CMSC5733 Social Computing

Community Question Answering
and
Community Detection

Department of Computer Science and Engineering
The Chinese University of Hong Kong

October 7, 2013

Outline

1. Community Question Answering
   - Introduction
   - Question Subjectivity Analysis
   - Question Retrieval
2. Community Detection
   - Introduction
   - Methods
   - Summary
3. References

Outline

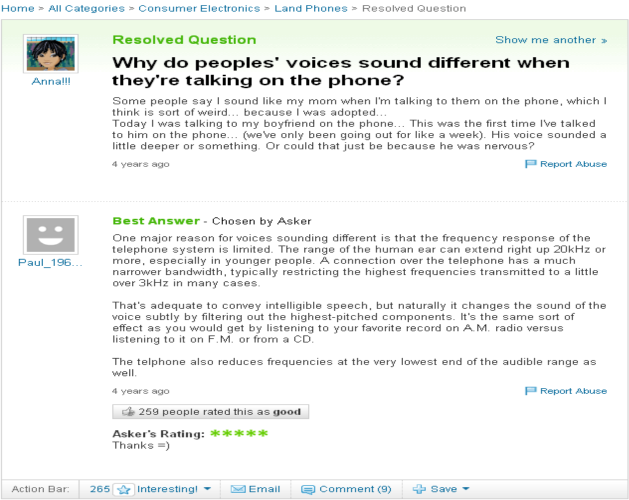
Outline

1. Community Question Answering
   - Introduction
   - Question Subjectivity Analysis
   - Question Retrieval
2. Community Detection
   - Introduction
   - Methods
   - Summary
3. References

Community Question Answering

- Knowledge dissemination, information seeking
- Natural language questions
- Explicit, self-contained answers

Community Question Answering



Community Question Answering

Artificial Neural Networks | Machine Learning | Edit

How do convolutional neural networks work? Edit

Especially, what kind of benefits does convolution give you? Edit

Comment · Post (1) · Wiki · Options · Redirect Question

2 Answers · **Create Answer Wiki**

Mikio L. Braun, Ph.D. in machine learning, 10+ years ...

3 votes by Kat Li, Barak Cohen, and Lucian Sasu

Convolutional neural networks work like learnable local filters.

The best example is probably their application to computer vision. The first step in image analysis is often to perform some local filtering of the image, for example, to enhance edges in the image.

You do this by taking the neighborhood of each pixel and convolve it with a certain mask (set of weights). Basically you compute a linear combination of those pixels. For example, if you have a positive weight on the center pixel and negative weights on the surrounding pixels you compute the difference between the center pixel and the surrounding, giving you a crude kind of edge detector.

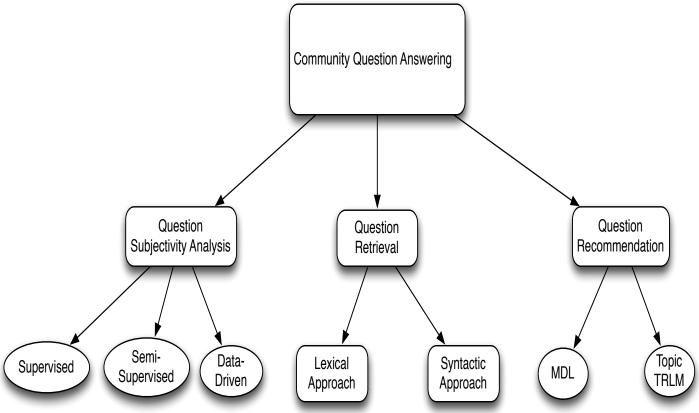
Now you can either put that filter in there by hand or learn the right filter through a convolutional neural network. If we consider the simplest case, you have an input layer representing all pixels in your image while the output layer representing the filter responses. Each node in the output layer is connected to a pixel and its neighborhood in the input layer. So far, so good. What makes convolutional neural networks special is that the weights are shared, that is, they are the same for different pixels in the image (but different with respect to the position relative to the center pixel). That way you effectively learn a filter, which also turns out to be suited to the problem you are trying to learn.

Comment · Post · Thank · Sep 29, 2011

Advantages of Community Question Answering

- Could solve information needs that are personal, heterogeneous, specific, open-ended, and cannot be expressed as a short query
- No single Web page will directly answer these complex and heterogeneous needs, CQA users should understand and answer better than a machine
- Have accumulated rich knowledge
  - More than one billion posted answers in Yahoo! Answers http://yanswersblog.com/index.php/archives/2010/05/03/1-billion-answers-served/
  - More than 190 million resolved questions in Baidu Zhidao
  - In China, 25% of Google's top-search-results page contain at least one link to some Q&A site, Si et al., VLDB, 2010

Community Question Answering



Outline

Question Subjectivity Analysis

- Question Analysis is to analyze characteristics of questions
- Understand User Intent
- Provide rich information to question search, question recommendation, answer quality prediction, etc.
- Question Subjectivity Analysis is an important aspect of question analysis

Question Subjectivity Analysis: Definition

- Subjective question
  - Private statements
  - Personal opinion and experience
  - What's the difference between chemotherapy and radiation treatments?
- Objective question
  - Objective, verifiable information
  - Often with support from reliable sources
  - Has anyone got one of those home blood pressure monitors? and if so what make is it and do you think they are worth getting?

Question Subjectivity Analysis: Motivation

- More accurately identify similar questions, improve question search
- Better rank or filter the answers based on whether an answer matches the question orientation
- Crucial component of inferring user intent, a long-standing problem in Web search
- Route subjective questions to users for answer, trigger automatic factual question answering system for objective questions

Question Subjectivity Analysis: Challenge

- Ill-formatted, e.g., word capitalization may be incorrect or missing, consecutive words may be concatenated
- Ungrammatical, include common online idioms, e.g., using “u” means “you”, “2” means “to”

Question Subjectivity Analysis: Supervised Learning

- Baoli Li, Yandong Liu, Ashwin Ram, Ernest V. Garcia and Eugene Agichtein, Exploring Question Subjectivity Prediction in Community QA, SIGIR, 2008
- Support Vector Machine with linear kernel
- Features
  - Character 3-gram
  - Word
  - Word + character 3-gram
  - Word n-gram
  - Word POS n-gram, mix of word and POS tri-grams
- Term weighting schemes: binary, TF, TF*IDF

Question Subjectivity Analysis: Semi-Supervised Learning

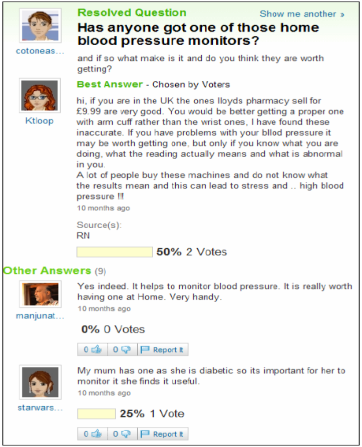


Figure: **Yahoo Answers Example.**

- Baoli Li, Yandong Liu and Eugene Agichtein, CoCQA: Co-Training Over Questions and Answers with an Application to Predicting Question Subjectivity Orientation, EMNLP, 2008
- Incorporate relationships between questions and corresponding answers
- Co-training, two views of the data, question and answer

Input:

- F_Q and F_A are *Question* and *Answer* feature views
- C_Q and C_A are classifiers trained on F_Q and F_A respectively
- **L** is a set of labeled training examples
- **U** is a set of unlabeled examples
- K: Number of unlabeled examples to choose on each iteration
- X: the threshold for increment
- R: the maximal number of iterations

