in E and/or links in W. The ratio of the number of noisy edges to the number of noisy links is controlled in the range of $0-0.5$. Finally, we evaluate the performance of the three methods based on different levels of missing data. We randomly drop $0-30\%$ of the synthetic data and report the results.

Fig. 3 shows the performance of the three overlapping community detection methods based on different levels of noise and missing data. The accuracy of SOC is higher than that of correlational learning by $10\%-12\%$ in case of noisy data and by up to 8% in case of missing data. SOC performs much better than COPRA in these two situations. The results also indicate that SOC has stable performance at the presence of noise or missing data.

6.1.2 Evaluation using an extension of modularity

We also evaluate the efficiency of the three methods using an extension of the modularity metric (EQ). The modularity metric (Newman and Girvan, 2004) measures the goodness of a clustering method in comparison with a random graph. Again, the modularity metric is only for classic clustering methods where the communities are assumed to be exclusive. To accommodate the overlapping communities we use an extension of the existing modularity metric proposed in Shen *et al.* (2009):

$$\text{EQ} = \frac{1}{2m}\sum_{i=1}^{K}\sum_{v,w\in C_i}\frac{1}{O_v O_w}\left(A_{vw} - \frac{k_v k_w}{2m}\right), \tag{16}$$

where O_v is the number of the clusters to which object v belongs and k_v is the degree of object v. The setting of the different noise levels and missing data levels is the same as in the previous section.

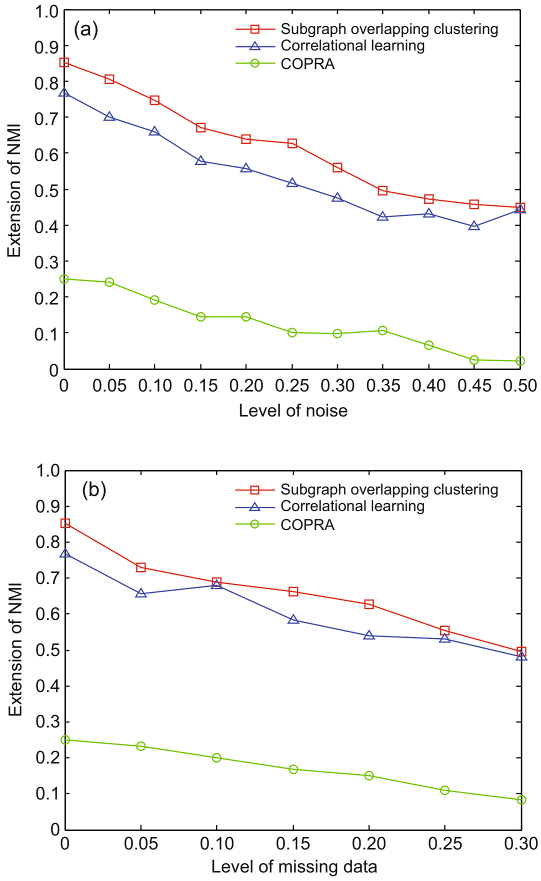
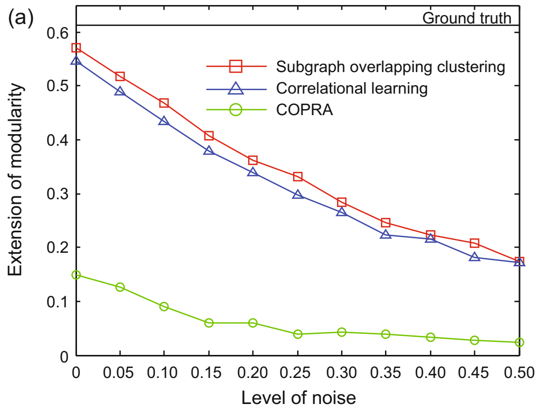


Fig. 3 Extension of normalized mutual information (NMI) for detecting the three approaches with different levels of noise (a) and different levels of missing data (b)

In Fig. 4, the black line is the ground truth, and can be considered as the baseline. The EQ of SOC is higher than that of correlational learning by up to 7% with respect to noisy data and by up to 5% with respect to missing data. COPRA fails to seek any overlapping cluster, as the synthetic dataset is not a real social network, even though it follows the power-law.



