# ターム共起に注目したグラフ構造に基づくドキュメントクラスタリング

# Graph Based Document Clustering with Term Co-Occurrence

藤田　真可†　　新美　礼彦††　　小西　修†††

†公立はこだて未来大学大学院　システム情報科学研究科　〒041-8655 北海道函館市亀田中野町 116-2
††公立はこだて未来大学　システム情報科学部　〒041-8655 北海道函館市亀田中野町 116-2
†††公立はこだて未来大学　システム情報科学部　〒041-8655 北海道函館市亀田中野町 116-2

E-mail:　†g2109040@fun.ac.jp　††niimi@fun.ac.jp,　†††okonishi@fun.ac.jp

**あらまし**　ドキュメントクラスタリングは，テキストマイニングにおける最も活発な研究課題のひとつである．ドキュメントクラスタリングは，タームの出現頻度の統計を使って類似なドキュメントに分類するものである．このドキュメントクラスタリングという問題では，二つの異なる分野からの新しいアプローチがある．一つは，複雑ネットワークのコミュニティ抽出，もう一つは，スペクトラルクラスタリングであり，これら二つは，ドキュメント集合を一つのグラフとして表すものである．本研究では，大規模ドキュメント集合を実時間でクラスタリングする方法を提案する．ドキュメントの共起ターム対に注目したグラフを構築し，ハブに基づくクラスタリングを行いサブグラフに分割する. さらに, このサブグラフを基にスペクトラルクラスタリングを適用し概念マップを抽出する．この概念マップをドキュメントのインデックスとして使用した検索システムを構築することができる．これは，検索システムの得られた大きな検索結果集合をダイナミックスクラスタリングできるアルゴリズムである．

**キーワード**　ドキュメントクラスタリング，ハブに基づくクラスタリング，スペクトラルクラスタリング，ターム共起

Abstract：　Document clustering is one of the most active research topics in text mining. Document clustering groups similar documents using statistical computations on term frequencies. Ideally, related documents within the document collection are clustered. In this work two approaches issued from very different fields are explored for document clustering: community detection in complex networks and spectral clustering. Both approaches are based on a representation of the document collection as a graph, of which the nodes represent the documents and the edges represent the similarities between each pair of documents, such that the two approaches have many issues in common. These graph based approaches are complementary and are useful for finding structure in large collections of documents.　We present a novel method for semantically clustering a large collection of documents using community detection in graphs. A term network based on term co-occurrence is generated from the documents collection, the terms in the complex network are clustered into some communities by means of hub based clustering and spectral clustering, the semantic term clusters as conceptual maps are used to generate overlapping document clusters. The terms resulting from clusters as queries are used to map the highest ranked documents to clusters. Our algorithm occupies a middle ground between speed and quality. Our method provides a way to segment large document collection in fast running times. The algorithm presented can also be incorporated into a search system that enables the dynamic clustering of large numbers of search results.

**Keywords**:　document clustering, graph based clustering, community detection, term co-occurrence

## 1. はじめに

現在，テキストデータなど大規模な情報をコンピュータで扱うことが多くなっている．しかし，大量のデータ情報の各情報がどのような関連性があるかということは分かりにくくなっている．その大量の情報の中から必要な情報を取り出せることが必要である．

情報検索において，キーワード検索は，キーワード

と関連しているドキュメントでもユーザーが意図しないドキュメントが検索結果として出てくることがある．これはインターネットコンテンツの発展や普及により情報が多様化しているためである．

また，従来の研究では，大規模なドキュメントクラスタリングに k-means 法が使われており，ドキュメントキーワードを記述するときにはベクトル空間モデルで記述されていた．そのため，対応するデータ量が多くなると結果が煩雑になってしまっていた．

これを解決するためにドキュメントをグラフ表現し，グラフマイニングを行う．

## 2. 関連研究

Barabasi らの研究でスケールフリーネットワークの度数分布が平均に一致しないことが発見された．そのネットワークは 80 パーセント以上のノードがリンク数 4 未満であり，度数分布の上位個数のノード(全体の 0.001 パーセントほどのノード)が 1000 本以上のリンクを持っているべき乗則であることがわかっている．すｔまり，ランダムネットワークの分布では平均から外れるとノード数が少なくなるが，べき乗足に従うスケールフリーネットワークではそのような系に従う尺度が存在していないという特徴がある．本研究では，使用するデータにより構築したネットワークがこのべき乗則に従うかを検証する[1,2].