Algorithm ***CoCQA***

1. Train $C_{Q,0}$ on **L**: F_Q, and record resulting $ACC_{Q,0}$
2. Train $C_{A,0}$ on **L**: F_A, and record resulting $ACC_{A,0}$
3. **for** j=1 to R **do**:
   Use $C_{Q,j-1}$ to predict labels for **U** and choose
   top K items with highest confidence → $E_{Q,j-1}$
   Use $C_{A,j-1}$ to predict labels for **U** and choose
   top K items with highest confidence → $E_{A,j-1}$
   Move examples $E_{Q,j-1} \cup E_{A,j-1}$ → **L**
   Train $C_{Q,j}$ on **L**: F_Q and record training $ACC_{Q,j}$
   Train $C_{A,j}$ on **L**: F_A and record training $ACC_{A,j}$
   if $Max(\Delta ACC_{Q,j}, \Delta ACC_{A,j}) < X$ **break**
4. **return** final classifiers $C_{Q,j}$ → C_Q and $C_{A,j}$ → C_A

- At step 1,2, each category has top K_j most confident examples chosen as additional "labeled" data
- Terminate when the increments of both classifiers are less than threshold X or maximum number of iterations are exceeded

Question Subjectivity Analysis: Data-driven Approach

- Tom Chao Zhou, Xiance Si, Edward Y. Chang, Irwin King and Michael R. Lyu, A Data-Driven Approach to Question Subjectivity Identification in Community Question Answering, AAAI, 2012
- Li et al. 2008 (supervised), Li et al. 2008 (CoCQA, semi-supervised) based on manual labeling data
- Manual labeling data is quite expensive

Question Subjectivity Analysis: Data-driven Approach

Web-scale learning is to use available large-scale data rather than hoping for annotated data that isn't available

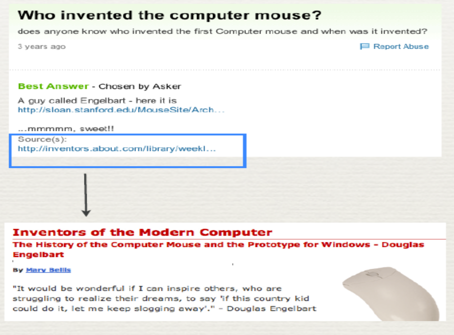
- Halevy, Norvig and Pereira

Question Subjectivity Analysis: Data-driven Approach

Whether we can utilize social signals to collect training data for question subjectivity identification with NO manual labeling?

- Like Signal: like an answer if they find the answer useful
- Intuition
  - Subjective: answers are opinions, different tastes; best answer receives similar number of likes with other answers
  - Objective: like an answer which explains universal truth in most detail; best answer receives high likes than other answers

- Vote Signal: users could vote for best answer
- Intuition
  - Subjective: vote for different answers, support different opinions; low percentage of votes on best answer
  - Objective: easy to identify answer contain the most fact; percentage of votes of best answer is high



- Source Signal: reference to authoritative resources
- Intuition
  - Only available for objective question that has fact answer

Question Subjectivity Analysis: Data-driven Approach

- Poll and Survey signal
- User intent is to seek opinions
- Very likely to be subjective

- What is something you learned in school that you think is useful to you today?
- If you could be a cartoon character, who would you want to be?

Question Subjectivity Analysis: Data-driven Approach

- Answer Number signal: the number of posted answers to each question
- Intuition
  - Subjective: alertpost opinions even they notice there are other answers
  - Objective: may not post answers to questions that has received other answers since an expected answer is usually fixed
  - A large answer number indicate subjectivity
  - A small answer number may be due to many reasons, such as objectivity, small page views

Question Subjectivity Analysis: Data-driven Approach

Summary of Social Signals		
Name	Description	Training Data
Like	Capture users' tastes	Positive && Negative
Vote	Reflect users' judgments	Positive && Negative
Source	Measure confidence on authoritativeness	Negative
Poll and Survey	Indicate users' intent	Positive
Answer Number	Imply users' willingness to answer a question	Positive

Question Subjectivity Analysis: Data-driven Approach

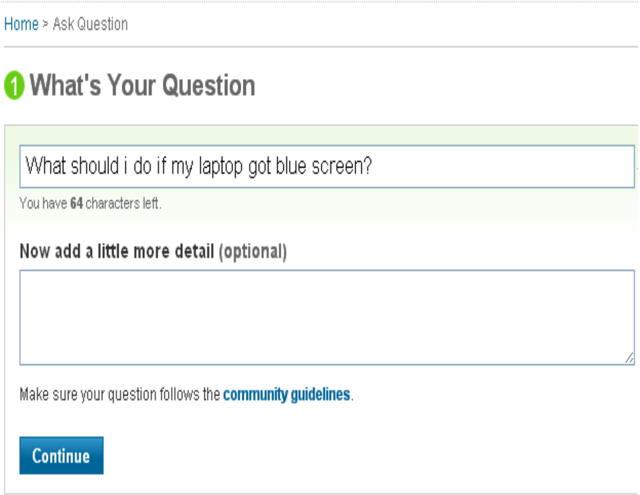
- Features
  - Word: term frequency
  - Word n-gram: term frequency
  - Question length: information needs of subjective questions are complex, users use descriptions to explain, larger question length
  - Request word: particular words to explicitly indicate their request for seeking opinions; manual list of 9 words

Question Subjectivity Analysis: Data-driven Approach

- Subjectivity clue: external lexicon, over 8000 clues, manually compiled word list from news to express opinions
- Punctuation density: density of punctuation marks
- Grammatical modifier: inspired by opinion mining research of using grammatical modifiers on judging users' opinions, adjective and adverb
- Entity: objective question expects fact answer, leading to less relationships among entities, subjective questions contains more descriptions, may involve relatively complex relations

Outline

Ask a Question



Problem and Opportunity

- Problem
  - Askers need to wait some time to get an answer, time lag
  - 15% of the questions do not receive any answer in Yahoo! Answers, which is one of the first CQA sites on the Web
- Opportunity
  - 25% questions in certain categories are recurrent, Anna, Gideon and Yoelle, WWW, 2012
- Answer new questions by reusing past resolved questions
- Question Retrieval: find semantically similar past questions for a new question

Question Retrieval Example

Search

What should i do if my laptop got blue screen?

Search Y! Answers

Sort by: Relevance | Newest | Most Answers

What should I do if I keep **getting** the "**blue screen** of death" for **my** Windows7 **laptop**?

...**I** keep **getting** the **blue screen** of death telling **me** that the pc is **getting** prepared for a... scary to imagine **what** would happen **if I** wasn't. **I** just bought this Windows7 Toshiba **laptop** from office depot in the summer...to crash (so early)? **What should I do**?

In Laptops & Notebooks - Asked by nelson316@verizon.net - 4 answers - 4 months ago

I just **got** a random **blue screen** of death, **should I** be worried?

...just suddenly **got** a random **blue screen** of death. **I**'ve never **got** one on this **laptop** before, and **I**'ve had no problems with **my laptop** at all until this bsod.... It said that **if** it was **my** first time...free. **I** don't even remember **what** sites **I** was on...with no problems. **I do** remember that the programs...off bsod like **my** old **laptop**? Or **should I** be worried? ...

1 In Other - Hardware - Asked by Kaylee - 6 answers - 2 weeks ago

why is **my laptop** showing the **blue screen**?

...to it, so **I**'m not sure **if** they could have **done** anything, but now when **i** turn **my laptop** on **i** would **get** a **blue screen** saying all this jumble...a boot disk, so **i** dont know **what** else **should i do**. Any help/advice?

In Laptops & Notebooks - Asked by doodiec - 6 answers - 5 years ago

Laptop blue screen problem!!!?

