IBM Research

Automatic Modeling of Human Behavior and Social Network

Ching-Yung Lin

IBM T. J. Watson Research Center

Univ. of Washington / Columbia Univ.

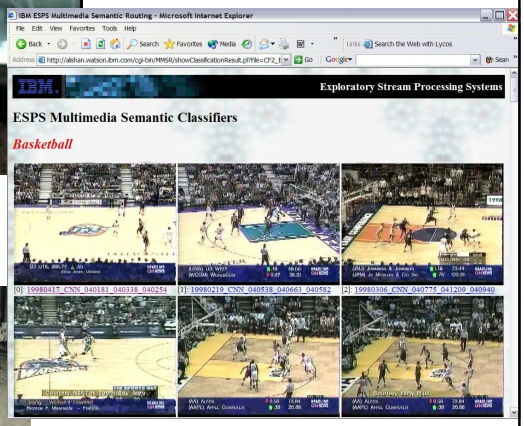
December 28, 2005

Outline

- **Motivation**
- **Social Network Analysis and Modeling**
  - **-- Who do you know?**
- **Expertise Modeling**
  - **-- What do you know?**
- **Personal / Community Interest Modeling**
  - **-- How do you like it?**
- **Sleep Quality Inference**
  - **-- What did you do?**
- **Smart Wearable Audio-Visual-Location Sensors**
  - **-- Where, When, What and Who did you see?**
- **Conclusion**

From multimedia understanding to multimodality understanding

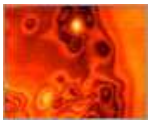
A picture is worth 1000 words
– which one thousand?



NIST TREC Video Concept Retrieval Benchmarking

Human – a complex multimodality subject/object

- "*Human and Social Dynamics (HSD)*" is identified as one of the five NSF key priorities among:
  - Nanoscale Science and Engineering
  - Biocomplexity in the Environment
  - Human and Social Dynamics
  - Mathematical Sciences
  - Cyberinfrastructure
  - (http://www.nsf.gov/news/priority_areas/)



Person, Community, Society

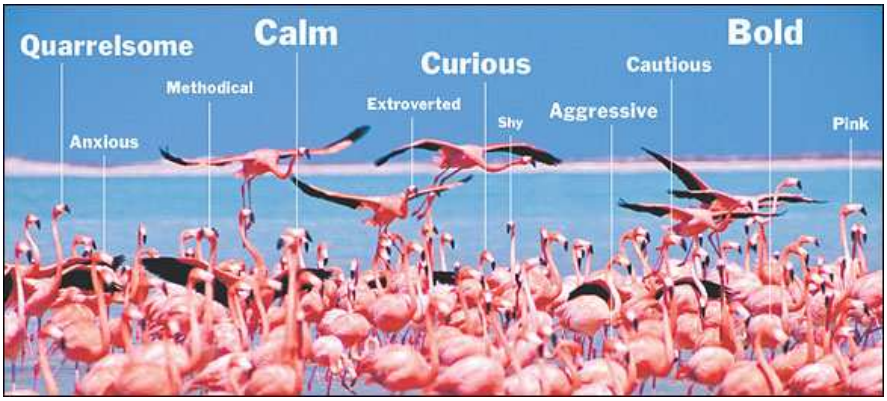


Photo Source: New York Times, 3/2/2005

Social Computing – when computer science meets sociology

Social Computing – when computer science meets sociology

- Computing-based Human Modeling (person)
- Computing-based Society Modeling (person <-> person)
- Computing-based Trust Management (person -> society, person -> person, person -> information)
- Computing-based Information Organization and Management (information -. person, information -> society)

Social Computing – when computer science meets sociology

- Computing-based Human Modeling (person)
  - Biometric-based modeling
  - Behavior modeling
  - Knowledge and Interest modeling
- Computing-based Society Modeling (person <-> person)
  - Community modeling
  - Social network modeling
- Computing-based Trust Management (person -> society, person -> person, person -> information)
  - Information trustworthiness
  - Human trustworthiness
  - System trustworthiness
- Computing-based Information Organization and Management (information -. person, information -> society)
  - Personalized information
  - Community-oriented information

Social Computing (I) -- Computing-based Human Modeling

- Biometric-based modeling:
  - Face Recognition
  - Speaker Recognition
  - Fingerprint Recognition
  - Iris Recognition
  - Hand, Dental Recognition
  - DNA Recognition
- Behavior modeling:
  - Painting and Speaking Style Authentication
  - Walking and Typing Features Authentication
  - Emailing Behavior Modeling
  - Personal Information Propagation Behavior Modeling
- Knowledge and Interest modeling:
  - Expertise Modeling
  - Interest Modeling
  - Personal Social Network Modeling