Rohinski らの研究では複雑なネットワークに対し，そのネットワークを複数の改装・種類を用いて自動的に分類するものを設定し，これを用いて一定のノードを持つ部分グラフを要約することで概念マップのクラスタリングを行っている．本研究ではスケールフリーネットワーク性を用いてハブ構造ネットワーク内のノードから構成される部分グラフを用いてクラスタリングを行う．[3,7]

Illhoi Yoo らの研究ではそれぞれのドキュメントクラスタが重要度の高いネットワーク構造と定義することで，それぞれのドキュメントクラスタについて意味的関連性のある情報の核を見つけ部分グラフを分類するモデルを生成している．この部分グラフのモデルをもとにし，各実験のドキュメントデータを関連付けてネットワークにすることでクラスタリングをおこなっている．本研究で抽出するハブノードとこのハブノードと接続するサブノードから構成されるネットワークもツリー構造のネットワークである．複数のハブ構造ネットワークも同様に用いているが，本研究ではハブ構造ネットワーククラスタリングではなくそのハブ構造ネットワークで構成されるノードで構築されるネットワークでクラスタリングを行う[4,5].

ドキュメント集合内のタームをノードとしたネットワーク構築し，自然言語であるドキュメント内のタームそれぞれがスモールワールドネットワーク構造を示すことから，スモールワールドコミュニティを使った意味的にクラスタリングする方法がある．ドキュメントを語彙のネットワークグラフにし，相互情報量によってグラフカットしクラスタリングする方法である．[6]

## 3. 提案手法

従来までは専門的なドキュメントの分類は専門家が手作業で分類していた．これを自動的に分類できるようシステムを構築する．また，従来のベクトル表現のドキュメントクラスタリングより膨大なドキュメントを直感的に理解できるような表現と，実時間で分類する方法を提案する．ドキュメント集合をグラフ表現し，このネットワークグラフを分割することでインデックスとなる概念マップを抽出し，ドキュメントクラスタリングを行う．

本研究では，大規模ドキュメント集合を実時間で分割する方法を提案する．ドキュメント集合をグラフ表現し，ドキュメント集合からなる複雑なネットワークの分類を効果的に行うアプローチをとる．これによってより効果的な表現となる概念マップに基づくドキュメントのクラスタリングを行う．

まず、ネットワークからハブ構造ネットワークを構築するアプローチを示す．このアプローチでは重要なハブノードとそれにつながるノードからなるネットワークを構築することで，そのドキュメント集合内でのキーワードから多くのキーワード，もしくは多くのドキュメントと関連しているキーワードを見つけることを目的とする．

### 3.1 提案手法の流れ

図 1 ではドキュメント集合から作られたネットワークグラフから概念マップ抽出までの流れを示している．はドキュメント集合全体のタームと共起タームからなるネットワークグラフである．b は共起タームの出現回数に閾値で制限したものである．c はｂのグラフ構造のハブを取り出したサブグラフ(クラスタ)である．d は，Hub Based Clustering を用いて抽出したクラスタにさらに Spectral Clustering [9,10]を用いてクラスタを抽出する．そして，cohesion を使ってそのタームの出現頻度に対してそのタームに接続するエッジとなる共起タームの出現頻度の割合の高いものを抽出して概念マップとする．これらによってドキュメントクラスタ

リング e ができる．

スペクトラルクラスタリングを行うために、スペクトラルクラスタリングの固有値問題を解決する必要がある。そこで、大規模なネットワークを意味のある分割でスペクトラルクラスタリングが行えるようなサイズにハブクラスタリングを使ってネットワークをクラスタリングする。このハブクラスタリングを行うことによって、大規模なネットワークに対してもスペクトラルクラスタリングを行うことができる。