...malicious URL block and then this **blue screen** comes up and **my laptop** turns off and asks **me if I** want to go into safe mode. **What should I do**? Is there any way...for a new **laptop** cause **I got** low practice SAT scores...

In Laptops & Notebooks - Asked by Mathhew Colman - 5 answers - 10 months ago

sony vaio **blue screen** problem, **what should i do**? please help?

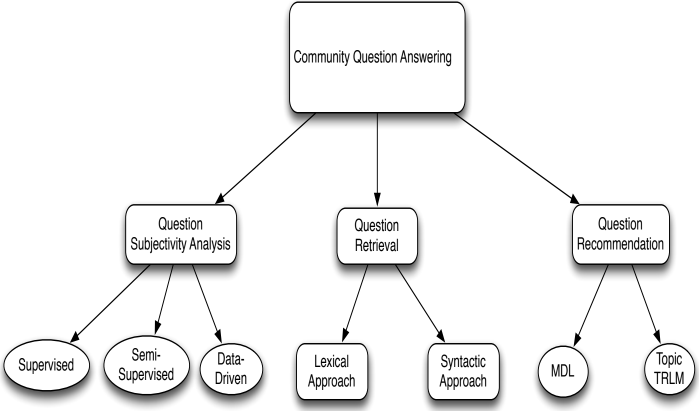
Benefit of Question Retrieval

- Provide an alternative to automatic question answering
- Help askers get an answer in a timely manner
- Guide answerers to answer unique questions, better utilize users' answering passion

Notations

Symbol	Description
Q	A new question
D	A candidate question
$\lvert \cdot \rvert$	Length of the text
C	Background collection
w	A term in the new question
t	A term in a candidate question

Community Question Answering



Lexical-based Approach: Language Model

- In language modeling, similarity between a query and a document is given by the probability of generating the query from the document language model
- Unigram language model, i.i.d. sampling

$$P(Q|D) = \prod_{w \in Q} P(w|D)$$

- In question retrieval syntax, query is the new question, document is a candidate question

Lexical-based Approach: Language Model

- To avoid zero probabilities and estimate more accurate language models, documents are smoothed using a background collection

$$P(w|D) = (1-\lambda)P_{ml}(w|D) + \lambda P_{ml}(w|C)$$

- λ is a smoothing parameter, $0 \leq \lambda \leq 1$

$$P_{ml}(w|D) = \frac{termfrequency(w, D)}{\sum_{w' \in D} termfrequency(w', D)}$$

- Maximum likelihood estimator to calculate $P_{ml}(\cdot)$

Language Model Example

- Query (q): revenue down
- Document 1 (d_1): xyzzy reports a profit but revenue is down
- Document 2 (d_2): quorus narrows quarter loss but revenue decreases further
- $\lambda = 0.5$

$$P(Q|D) = \prod_{w \in Q} P(w|D)$$

$$P(w|D) = (1 - \lambda) P_{ml}(w|D) + \lambda P_{ml}(w|C)$$

$$P(q|d_1) = [(1/8 + 2/16)/2] \times [(1/8 + 1/16)/2] = 3/256$$

$$P(q|d_2) = [(1/8 + 2/16)/2] \times [(0/8 + 1/16)/2] = 1/256$$

- Ranking: $d_1 > d_2$

Lexical-based Approach: Translation Model

LM	
Advantage	Simple
Disadvantage	Lexical Gap

- Lexical Gap, two questions that have the same meaning use very different wording
  - Is downloading movies illegal?
  - Can I share a copy of a DVD online?
- Jiwoon Jeon, W. Bruce Croft and Joon Ho Lee, Finding Similar Questions in Large Question and Answer Archives, CIKM, 2005

Lexical-based Approach: Translation Model

Language Model
$P(w\|D) = (1-\lambda)P_{ml}(w\|D) + \lambda P_{ml}(w\|C)$
Translation Model
$P(w\|D) = (1-\lambda)\sum_{t \in D}(T(w\|t)P_{ml}(t\|D)) + \lambda P_{ml}(w\|C)$

- $T(w|t)$ is the probability that word w is the translation of word t, denotes semantic similarities between words

Lexical-based Approach: Translation Model

Table: Questions share few common words, but may have high semantic relatedness according to translation model

Id like to insert music into PowerPoint. How can I link sounds in PowerPoint?
How can I shut down my system in Dos-mode. How to turn off computers in Dos-mode.
Photo transfer from cell phones to computers. How to move photos taken by cell phones.
Which application can run bin files? I download a game. How can I execute bin files?

Rank	bmp	format	music	intel	excel	font	watch	memory
1	bmp	format	music	pentium	excel	font	watch	memory
2	jpg	format*	file	4	korean	korean	time	virtual
3	gif	xp	tag	celeron	function	97	background	shortage
4	save	windows	sound	amd	novice	add	start	ram
5	file	hard	background	intel	cell	download	date	message
6	picture	98	song	performance	disappear	control-panel	display	configuration
7	change	partition	play	support	convert	register	tray	256
8	ms-paint	drive	mp3	question	if	install	power	extend
9	convert	disk	cd	buy	xls	default	screen	system
10	photo	C	source	cpu	record	photoshop	wrong	windows

Figure: The first row shows the source words and each column shows top 10 words that are most semantically similar to the source word. A higher rank means a larger $T(w|t)$ value

Lexical-based Approach: Translation Model

- How to learn $T(w|t)$?
  - Prepare a monolingual parallel corpus of pairs of text, each pair should be semantically similar
  - Employ machine translation model IBM model 1 on the parallel corpus to learn $T(w|t)$
  - IBM model 1: Brown et al., Computational Linguistics, 1990
- How this paper prepares monolingual parallel corpus
  - Each pair contains two questions whose answers are very similar

Lexical-based Approach: Translation Model

- Delphine Bernhard and Iryna Gurevych, Combining Lexical Semantic Resources with Question & Answer Archives for Translation-Based Answer Finding, ACL, 2009
- Propose several methods to prepare parallel monolingual corpora
  - Question answer pairs: question $\leftrightarrow$ answer
  - Question reformulation pairs: question $\leftrightarrow$ question reformulation by user

Lexical-based Approach: Translation Model

RUClimate (supervisor) [332] merged the question **Why iare clouds white** into **Why are clouds white** 9 Feb 2012 17:03

RUClimate (supervisor) [332] merged the question **What makes the clouds appeared to be white** into **Why are clouds white** 9 Feb 2012 16:44

RUClimate (supervisor) [332] merged the question **Why does Clouds appear white** into **Why are clouds white** 9 Feb 2012 16:44

RUClimate (supervisor) [332] merged the question **Why do clouds appear white** into **Why are clouds white** 9 Feb 2012 16:43

RUClimate (supervisor) [332] merged the question **Why do clouds look white** into **Why are clouds white** 9 Feb 2012 16:43

RUClimate (supervisor) [332] merged the question **Why do clouds in the sky appear white** into **Why are clouds white** 9 Feb 2012 16:43

RUClimate (supervisor) [332] merged the question **How does the cloud is white** into **Why are clouds white** 9 Feb 2012 16:43

Lexical-based Approach: Translation Model

- Lexical Semantic Resources: glosses and definitions for the same lexeme in different lexical semantic and encyclopedic resources can be considered as near-paraphrases, since they define the same terms and hence have the same meaning
- moon
  - Wordnet: the natural satellite of the Earth
  - English Wiktionary: the Moon, the satellite of planet Earth
  - English Wikipedia: the Moon (Latin: Luna) is Earth's only natural satellite and the fifth largest natural satellite in the Solar System