Social Computing (II) -- Computing-based Society Modeling

- Community modeling:
  - Communication-based Community Identification
  - Link-based Community Identification (e.g., blogs, personal webpages, citations)
  - Access-based Community Identification (e.g., e-commerce sites, digital archive, organization database)
  - Opinion-based Community Identification (e.g., collaborative filtering)
- Social Network modeling:
  - Informal Network in Organization
  - Information Propagation Network
  - Epidemic Network
  - Friendship Network

Social Computing (III) -- Computing-based Trust Management

- Information Trustworthiness:
  - Multimedia Authentication
  - Data Security in Communication
- Human Trustworthiness:
  - Collaborator Identification
  - Opinion Acceptance
- System Trustworthiness:
  - System Reliability
  - Fault-Tolerance System

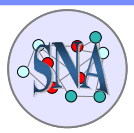
Social Computing (IV) -- Computing-based Information Organization and Management

- Personalized Information:
  - Personalized Multimedia Summarization
  - Personalized Data Abstraction
  - Personalized Data Representation and Visualization
- Community-Oriented Information:
  - Community-Profiling for Multimedia Summarization
  - Community-Profiling for Data Abstraction
  - Community-Profiling for Data Representation and Visualization

Outline

- Motivation
- Social Network Analysis and Modeling
- Expertise Modeling
- Personal / Community Interest Modeling
- Sleep Quality Inference
- Conclusion

What is Social Network Analysis?

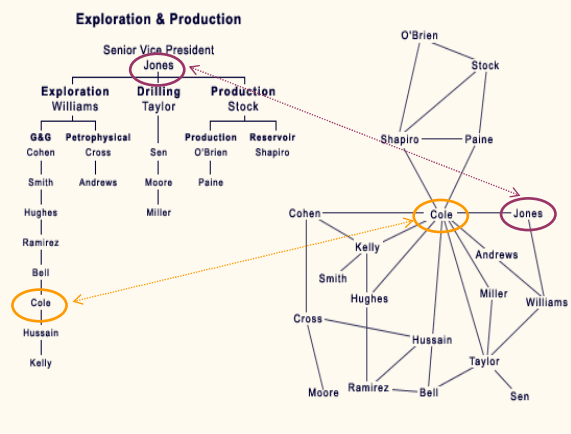


- Social Network Analysis (SNA) is a set of methods and tools for revealing relations between entities – communities, people (if they are strong influential stakeholders), teams, departments, organizations and even countries.
- Social networks assumes interdependence between people.
- Behaviors and outcomes are understood through our relationship with others
  - Academic roots in sociology, anthropology, organizational behavior and medicine
  - Recent application to problems in Knowledge Management and Collaboration
  - Also called Organizational Network Analysis

Provided by Drs. Tony Mobbs and Kate Ehrlich, IBM

Beyond the organizational chart

- Organization charts are not the best indicator of how work gets done
- Senior people are not always central; peripheral people can represent untapped knowledge
- Making the network visible makes it actionable and becomes the basis for a collaboration action plan



Source: Cross, R., Parker, A., Prusak, L. & Borgatti, S.P. 2001. Knowing What We Know: Supporting Knowledge Creation and Sharing in Social Networks. Organizational Dynamics 30(2): 100-120. [pdf]

Provided by Drs. Tony Mobbs and Kate Ehrlich, IBM

Relationships are multi-dimensional and uncovered through network questions

Actions

Communication
How often do you communicate with this person?

Innovation
How often do you turn to this person for new ideas

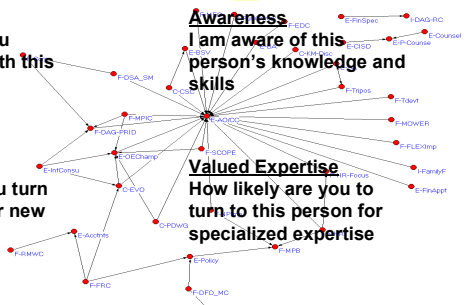
Advice
How often do you seek advice from this person before making an important decision?

Awareness

Awareness
I am aware of this person's knowledge and skills

Valued Expertise
How likely are you to turn to this person for specialized expertise

Learning
How likely are you to rely on this person for advice on new methods and processes



Emotional

Trust
I believe there is a high personal cost in seeking advice or support from this person

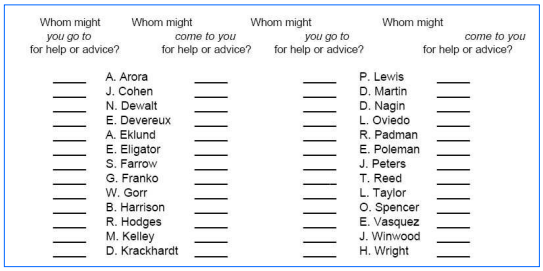
Access
I believe this person will respond to my request in a reasonable and timely manner

Energy
I generally feel energized when I interact with this person

Provided by Drs. Tony Mobbs and Kate Ehrlich, IBM

What can computer scientists do?