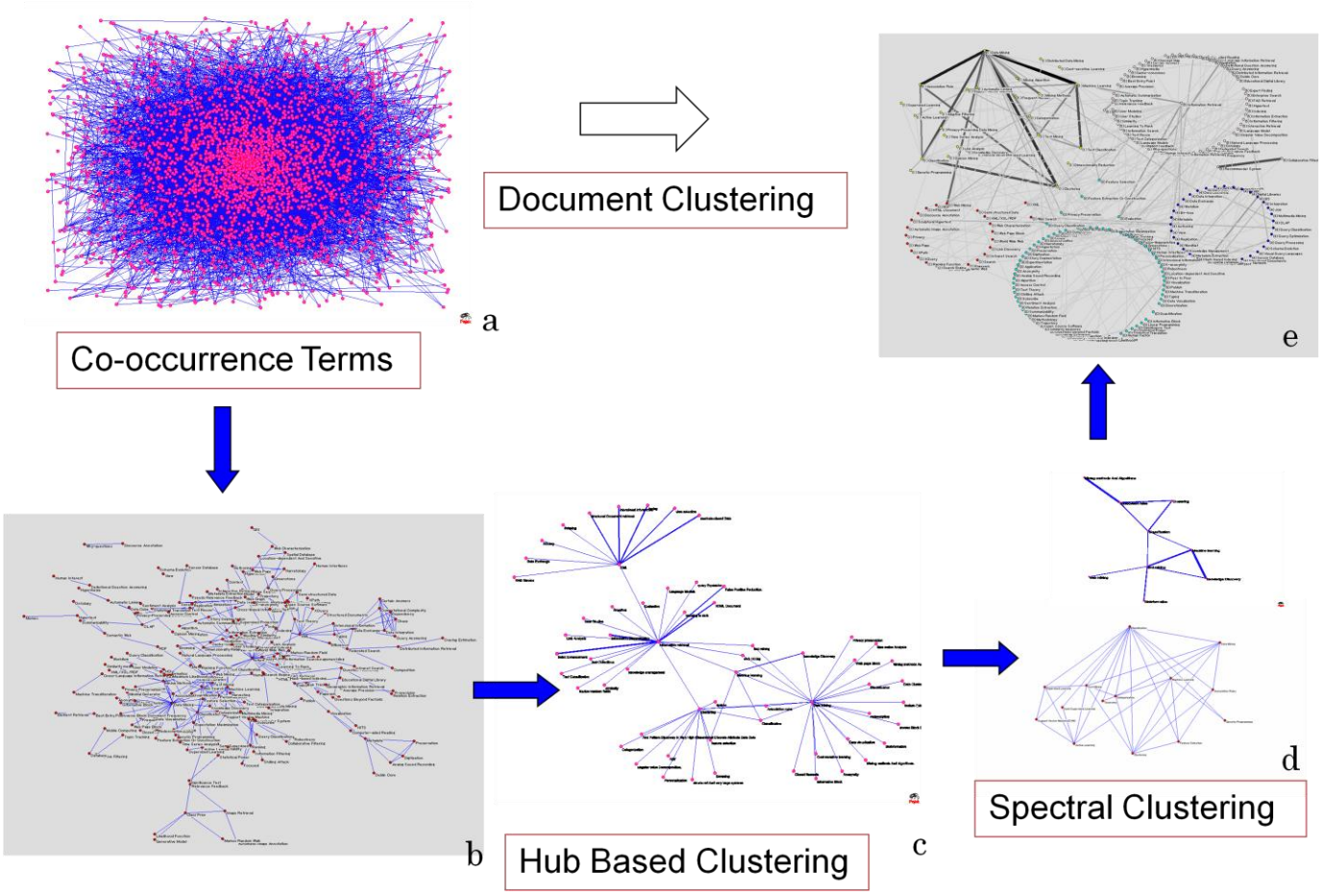



図 1　グラフ構築からの概念マップ抽出までの流れ

本研究では、ドキュメント集合をひとつの世界としてとらえ．各ドキュメントのキーワードに注目する．このとき，ドキュメントのキーワードの欄よりタームを抽出した．

[Co-occurrence Term]