Lexical-based Approach: Translation-based Language Model

TM	
Advantage	Tackle lexical gap to some extent
Disadvantage	$T(w\|w) = 1$ for all w while maintaining other word translation probabilities unchanged, produce inconsistent probability estimates and make the model unstable

- Xiaobing Xue, Jiwoon Jeon and W. Bruce Croft, Retrieval Models for Question and Answer Archives, SiGIR, 2008
- Translation-based Language Model

Lexical-based Approach: Translation-based Language Model

Translation Model
$P(w\|D) = (1-\lambda)\sum_{t\in D}(T(w\|t)P_{ml}(t\|D)) + \lambda P_{ml}(w\|C)$
Translation-based Language Model
$P(w\|D) = \frac{\|D\|}{\|D\|+\lambda}P_{mx}(w\|D) + \frac{\lambda}{\|D\|+\lambda}P_{ml}(w\|C)$ $P_{mx}(w\|D) = (1-\beta)P_{ml}(w\|D) + \beta\sum_{t\in D}T(w\|t)P_{ml}(t\|D)$

- Linear combination of language model and translation model
- Answer part should provide additional evidence about relevance, incorporating the answer part

$$P_{mx}(w|(D,A)) = \alpha P_{ml}(w|D) + \beta\sum_{t\in D}T(w|t)P_{ml}(t|D) + \gamma P_{ml}(w|A)$$

$$\alpha + \beta + \gamma = 1$$

Syntactic-based Approach: Syntactic Tree Matching

- Some similar questions neither share many common words, nor follow identical syntactic structure
  - How can I lose weight in a few months?
  - Are there any ways of losing pound in a short period?
- Kai Wang, Zhaoyan Ming and Tat-Seng Chua, A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services, SIGIR, 2009
- Syntactic tree matching

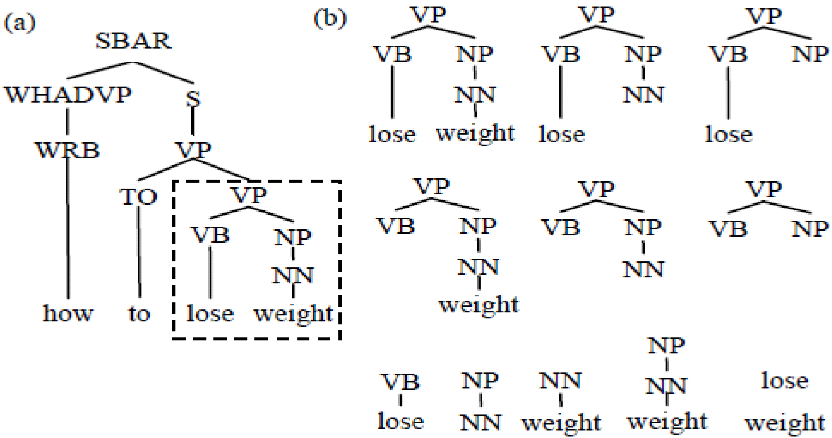


Figure: (a) The Syntactic Tree of the Question "How to lose weight?". (b) Tree Fragments of the Sub-tree covering "lose weight".

Syntactic-based Approach: Syntactic Tree Matching

- Tree kernel: utilize structural or syntactic information to capture higher order dependencies between grammar rules

$$k(T_1, T_2) = \sum_{n_1 \in N_1} \sum_{n_2 \in N2} C(n_1, n_2)$$

- N_1, N_2 are sets of nodes in two syntactic trees T_1 and T_2, and $C(n_1, n_2)$ equals to the number of common fragments rooted in n_1 and n_2

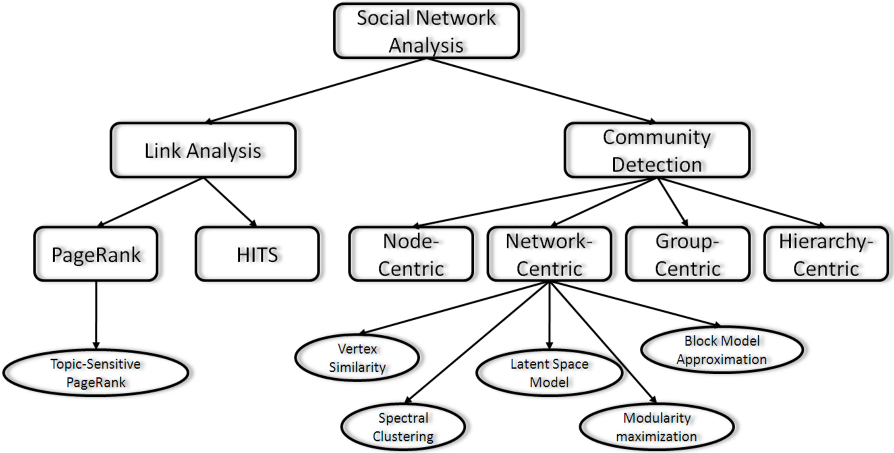
Syntactic-based Approach: Syntactic Tree Matching

- Limitation of tree kernel
  - Tree kernel function merely replies on the intuition of counting the common number of sub-trees, whereas the number might not be a good indicator of the similarity between two questions
  - Two evaluated sub-trees have to be identical to allow further parent matching, for which semantic representations cannot fit in well
- Syntactic tree matching
  - A new weighting scheme for tree fragments that are robust against some grammatical errors
  - Incorporate semantic features

Outline

Outline

Content


Social Network Analysis
Link Analysis
Community Detection
PageRank
HITS
Node-Centric
Network-Centric
Group-Centric
Hierarchy-Centric
Topic-Sensitive PageRank
Vertex Similarity
Latent Space Model
Block Model Approximation
Spectral Clustering
Modularity maximization

Communities

Community

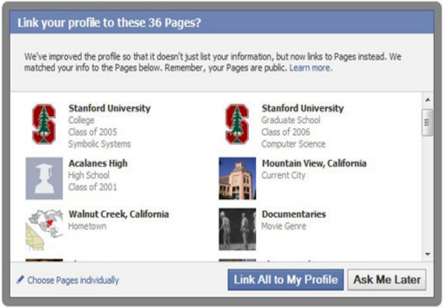
A community is formed by individuals such that those within a group interact with each other more frequently than with those outside the group.

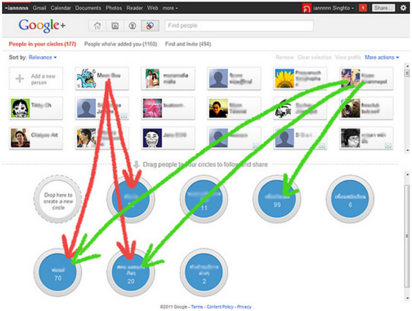
- Users form communities in social media
- Community is formed through frequent interacting
- A set of users who do not interact with each other is not a community

Why Communities Are Formed?