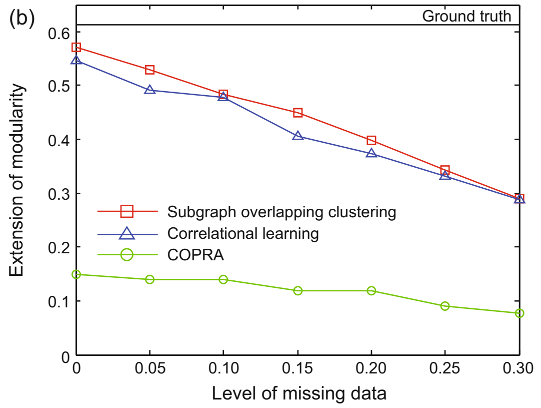


Fig. 4 Extension of modularity for detecting the three approaches with different levels of noise (a) and different levels of missing data (b)

6.2 DBLP

We use the DBLP database from Sun *et al.* (2009b) and extract two object-attribute graphs, i.e., an author-conference graph and a paper-word graph.

6.2.1 Author-conference graph

To build the author-conference graph, we assume that if an author has at least one paper in a certain conference, there exists an edge from this author to the conference and that if two authors have jointly published one paper, there is a link between them. In this experiment, we set the weight as a binary value. The author-conference graph is composed of 28 702 authors and 20 conferences. These 20 conferences mainly come from four fields:

Database: EDBT, ICDE, PODS, SIGMOD, and VLDB.

Data mining: ICDM, KDD, PAKDD, PKDD, and SDM.

Machine learning: AAAI, ECML, ICML, and IJCAI.

Information retrieval: ECIR, SIGIR, WWW, and WSDM.

Others: CVPR and CIKM (As we know, CVPR and CIKM belong to multiple fields).

Figs. 5a and 5b give a detailed description of this graph. The distribution of authors based on their behavior or degrees almost follows the power-law.

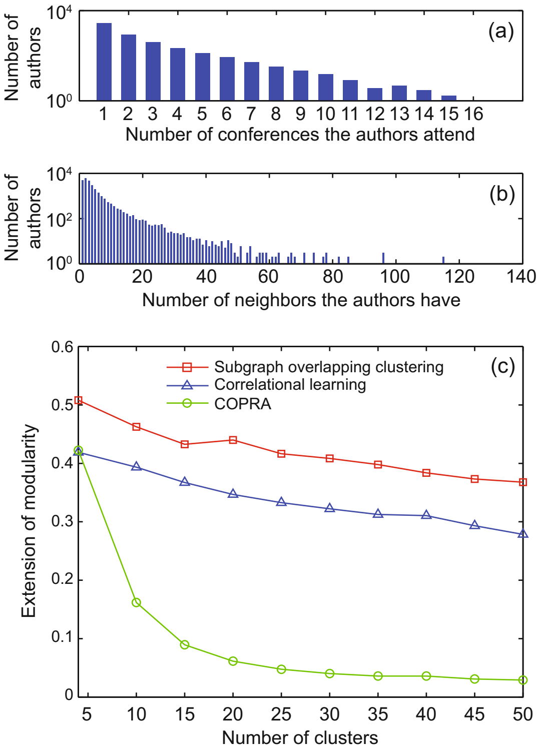


Fig. 5 Author-conference graph. (a) Distribution of authors based on the number of conferences they attend; (b) Distribution of authors based on the number of neighbors they own; (c) Extension of modularity for the different numbers of clusters in the author-conference graph

We evaluate the efficiency of the three approaches using the extension of modularity (EQ) with different numbers of clusters. Each evaluation is run 50 times.

Fig. 5c shows that SOC performs much better than the other approaches, indicating that the approach combining content and link information reflects much more accurately the structure of the real network than methods using either content or link information. Specifically, we have the following two observations:

1. SOC is more sensitive to overlapping communities. For example, we cluster the author-conference graph into four overlapping communities and find that the correlational learning approach assigns several authors who work mainly in two fields to all the communities with the total number of overlapping objects much higher than that of SOC. COPRA sometimes assigns these authors to only one community, which is an obvious mistake.

2. SOC uses the concept of candidate subgraphs based on the two types of information to capture the 'local closeness' of a real social network. In contrast, correlational learning considers only the global information while COPRA considers only the neighborhood information of the objects in a network.