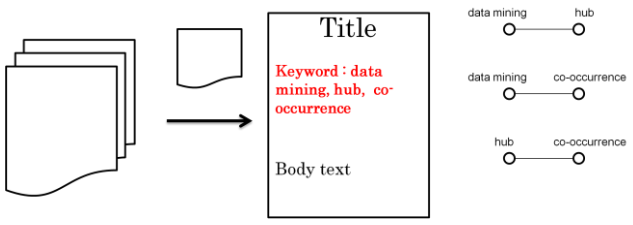



図 2　タームペア生成

1. ID とタームのテーブルからタームペアを作る．具体例を図 2 に示す．“data mining”と“hub”と“co-occurrence”がタームとなり，グラフのノードとなる．エッジは “data mining”と“hub”，“data mining”と“co-occurrence”，“hub”と“co-occurrence”のノード間に付くことになる．

2. タームをノードとし，1 で出来たタームペアにエッジを付ける．これをもとにグラフを構築する．たまたまできたタームペアの使用を避けるために，出現回数に閾値を指定して閾値を越えたタームペアを使う．

[Hub Based Clustering]

3. ２で再構築されたグラフに対してエッジの重みをつける．Cohesion を使い，全体の出現頻度に対するペアの出現頻度をエッジの重みとする．
4. ハブの抽出を行う．ハブには，各ノードに対して接続しているエッジの cohesion で付けた重みの総和を求め，値が大きいものからハブとして取りだす．
5. 取りだした各ハブそれぞれに隣接しているノード同士のグラフを抽出する．これは１，２で構築したグラフの部分グラフにあたる．

$Algorithm:$ *Hub Based Clustering*

*Input* : $a\ graph\ G=(V,E)\ (co-occurrence\ term\ sets)$
$k\ (the\ number\ of\ graph\ partition)$

*Output* : $k\ clusters\ (k\ HNSs\ (Hub\ Node\ Sets))$

*For each edge* $e_j$ *in* $E$
  *For each* $v_i$ *in* $e_j$
    $N\deg ree(v_i) += weight(e_j)$
  *End For*
*End For*
*Sort* $(V, N\deg ree(v), desend)$
$G' = \{v_1, v_2, \cdots, v_k \mid Top\ k\ of\ N\deg ree\}$
*For each* $HNS_i$, $i = 1\ to\ k$
  $HNS_i = HNS_i \cup \{v_i\}$
  *For each* $v_j$ *in* $V$
    *If* $Linking(HNS_i, v_j) = true$
      $HNS_i = HNS_i \cup \{v_j\}$
    *End If*
  *End For*
*End For*

図３　ハブに基づくクラスタリング

[Spectral Clustering]

6. この５で抽出した部分グラフをもとにスペクトラルクラスタリングアルゴリズムを用いてクラスタを作る．
7. ６で行われたグラフカットで出来た部分グラフを概念マップとし，ドキュメント間の関連の特徴づけを行う．

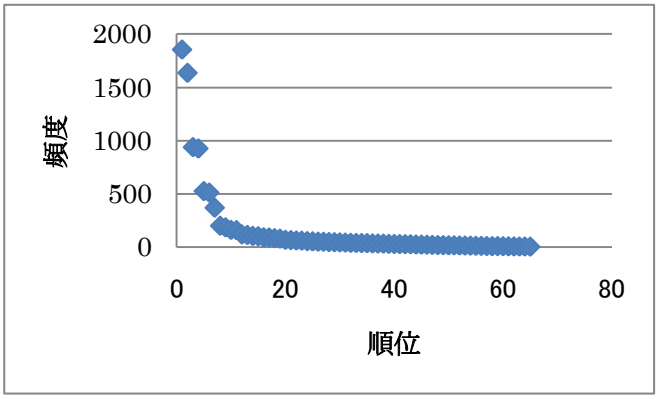
## 4. 実験

今回の実験に使用したデータを表に示す．

表１　使用データ（論文数）

| | |
|---|---|
| SIGIR | ２１５ |
| ACM/SIGMOD | ８２９ |
| DD-2006 | １２１ |
| **総数** | １１６５ |

この実験データより抽出した
タームの総数は 6971 個，タームの種類は 3037 個，タームペアの総数は 20046 個，タームペアの種類は 15283 個となった．この抽出結果からノードは 3037 個，エッジは 15283 個となるのでグラフを構築した．
このネットワークグラフは zipf's low に従っており，スケールフリー性を示した．