- Human beings are social
- Social media are easy to use
  - People's social lives are easy to extend with the help of social media
- People connect with friends, relatives, colleges, etc. in the physical world as well as online

Examples of Communities





Community Detection

Two Types of Users

1. Explicit Groups: Formed by user subscriptions
   - E.g., Groups in Facebook
2. Implicit Groups: implicitly formed by social interactions
   - E.g., Community question answering

Community Detection

Discovering groups in a network where individuals' group memberships are not explicitly given

Outline

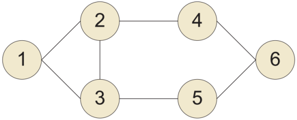
Approaches

Four categories

- Node-centric approach
  - Each node in a group satisfies certain properties
- Group-centric approach
  - Consider the connections inside a group as a whole
- **Network-centric approach**
  - Partition nodes of a network into several disjoint sets
- Hierarchy-centric approach
  - Build a hierarchical structure of communities based on network topology

Node-Centric Community Detection

- Nodes satisfying certain properties within a group
  - Complete mutuality
    - cliques: A clique is a maximum complete subgraph in which all nodes are adjacent to each other
  - Reachability of members
    - k-clique: A k-clique is a maximal subgraph in which the largest geodesic distance between any two nodes is no greater than k
    - k-club: The geodesic distance within the group (i.e., diameter) to be no greater than k
    - A k-clique might have diameter larger than k in the subgraph! E.g. {1, 2, 3, 4, 5}



cliques: {1, 2, 3}
2-cliques: {1, 2, 3, 4, 5}, {2, 3, 4, 5, 6}
2-clubs: {1, 2, 3, 4}, {1, 2, 3, 5}, {2, 3, 4, 5, 6}

Group-Centric Community Detection

Density-Based Groups

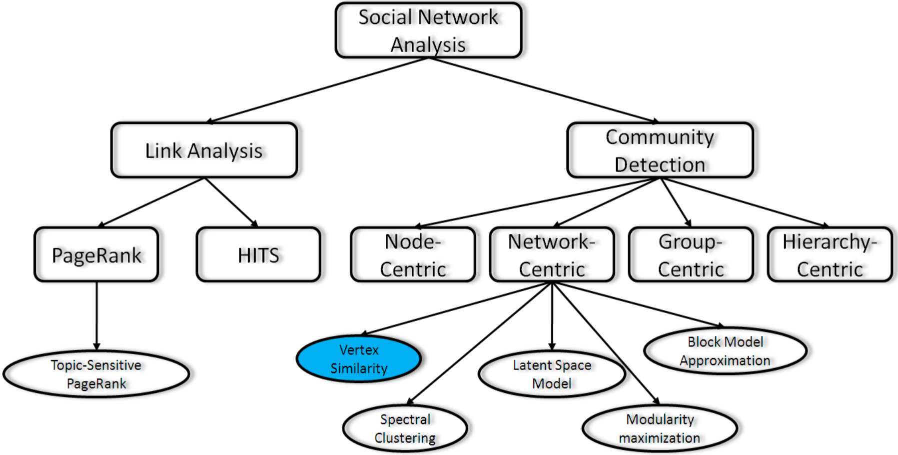
- It is acceptable for some nodes to have low connectivity
- The whole group satisfies a certain condition
  - E.g., the group density $\geq$ a given threshold
- A subgraph $G_s(V_s, E_s)$ is $\gamma - dense$ (*quasi-clique*, Abello et al., 2002) if

$$\frac{E_s}{V_s(V_s-1)/2} \geq \gamma$$

- Greedy search through recursive pruning
  - Local search: sample a subgraph and find a maximum $\gamma - dense$ quasi-clique (say, of size k)
  - Heuristic pruning: remove nodes with degree less than $k \cdot \gamma$

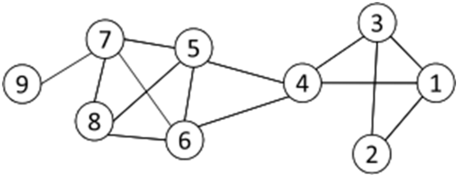
Network-Centric Community Detection

- Consider the global topology of a network
- Partition nodes of a network into disjoint sets
- Optimize a criterion defined over a partition rather than over one group
- Approaches:
  - Clustering based on vertex similarity
  - Latent space models (multi-dimensional scaling)
  - Block model approximation
  - Spectral clustering
  - Modularity maximization


Social Network Analysis
Link Analysis
Community Detection
PageRank
HITS
Node-Centric
Network-Centric
Group-Centric
Hierarchy-Centric
Topic-Sensitive PageRank
Vertex Similarity
Spectral Clustering
Latent Space Model
Modularity maximization
Block Model Approximation

Clustering Based on Vertex Similarity

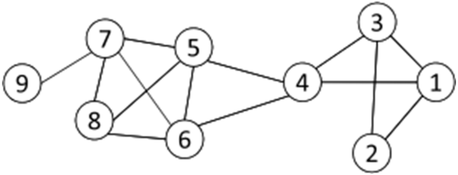
- Vertex similarity is defined in terms of the similarity of their social circles
- Structural equivalence: two nodes are structurally equivalent iff they are connecting to the same set of actors



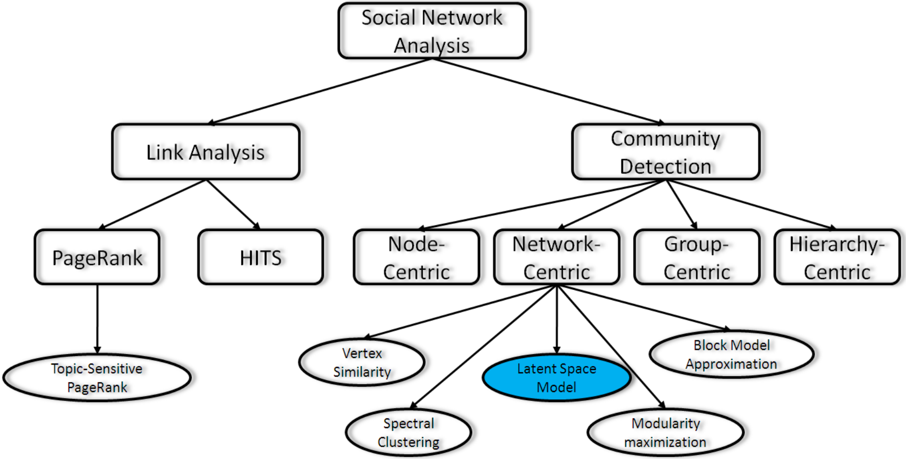
  - Nodes 1 and 3 are structurally equivalent; So are nodes 5 and 6.
  - Structural equivalence is too restrict for practical use
- Apply k-means to find communities

Vertex Similarity Measurements

- Cosine Similarity: $Cosine(v_i, v_j) = \frac{|N_i \bigcap N_j|}{\sqrt{|N_i| \cdot |N_j|}}$
- Jaccard Similarity: $Jaccard(v_i, v_j) = \frac{|N_i \bigcap N_j|}{|N_i \bigcup N_j|}$



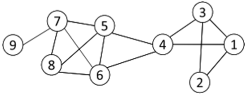
$$
\begin{aligned}
Cosine(4,6) &= \frac{1}{\sqrt{4 \cdot 4}} = \frac{1}{4} \\
Jaccard(4,6) &= \frac{|\{5\}|}{|\{1,3,4,5,6,7,8\}|} = \frac{1}{7}
\end{aligned}
$$


Social Network Analysis
Link Analysis
Community Detection
PageRank
HITS
Node-Centric
Network-Centric
Group-Centric
Hierarchy-Centric
Topic-Sensitive PageRank
Vertex Similarity
Latent Space Model
Block Model Approximation
Spectral Clustering
Modularity maximization

Latent Space Models

- Map nodes into a low-dimensional Euclidean space such that the proximity between nodes based on network connectivity are kept in the new space
- Multi-dimensional scaling (MDS)
  - Given a network, construct a proximity matrix $P \in \mathbb{R}^{n \times n}$ representing the pairwise distance between nodes
  - Let $S \in \mathbb{R}^{n \times k}$ denote the coordinates of nodes in the low-dimensional space
  $$SS^T \approx -\frac{1}{2}(I - \frac{1}{n}ee^T)(P \circ P)(I - \frac{1}{n}ee^T) = \tilde{P},$$
  where $\circ$ is the element-wise matrix multiplication
  - Objective: $\min ||SS^T - \tilde{P}||_F^2$
  - Let $\Lambda = diag(\lambda_1, ..., \lambda_k)$ (the top-k eigenvalues of $\tilde{P}$), V the top-k eigenvectors
  - Solution: $S = \Lambda V^{1/2}$
- Apply k-means to S to obtain communities