Fig. 6 is another evaluation for measuring the similarity among the conferences. The four labels around the circle represent the four different fields; the central labels denote CVPR and CIKM, which belong to multiple fields. It is obvious that the similarity among the conferences within the same field is much higher than that among the conferences across different fields. In this experiment, we define the similarity as follows:

$$\mathbf{Conf}_{\cdot i} = \left\{ \frac{|\mathrm{Conf}_{\cdot i} \cap C_k|}{|\mathrm{Conf}_{\cdot i}|} \right\}_{k=1}^{K},$$

$$\mathrm{sim} < \mathbf{Conf}_{\cdot i}, \mathbf{Conf}_{\cdot j} >= \frac{\mathbf{Conf}_{\cdot i}}{\|\mathbf{Conf}_{\cdot i}\|} \cdot \frac{\mathbf{Conf}_{\cdot j}}{\|\mathbf{Conf}_{\cdot j}\|},$$

where $|\mathrm{Conf}_{\cdot i}|$ denotes the number of the authors publishing papers in the ith conference, $|\mathrm{Conf}_{\cdot i} \cap C_k|$ denotes the number of the authors assigned to the kth community, and $\mathbf{Conf}_{\cdot i}$ is a K-dimensional vector representing the ith conference. In addition, we design a threshold which distinguishes between a strong similarity and a weak similarity, and set its value as an average in between. The number of the clusters K does not influence the final result. We set it as 4 here.

Fig. 6 shows that SOC and correlational learning can both cluster the conferences and assign them to proper fields. They assign CVPR to both 'data ming' and 'machine leaning' fields. We also find that SOC assigns CIKM to 'data mining', 'machine learning', and 'information retrieval' fields at the same time but correlational learning assigns CIKM only to the 'information retrieval' field. This fact means that the dominant attributes (i.e., the conferences) in the corresponding clusters are not strong. SOC assigns the conferences to appropriate clusters based on the dominate attributes. Since COPRA does not have the content information, we cannot include it in this experiment.

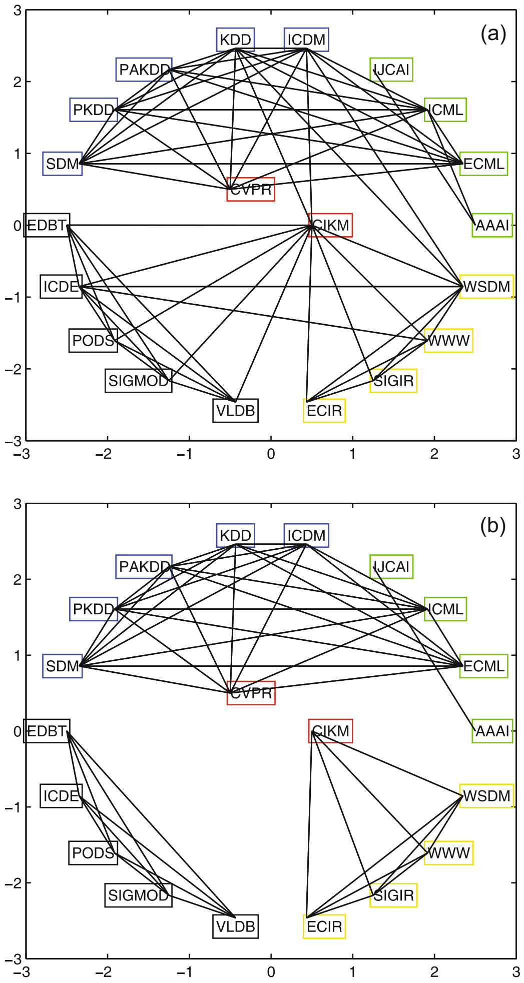


Fig. 6 Similarity among the conferences in DBLP. (a) Performance of SOC; (b) Performance of correlational learning

6.2.2 Paper-word graph

In the process of building the paper-word graph, we set that if two papers are written by the same author, there exists a link between them. We assume that an author works in one specific field in general, which means that if two papers are in different fields and are written by the same author, these two fields may have a relationship. There are about 28 500 papers and 9500 effective words in this graph. To simply extract the semantics, we use a simple technique of dimensionality reduction to transform an ordinary set of words into 50 bags of words. We then use the topic words to replace the words to describe the papers. In this experiment, we do not use these bags of words for correlational learning because it can directly deal with high-dimensional data by singular value decomposition (SVD).