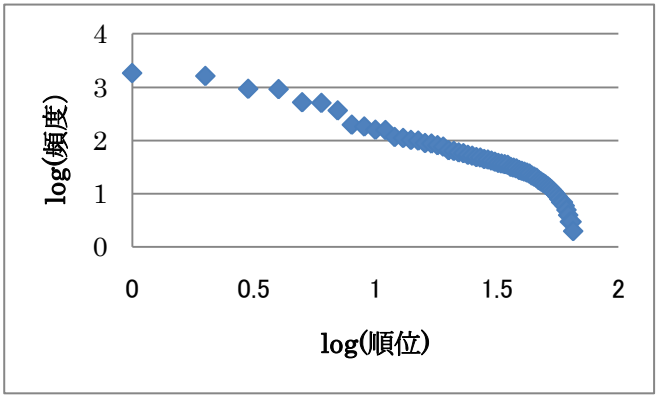




図４　zipf's law に基づく分布

ハブを取り出し，一つのハブから概念マップをいくつか作る．図４は“Data mining”のハブノードからの概念マップである．

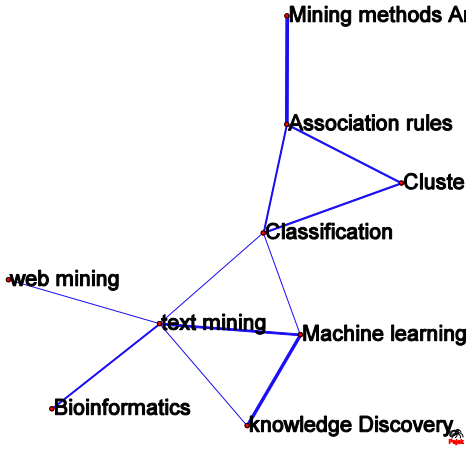



図４　概念マップ例

各ハブ構造ネットワーク内の全てのノードについてそれらのノード間のリンクを全て抽出し、ネットワークを構築することで、概念マップを抽出する。そのネットワーク兄での各２点のノードの平均距離とクラスター度を調べることで、スモールワールド性を調べた。（表２）

表２　クラスター係数

| 順位 | 重みの総和 | keyword | リンク数 | リンクしているノード同士のリンク数 | クラスター係数 |
|---|---|---|---|---|---|
| 1 | 19.67719 | Algorithms | 284 | 701 | 0.001158 |
| 2 | 16.55079 | Experimentation | 216 | 643 | 0.001046 |
| 3 | 11.77862 | Design | 134 | 447 | 0.001344 |
| 4 | 9.777505 | Performance | 116 | 451 | 0.001143 |
| 5 | 8.761626 | Human Factors | 83 | 312 | 0.001711 |
| 6 | 7.852828 | Measurement | 76 | 292 | 0.001789 |
| 7 | 6.762 | Theory | 67 | 284 | 0.001667 |
| 8 | 6.058881 | Certain answers | 9 | 36 | 0.014286 |
| 9 | 6.058881 | chase | 9 | 36 | 0.009524 |
| 10 | 6.058881 | conjunctive queries | 9 | 36 | 0.014286 |

実験結果の例とし”Data mining”の概念マップからの部分グラフとその文献数を表３に示す。

表３　ハブノード“Data mining”の結果

| 共通するノード | 接続ノード | 文献件数 |
|---|---|---|
| Data maining | association rule | 4 |
| | clustering | 4 |
| Data mining | association rule | 3 |
| | mining methods | 3 |
| Data mining | association rule | 2 |
| | mining methods | 2 |
| Data mining | closed itemset | 2 |
| | minimal generator | 2 |
| Data mining | clustering | 2 |
| | singular value decomposition | 2 |
| Data mining | clustering | 2 |
| | classification | 2 |
| Data mining | knowledge discovery | 3 |
| | text mining | 3 |
| Data mining | knowledge discovery | 2 |
| | text mining | 2 |
| Data mining | knowledge discovery | 2 |
| | text classification | 2 |

## 5. 分析と考察

ハブを抽出したときに、抽出したハブに隣接しているノード(サブノード)が、ハブの場合がある。(図５)このようなノードはネットワークの中で特に強い概念を持つノードではないかと考えられる。