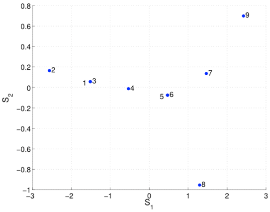
Example of MDS



Two communities:
{1, 2, 3, 4} and {5, 6, 7, 8, 9}

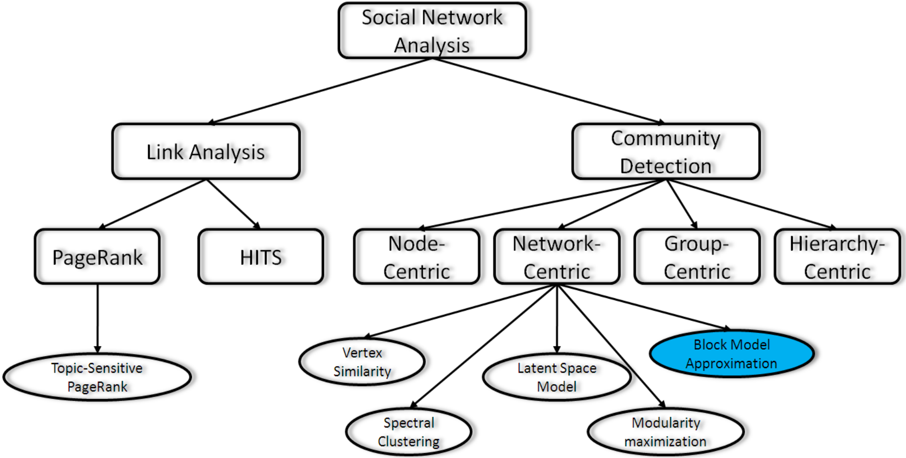
geodesic distance

$$P = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 & 2 & 3 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 & 3 & 4 & 4 & 5 \\ 1 & 1 & 0 & 1 & 2 & 2 & 3 & 3 & 4 \\ 1 & 2 & 1 & 0 & 1 & 1 & 2 & 2 & 3 \\ 2 & 3 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \\ 2 & 3 & 2 & 1 & 1 & 0 & 1 & 1 & 2 \\ 3 & 4 & 3 & 2 & 1 & 1 & 0 & 1 & 1 \\ 3 & 4 & 3 & 2 & 1 & 1 & 1 & 0 & 2 \\ 4 & 5 & 4 & 3 & 2 & 2 & 1 & 2 & 0 \end{bmatrix}$$



$$V = \begin{bmatrix} -0.33 & 0.05 \\ -0.55 & 0.14 \\ -0.33 & 0.05 \\ -0.11 & -0.01 \\ 0.10 & -0.06 \\ 0.10 & -0.06 \\ 0.32 & 0.11 \\ 0.28 & -0.79 \\ 0.52 & 0.58 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 21.56 & 0 \\ 0 & 1.46 \end{bmatrix}, \quad S = V\Lambda^{1/2} = \begin{bmatrix} -1.51 & 0.06 \\ -2.56 & 0.17 \\ -1.51 & 0.06 \\ -0.53 & -0.01 \\ 0.47 & -0.08 \\ 0.47 & -0.08 \\ 1.47 & 0.14 \\ 1.29 & -0.95 \\ 2.42 & 0.70 \end{bmatrix}$$

Figure: From http://dmml.asu.edu/cdm/slides/chapter3.pdf


Social Network Analysis
Link Analysis
Community Detection
PageRank
HITS
Node-Centric
Network-Centric
Group-Centric
Hierarchy-Centric
Topic-Sensitive PageRank
Vertex Similarity
Latent Space Model
Block Model Approximation
Spectral Clustering
Modularity maximization

Block Model Approximation

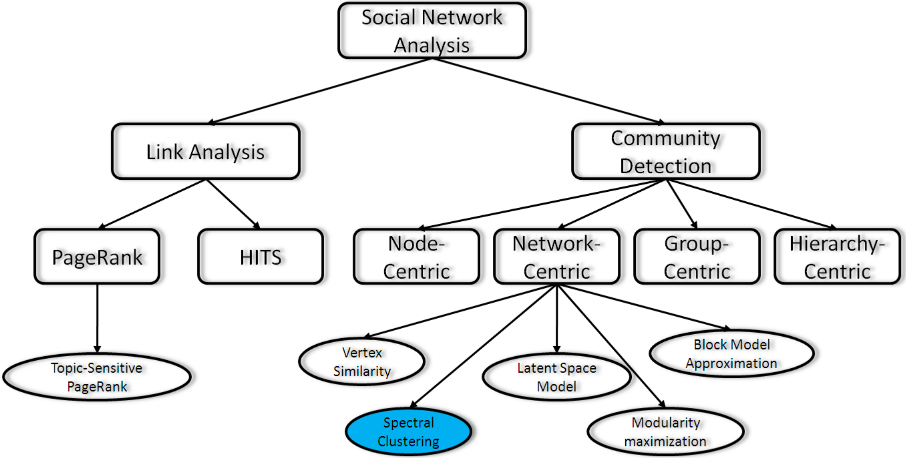
Adjacency Matrix

-	1	1	1	0	0	0	0	0
1	-	1	0	0	0	0	0	0
1	1	-	1	0	0	0	0	0
1	0	1	-	1	1	0	0	0
0	0	0	1	-	1	1	1	0
0	0	0	1	1	-	1	1	0
0	0	0	0	1	1	-	1	1
0	0	0	0	1	1	1	-	0
0	0	0	0	0	0	1	0	-

Ideal Block Structure

1	1	1	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0
0	0	0	0	1	1	1	1	1
0	0	0	0	1	1	1	1	1
0	0	0	0	1	1	1	1	1
0	0	0	0	1	1	1	1	1
0	0	0	0	1	1	1	1	1

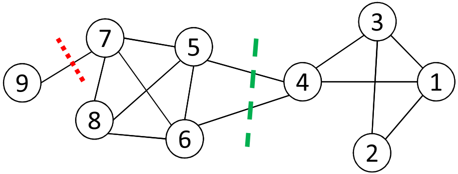
- Objective: Minimize the difference between an adjacency matrix and a block structure
$$\min_{S,\Sigma} ||A - S\Sigma S^T||_F^2$$
where $S \in \{0,1\}^{n\times k}$, and $\Sigma \in R^{k\times k}$ is diagonal
- Challenge: S is discrete, difficult to solve
- Relaxation: Allow S to be continuous satisfying $S^T S = I_k$
- Solution: the top k eigenvectors of A
- Apply k-means to S to obtain communities


Social Network Analysis
Link Analysis
Community Detection
PageRank
HITS
Node-Centric
Network-Centric
Group-Centric
Hierarchy-Centric
Topic-Sensitive PageRank
Vertex Similarity
Latent Space Model
Block Model Approximation
Spectral Clustering
Modularity maximization