We give the statistics about the paper-word graph in Figs. 7a and 7b. The majority of the papers has 2–6 topic words and the distribution based on the neighbors follows the power-law. Every paper has about 26 neighbors.

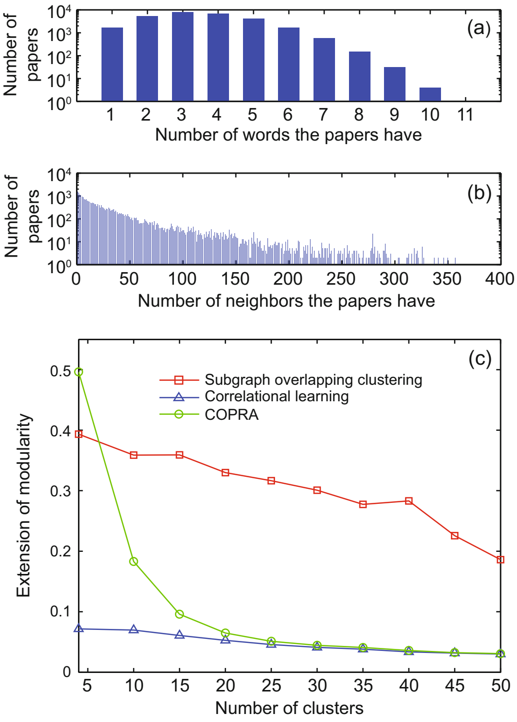


Fig. 7 Paper-word graph. (a) Distribution of papers based on the number of words they have; (b) Distribution of papers based on the number of neighbors they own; (c) Extension of modularity for the different numbers of clusters in the paper-word graph

We evaluate the efficiency of the three approaches using the extension of modularity (EQ) in the paper-word graph with different numbers of clusters, and each evaluation is run 50 times. Fig. 7c shows that when the number of clusters is larger than 10, the performance of SOC is much better than those of the other two approaches. COPRA has the best result when the clustering number is four, because in this situation the papers come mainly from four fields and this setting is useful for propagating the proper labels to the corresponding objects. Correlation learning fails to find the overlapping clusters. There are a huge number of attributes in this experiment and some of them may not be useful although all the words are filtered at first. Therefore, using only the global information which is the mapping based on SVD fails to build the relationship among the objects.

6.3 Newmovies

Newmovies is a heterogeneous network consisting of movies, actors, directors, writers, and various relationships among them. It is divided mainly into two parts. The first part is content information. Every actor, actress, or film (object) has a short presentation from Wikipedia pages and the total number of objects is about 34 200. The second part is link information. If two actors, actresses, or films exist in the same Wikipedia page, there is a link between them. To analyze this huge social network, we build an object-word graph based on both content and link information. In this experiment we also use the dimensionality reduction technique to transform the ordinary set of words into 100 bags of words.

We first give the statistics of the object-word graph in Figs. 8a and 8b. It is shown that the majority of the objects has 1–4 topic words. The distribution based on the neighborhood follows the power-law and the average degree of the objects is about 7.

We use extension of modularity to evaluate the performance of these three approaches with different numbers of clusters, and each evaluation is run 50 times. Fig. 8c shows that the result of overlapping clusters found by SOC is much better than by the other two approaches.

6.4 Algorithm complexity

To perform the evaluation of algorithm complexity, we compute the running time of 50 iterations for the three overlapping clustering algorithms in the author-conference graph. Fig. 9 shows that the running time of correlational learning is the least and that of subgraph overlapping clustering is the longest. The reason is that our approach uses both content and link data and a complex statistical model.