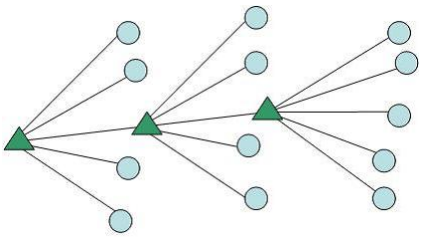


図５　ハブノードの派生

また、図６のように２点のハブノードに隣接するノードが複数ある場合、抽出数をわずかに増加するだけでハブノードになるノードがある一方で、膨大な数の抽出数でもハブノードに変化しないノードが存在する。これは重みの総和が高いだけでなく、その歳のそれらのノードが持つリンクの本数にも影響があると考えられる。

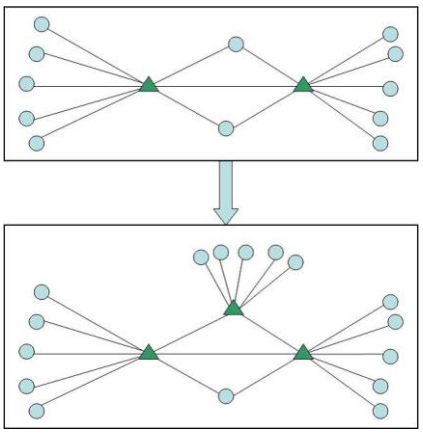


図６　ハブノードの派生

## 6. おわりに

本研究では，大規模なデータを扱うのに適したグラ

フ表現を用い，ドキュメント集合からなる複雑なネットワークの分類を効果的に行うアプローチをとった．これにより，概念マップに基づくドキュメントクラスタリングができた．

本論文の特徴としては、共起タームを利用したグラフからのハブクラスタリングとスペクトラルクラスタリングを行い、概念マップを抽出するところである。概念マップ抽出によりその概念マップをインデックスとした検索システムを可能とする。

共起タームからネットワークグラフを構築することで言語世界のスケールフリー性に注目する。そして，そのスケールフリー性がもつハブというか概念を用いて、ハブに基づくグラフクラスタリングによるサブグラフの作成する．ハブクラスタリングすることで，大規模なデータを意味のあるクラスタに分割し、クラスタサイズを小さくすることでスペクトラルクラスタリングを適応できるようになる。スペクトラルクラスタリングは高い質でクラスタリングを行うことができるクラスタリング手法である。

今後，より大規模なドキュメント集合に適応できる効果的な高速スペクトラルクラスタリングアルゴリズム（乱択アルゴリズムを含む）の開発を進める．[8]

また、実験データのドキュメントからのキーワード抽出を工夫することでより精度の高い結果が得られると考えられる．そして，より高速な処理を可能とするためにスペクトラルクラスタリングの改良が必要となる．

## 参　考　文　献


[1] A.L.Barabasi, R Albert, H.Jeong, and G.Bianconi: "Power-law distribution of the world wide web.Science", 287, 2000.

[2] A.L.Barabasi, Reka Albert: "Emergence of Scaling in Random network", SCIENCE Vol 286 p509-512, 1999.

[3] Rohini K. Srihari, Sudarshan Lamkhede, Anmol Bhasin: “Unapparent Information Revelation: A Concept Chain Graph Approach”, CIKM'05, 2005.

[4] Illhoi Yoo, Xiaohua Hu, Il Yeol Song "Integrating Biomedical Literature Clustering nd Summriztion Approches using Biomedical Ontology", ACM, 2006.

[5] Illhoi Yoo, Xiaohua Hu, Il Yeol Song: "Clustering Ontology-enriched Graph Representation for Biomedical Documents based on Scale-Free Network Theory", 2006 3rd International IEEE conference on volume, p851-858, 2006.

[6] Brant Chee, Bruce Schatz: "Document Clustering using Small world community", JCDL’07, 53-60, 2007.

[7] L. da F. Costa, Hub-Based Community Finding, arXiv:cond-mat/0405022v1, 2004.

[8] Y.Wng, H.Song and W.Wang, A Microscopic View on Community Detection in Complex Networks, PIKM’08, 57-64, 2008.

[9] Y.Chi, X.Song, D.Zhou, K.Hino, and B.Tseng, Evolutionary Spectral Clustering by Incorporating Temporal Smoothness, KDD’07.

[10] X.Wang and I.Davidson, Flexible Constrained Spectral Clustering, KDD’10, 563-572, 2010.