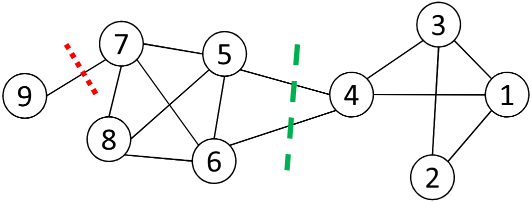
Cut

- Community detection → graph partition → minimum cut problem
- Cut: A partition of vertices of a graph into two disjoint sets
- Minimum cut: Find a graph partition such that the number of edges among different sets is minimized
  - Minimum cut often returns an imbalanced partition, e.g., node 9
  - Consider community size
    - Let C_i denote a community, $|C_i|$ represent the number of nodes in C_i, and $vol(C_i)$ measure the total degrees of nodes in C_i

$$RatioCut(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i, \bar{C}_i)}{|C_i|} \quad NormalizedCut(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i, \bar{C}_i)}{vol(C_i)}$$



Ratio Cut & Normalized Cut Example



$$RatioCut(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i,\bar{C}_i)}{|C_i|} \quad NormalizedCut(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i,\bar{C}_i)}{vol(C_i)}$$

- For partition in red (π_1)
  - $RatioCut(\pi_1) = \frac{1}{2}(\frac{1}{1} + \frac{1}{8}) = 0.56$
  - $NormalizedCut(\pi_1) = \frac{1}{2}(\frac{1}{1} + \frac{1}{27}) = 0.52$
- For partition in green (π_2)
  - $RatioCut(\pi_2) = \frac{1}{2}(\frac{2}{4} + \frac{2}{5}) = 0.45 < RatioCut(\pi_1)$
  - $NormalizedCut(\pi_2) = \frac{1}{2}(\frac{2}{12} + \frac{2}{16}) = 0.15 < NormalizedCut(\pi_1)$
- Smaller values mean more balanced partition

Spectral Clustering

- Finding the minimum ratio cut and normalized cut are NP-hard
- An approximation is spectral clustering

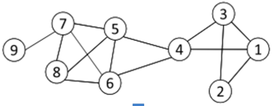
$$\min_{S \in \{0,1\}^{n \times k}} Tr(S^T \tilde{L} S) \qquad s.t., S^T S = I_k$$

- $\tilde{L}$ is the Graph Laplacian for ratio cut

$$\begin{aligned} \tilde{L} &= D - A \\ Normalized \quad \tilde{L} &= I - D^{-1/2} A D^{-1/2} \\ D &= diag\{d_1, d_2, ..., d_n\} \end{aligned}$$

- Solution: S is the eigenvector of $\tilde{L}$ (or normalized $\tilde{L}$) with smallest eigenvalues (except the first one)
- Apply k-means to S to obtain communities

Spectral Clustering Example



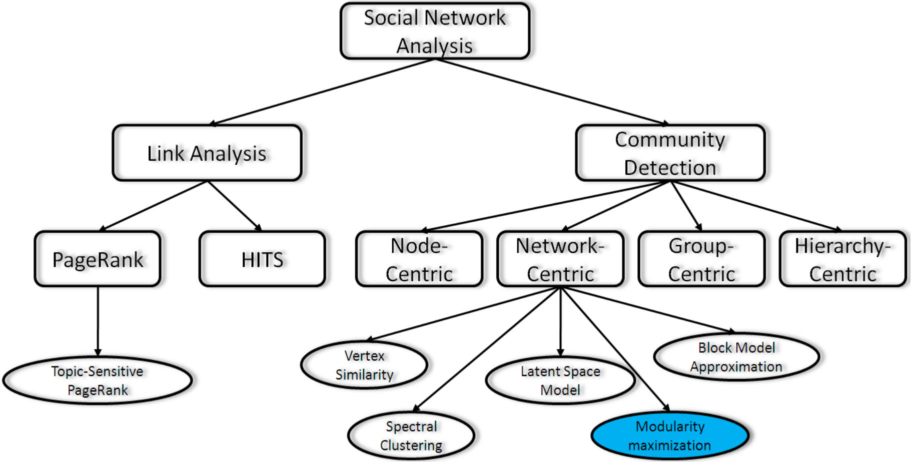
Two communities:
{1, 2, 3, 4} and {5, 6, 7, 8, 9}

k-means

$$D = diag(3, 2, 3, 4, 4, 4, 4, 3, 1)$$

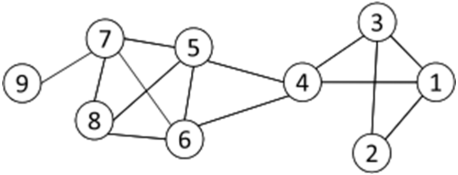
$$\widetilde{L} = D - A = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 4 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 4 & -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & -1 & 4 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 4 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix} \Rightarrow S = \begin{bmatrix} 0.33 & -0.38 \\ 0.33 & -0.48 \\ 0.33 & -0.38 \\ 0.33 & -0.12 \\ 0.33 & 0.16 \\ 0.33 & 0.16 \\ 0.33 & 0.30 \\ 0.33 & 0.24 \\ 0.33 & 0.51 \end{bmatrix}$$

Figure: From http://dmml.asu.edu/cdm/slides/chapter3.pdf


Social Network Analysis
Link Analysis
Community Detection
PageRank
HITS
Node-Centric
Network-Centric
Group-Centric
Hierarchy-Centric
Topic-Sensitive PageRank
Vertex Similarity
Latent Space Model
Block Model Approximation
Spectral Clustering
Modularity maximization

Modularity Maximization

- Modularity measures the network interactions compared with the expected random connections
- In a network with m edges, for two nodes with degree d_i and d_j, the expected random connections are $\frac{d_i d_j}{2m}$



  - The expected number of edges between nodes 1 and 2 is $3 \times 2/(2 \times 14) = 3/14$
- Strength of a community: $\displaystyle\sum_{i \in C, j \in C} (A_{ij} - d_i d_j / 2m)$
- Modularity: $Q = \dfrac{1}{2m} \displaystyle\sum_{C} \sum_{i \in C, j \in C} (A_{ij} - d_i d_j / 2m)$

Matrix Formation

- The modularity maximization can be reformed in the matrix form:

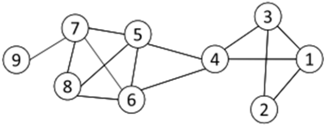
$$Q = \frac{1}{2m} Tr(S^T BS) \qquad s.t., S^T S = I_k$$

- B is the modularity matrix

$$B_{ij} = A_{ij} - d_i d_j / 2m$$

- Solution: S is the top eigenvector of the modularity matrix
- Modularity: $Q = \frac{1}{2m} \sum_C \sum_{i \in C, j \in C} (A_{ij} - d_i d_j / 2m)$
- Apply k-means to S to obtain communities

Modularity Maximization Example



$$B = \begin{bmatrix} -0.32 & 0.79 & 0.68 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.79 & -0.14 & 0.79 & -0.29 & -0.29 & -0.29 & -0.29 & -0.21 & -0.07 \\ 0.68 & 0.79 & -0.32 & 0.57 & -0.43 & -0.43 & -0.43 & -0.32 & -0.11 \\ 0.57 & -0.29 & 0.57 & -0.57 & 0.43 & 0.43 & -0.57 & -0.43 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & -0.57 & 0.43 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & 0.43 & 0.43 & -0.57 & 0.43 & 0.57 & -0.14 \\ -0.43 & -0.29 & -0.43 & -0.57 & 0.43 & 0.43 & -0.57 & 0.57 & 0.86 \\ -0.32 & -0.21 & -0.32 & -0.43 & 0.57 & 0.57 & 0.57 & -0.32 & -0.11 \\ -0.11 & -0.07 & -0.11 & -0.14 & -0.14 & -0.14 & 0.86 & -0.11 & -0.04 \end{bmatrix}$$