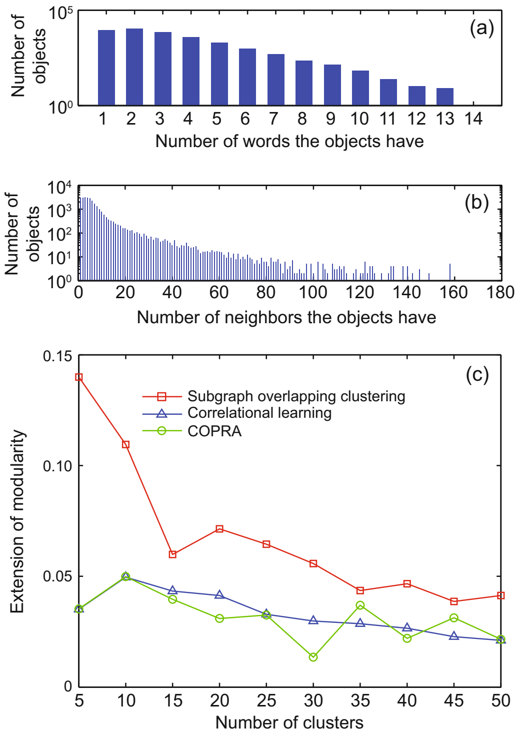


Fig. 8 Object-word graph. (a) Distribution of objects based on the number of words they have; (b) Distribution of objects based on the number of neighbors they own; (c) Extension of modularity for the different numbers of clusters in the object-word graph

7 Conclusions

We deal with the issue of combining content and link information for discovering overlapping communities and develop an effective approach, subgraph overlapping clustering (SOC), for community detection. SOC uses a candidate subgraph strategy that combines modularity theory and the Markov random field to develop an appropriate assignment for edges (social behavior or action). The candidate subgraph trained by an EM algorithm is a stable iterative process. We also demonstrate that SOC reflects the 'local denseness' for real social networks. Experiments on large databases of DBLP and Newmovies demonstrate that SOC successfully generates high quality overlapping communities in comparison with the peer methods in the literature that use either content or link information.

There are two directions for future research. First, we would seek a more powerful concept about the subgraph to make a more accurate description of social behavior. Second, we would seek a suitable concept to represent the overlapping communities in complex and huge social networks.

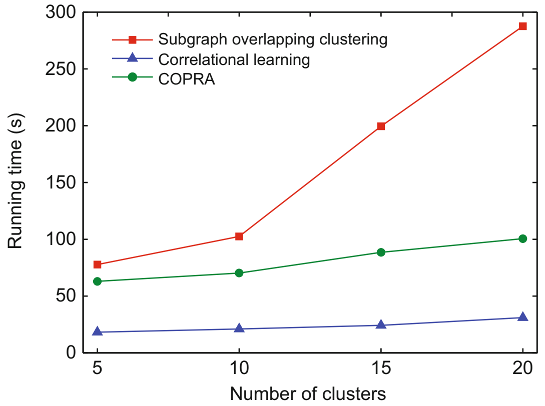


Fig. 9 Running times of 50 iterations for the different numbers of clusters in the author-conference graph

References

Ahn, Y.Y., Bagrow, J.P., Lehmann, S., 2010. Link communities reveal multiscale complexity in networks. *Nature*, **466**(7307):761-764. [doi:10.1038/nature09182]

Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P., 2008. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, **9**:1981-2014.

Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., Mooney, R.J., 2005. Model-Based Overlapping Clustering. KDD, p.532-537.

Baumes, J., Goldberg, M.K., Ismail, M.M., 2005. Efficient Identification of Overlapping Communities. ISI, p.27-36.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Chen, W.Y., Zhang, D., Chang, E.Y., 2008. Combinational Collaborative Filtering for Personalized Community Recommendation. KDD, p.115-123.

Cohn, D.A., Hofmann, T., 2000. The Missing Link—a Probabilistic Model of Document Content and Hypertext Connectivity. NIPS, p.430-436.

Fortunato, S., Castellano, C., 2009. Community Structure in Graphs. *In*: Encyclopedia of Complexity and Systems Science, Part 3, p.1141-1163. [doi:10.1007/978-0-387-30440-3_76]

Fu, Q., Banerjee, A., 2008. Multiplicative Mixture Models for Overlapping Clustering. ICDM, p.791-796.