Modularity Matrix

$$\begin{matrix} 0.4384 & -0.2709 \\ 0.3809 & 0.2671 \\ 0.4384 & -0.2709 \\ 0.1716 & 0.6063 \\ -0.2861 & -0.3487 \\ -0.2861 & -0.3487 \\ -0.3754 & 0.3355 \\ -0.3421 & 0.1855 \\ -0.1396 & -0.1552 \end{matrix}$$

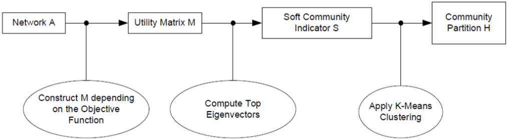
k-means

Two Communities:
{1, 2, 3, 4} and {5, 6, 7, 8, 9}

Figure: From http://dmml.asu.edu/cdm/slides/chapter3.pdf

A Unified Process

- Goal of network-centric community detection: Partition network nodes into several disjoint sets



$$\text{Utility Matrix } M = \begin{cases} \text{modified proximity matrix } \widetilde{P} & \textit{if} \text{ latent space models} \\ \text{adjacency matrix } A & \textit{if} \text{ block models} \\ \text{graph Laplacian } \widetilde{L} & \textit{if} \text{ spectral clustering} \\ \text{modularity maximization } \ B & \textit{if} \text{ modularity maximization} \end{cases}$$

- Limitation: The number of communities requires manual setting

Hierarchy-Centric Community Detection

- Goal: Build a hierarchical structure of communities based on network topology
- Facilitate the analysis at different resolutions
- Approaches:
  - Top-down: Divisive hierarchical clustering
  - Bottom-up: Agglomerative hierarchical clustering

Outline

Summary

- Goal: Discovering groups in a network where individuals' group memberships are not explicitly given
- Approaches
  - Node-centric approach
    - Each node in a group satisfies certain properties
  - Group-centric approach
    - Consider the connections inside a group as a whole
  - Network-centric approach
    - Partition nodes of a network into several disjoint sets
  - Hierarchy-centric approach
    - Build a hierarchical structure of communities based on network topology
- Which one to choose?
- Scalability issue in real applicants

Outline

1. Community Question Answering
   - Introduction
   - Question Subjectivity Analysis
   - Question Retrieval

2. Community Detection
   - Introduction
   - Methods
   - Summary

3. References

Notes

- Slides of community detection are based on http://www.cse.ust.hk/~qyang/621U/W2/621-week2.ppt, which includes more information about community evaluation
- Matlab code for community detection: http://www.cse.ust.hk/~weikep/notes/Script_community_detection.m

References

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, An Introduction to Information Retrieval (Book), 2008.
- Brin, S. and Page, L., The Anatomy of a Large-Scale Hypertextual Web Search Engine, 1998.
- Jon M. Kleinberg, Authoritative sources in a hyperlinked environment, 1999.
- Taher H. Haveliwala, Topic-sensitive PageRank, 2002.

References

- L. Tang and H. Li, Community Detection and Mining in Social Media (Book), 2010.
- H. Liu, L. Tang, and N. Agarwal, Community Detection and Behavior Study for Social Computing (Tutorial), 2009.
- R. Andersen and K. J. Lang, Communities from seed sets, WWW, 2006:
- S.Fortunato, Community detection in graphs, 2010.

References

- D. Gibson, R. Kumar, and A. Tomkins, Discovering large dense subgraphs in massive graphs, VLDB, 2005.
- M. S. Handcock, A. E. Raftery, and J. M. Tantrum, Model-based clustering for social networks, 2007.
- P. D. Hoff, A. E. Raftery, and M. S. Handcock, Latent space approaches to social network analysis, 2002.
- A. Java, A. Joshi, and T. Finin. Detecting Commmunities via Simultaneous Clustering of Graphs and Folksonomies, WebKDD, 2008.

References

- R. Kumar, J. Novak, and A. Tomkins, Structure and evolution of online social networks, KDD, 2006.
- Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Analyzing communities and their evolutions in dynamic social networks, TKDD, 2009.
- B. Long, Z.M. Zhang, X.Wu, and P. S. Yu, Spectral clustering for multi-type relational data, ICML, 2006.
- B. Long, P. S. Yu, and Z.M. Zhang, A general model for multiple view unsupervised learning, SDM, 2008.
- I. Borg and P. Groenen, Modern Multidimensional Scaling: theory and applications (2nd ed.) (Book), 2005.

References

- Chengxiang Zhai and John Lafferty, A Study of Smoothing Methods for Language Models Applied to Information Retrieval, ACM Transactions on Information Systems, 2004
- Jiwoon Jeon, W. Bruce Croft and Joon Ho Lee, Finding Semantically Similar Questions Based on Their Answers, SIGIR, 2005
- Jiwoon Jeon, W. Bruce Croft and Joon Ho Lee, Finding Similar Questions in Large Question and Answer Archives, CIKM, 2005
- Xiaobing Xue, Jiwoon Jeon and W. Bruce Croft, Retrieval Models for Question and Answer Archives, SIGIR, 2008
- Hu Wu, Yongji Wang and Xiang Cheng, Incremental Probabilistic Latent Semantic Analysis for Automatic Question Recommendation, RecSys, 2008

References

- Yunbo Cao, Huizhong Duan, Chin-Yew Lin, Yong Yu and Hsiao-Wuen Hon, Recommending Questions Using the MDL-based Tree Cut Model, WWW, 2008
- Baoli Li, Yandong Liu and Eugene Agichtein, CoCQA: Co-Training Over Questions and Answers with an Application to Predicting Question Subjectivity Orientation, EMNLP, 2008
- Delphine Bernhard and Iryna Gurevych, Combining Lexical Semantic Resources with Question & Answer Archives for Translation-Based Answer Finding, ACL, 2009
- Kai Wang, Zhaoyan Ming and Tat-Seng Chua, A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services, SIGIR, 2009
- Alon Halevy, Peter Norvig, and Fernando Pereira, The Unreasonable Effectiveness of Data, IEEE Intelligent Systems, 2009

References

- Kai Wang, Zhao-Yan Ming, Xia Hu and Tat-Seng Chua, Segmentation of Multi-Sentence Questions: Towards Effective Question Retrieval in cQA Services, SIGIR, 2010
- Xin Cao, Gao Cong, Bin Cui, Christian S. Jensen, A Generalized Framework of Exploring Category Information for Question Retrieval in Community Question Answer Archives, WWW, 2010
- Tom Chao Zhou, Chin-Yew Lin, Irwin King, Michael R. Lyu, Young-In Song and Yunbo Cao, AAAI, 2011
- Shuguang Li and Suresh Manandhar, Improving Question Recommendation by Exploiting Information Need, ACL, 2011
- Tom Chao Zhou, Xiance Si, Edward Y. Chang, Irwin King and Michael R. Lyu, A Data-Driven Approach to Question Subjectivity Identification in Community Question Answering, AAAI, 2012

References

- Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, Hakan Ferhatosmanoglu, A Large-Scale Sentiment Analysis for Yahoo! Answers, WSDM, 2012
- Xingliang Ni, Yao Lu, Xiaojun Quan, Liu Wenyin, Bei Hua, User interest modeling and its application for question recommendation in user-interactive question answering systems, Information Processing and Management, 2012
- Anna Shtok, Gideon Dror and Yoelle Maarek, Learning from the Past: Answering New Questions with Past Answers, WWW, 2012
- Xiance Si, Edward Y. Chang, Zoltan Gyongyi and Maosong Sun, Confucius and Its Intelligent Disciples: Integrating Social with Search, VLDB, 2010
- Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer and Paul S. Roosin, A Statistical Approach to Machine Translation, Computational Linguistics, 1990

QA

Thanks for your attention!