Gao, J., Liang, F., Fan, W., Wang, C., Sun, Y.Z., Han, J.W., 2010. On Community Outliers and Their Efficient Detection in Information Networks. KDD, p.813-822. [doi:10.1145/1835804.1835907]

Gregory, S., 2010. Finding overlapping communities in networks by label propagation. *New J. Phys.*, **12**(10): 103018. [doi:10.1088/1367-2630/12/10/103018]

Kovacs, I.A., Palotai, R., Szalay, M.S., Csermely, P., 2009. Community Landscapes: an Integrative Approach to Determine Overlapping Network Module Hierarchy, Identify Key Nodes and Predict Network Dynamics. CoRR, abs/0912.0161.

Lancichinetti, A., Fortunato, S., Kertesz, J., 2009. Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.*, **11**(3):033015. [doi:10.1088/1367-2630/11/3/033015]

Lee, C., Reid, F., McDaid, A., Hurley, N., 2010. Detecting Highly Overlapping Community Structure by Greedy Clique Expansion. SNA-KDD.

Lin, C.X., Zhao, B., Mei, Q.Z., Han, J.W., 2010. PET: a Statistical Model for Popular Events Tracking in Social Communities. KDD, p.929-938. [doi:10.1145/1835804.1835922]

Nallapati, R., Ahmed, A., Xing, E.P., Cohen, W.W., 2008. Joint Latent Topic Models for Text and Citations. KDD, p.542-550.

Newman, M., Girvan, M., 2004. Finding and evaluating community structure in networks. *Phys. Rev. E*, **69**(2):026113. [doi:10.1103/PhysRevE.69.026113]

Palla, G., Derenyi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**(7043):814-818. [doi:10.1038/nature03607]

Segal, E., Battle, A., Koller, D., 2003. Decomposing Gene Expression into Cellular Processes. Pacific Symp. on Biocomputing, p.89-100.

Shen, H.W., Cheng, X.Q., Cai, K., Hu, M.B., 2009. Detect overlapping and hierarchical community structure in networks. *Phys. A*, **388**(8):1706-1712. [doi:10.1016/j.physa.2008.12.021]

Sun, Y.Z., Han, J.W., Zhao, P.X., Yin, Z.J., Cheng, H., Wu, T.Y., 2009a. RANKCLUS: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis. EDBT, p.565-576. [doi:10.1145/1516360.1516426]

Sun, Y.Z., Yu, Y.T., Han, J.W., 2009b. Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema. KDD, p.797-806. [doi:10.1145/1557019.1557107]

Tang, J., Sun, J.M., Wang, C., Yang, Z., 2009. Social Influence Analysis in Large-Scale Networks. KDD, p.807-816. [doi:10.1145/1557019.1557108]

Tang, L., Liu, H., 2009. Scalable Learning of Collective Behavior Based on Sparse Social Dimensions. CIKM, p.1107-1116. [doi:10.1145/1645953.1646094]

Tantipathananandh, C., Berger-Wolf, T.Y., Kempe, D., 2007. A Framework for Community Identification in Dynamic Social Networks. KDD, p.717-726.

Wang, X., Tang, L., Gao, H., Liu, H., 2010. Discovering Overlapping Groups in Social Media. ICDM, p.569-578.

Yan, F., Xu, Z.L., Qi, Y., 2011. Sparse Matrix-Variate Gaussian Process Blockmodels for Network Modeling. UAI, p.745-752.

Yang, T.B., Jin, R., Chi, Y., Zhu, S.H., 2009. Combining Link and Content for Community Detection: a Discriminative Approach. KDD, p.927-936. [doi:10.1145/1557019.1557120]

Yu, S., de Moor, B., Moreau, Y., 2009. Clustering by Heterogeneous Data Fusion: Framework and Applications. NIPS Workshop.

Zhang, S.H., Wang, R.S., Zhang, X.S., 2007. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Phys. A*, **374**(1):483-490. [doi:10.1016/j.physa.2006.07.023]

Zhang, X.D., Hu, X.H., Zhou, X.H., 2008. A Comparative Evaluation of Different Link Types on Enhancing Document Clustering. SIGIR, p.555-562.

Zhou, Y., Cheng, H., Yu, J.X., 2009. Graph clustering based on structural/attribute similarities. *Proc. VLDB*, **2**(1):718-729.

Zhu, S.H., Yu, K., Chi, Y., Gong, Y.H., 2007. Combining Content and Link for Classification Using Matrix Factorization. SIGIR, p.487-494.