- Informal social network within formal organizations is a major factor affecting a company/society's performance.
  - Krackhardt (2005) showed that companies with strong informal networks perform five or six times better than those with weak networks.
  - Since Weber (1920s), decades of related social scientific researches have been mainly relying on questionnaires and interviews to understand individual's thoughts and behaviors.



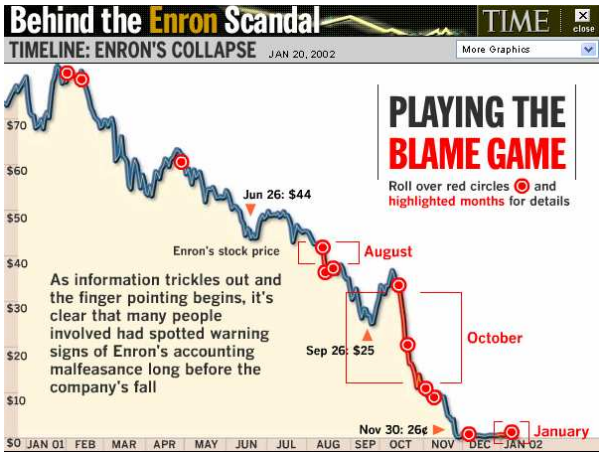
Whom might *you go to* for help or advice?		Whom might *come to you* for help or advice?		Whom might *you go to* for help or advice?		Whom might *come to you* for help or advice?	
____	A. Arora	____		____	P. Lewis	____	
____	J. Cohen	____		____	D. Martin	____	
____	N. Dewalt	____		____	D. Nagin	____	
____	E. Devereux	____		____	L. Oviedo	____	
____	A. Eklund	____		____	R. Padman	____	
____	E. Eligator	____		____	E. Poleman	____	
____	S. Farrow	____		____	J. Peters	____	
____	G. Franko	____		____	T. Reed	____	
____	W. Gorr	____		____	L. Taylor	____	
____	B. Harrison	____		____	O. Spencer	____	
____	R. Hodges	____		____	E. Vasquez	____	
____	M. Kelley	____		____	J. Winwood	____	
____	D. Krackhardt	____		____	H. Wright	____	

Sample questionnaire
(Prof. Krackhardt, CMU)

➔ Is it possible to 'acquire' social networks automatically?

➔ How about automatically building/updating 'personal profiles', 'social capitals'?

Case Study -- Enron Corpus



Enron Corpus

- Preprocessing
  - Original messages – 517,431
  - Remove empty messages – 493,391 remain
    - 1999 – 11196
    - 2000 – 196157
    - 2001 – 272875
    - 2002 – 35922
  - Remove repeated messages – 166,653 remain
  - Only keep intra-communications among 149 users within Enron – 25,428 remain
  - Number of terms: 84649
  - Number of users: 149

Name	Email	Position
Robert Badeer	robert.badeer@enron.com	Director
Eric Bass	eric.bass@enron.com	Trader
Sally Beck	sally.beck@enron.com	Employee
Rick Buy	rick.buy@enron.com	Manager
David Delainey	david.delainey@enron.com	CEO
James Derrick	james.derrick@enron.com	In House Lawyer
Mark Haedicke	mark.haedicke@enron.com	Managing Director
Steven Kean	steven.kean@enron.com	Vice President
Louise Kitchen	louise.kitchen@enron.com	President
Phillip Allen	phillip.allen@enron.com	N/A

Collected information about the emails

ID	Subject	Time	From	To
31265382.1075858640461	FW: California gas intrastate matte	2001-07-10T19:32:29.	k..allen@enron.com	matt.smith@enron.com,
14873812.1075858640483	FW: West Power Strategy Briefing	2001-07-11T12:56:41.	k..allen@enron.com	keith.holst@enron.com, mike
3650242.1075858640506.		2001-07-11T15:25:40.	k..allen@enron.com	barry.tycholiz@enron.com,
483924.1075858640549.J	FW: Party	2001-07-12T12:04:29.	k..allen@enron.com	s..shively@enron.com,
13141541.1075858640571		2001-07-12T19:55:20.	k..allen@enron.com	michael.l.brunner@rssmb.com
14620083.1075858640594	CA Instrate Gas matters	2001-07-13T13:44:25.	k..allen@enron.com	leslie.lawner@enron.com,
634023.1075858640616.J	FW: CA Instrate Gas matters	2001-07-13T13:45:39.	k..allen@enron.com	mike.grigsby@enron.com,
19282752.1075858640638	Analyst/Associate Program: 2 Minu	2001-07-13T19:47:41.	k..allen@enron.com	ramabile@execlead.com,
16515849.1075858640659	FW: American Express Letter	2001-07-16T13:20:48.	k..allen@enron.com	johnny.ross@enron.com,
19618808.1075858640681	Party	2001-07-16T13:22:52.	k..allen@enron.com	richard.toubia@truequote.com
27461841.1075858640705	FW: Party	2001-07-16T13:44:37.	k..allen@enron.com	s..shively@enron.com,

What happened?– collect the ground truth

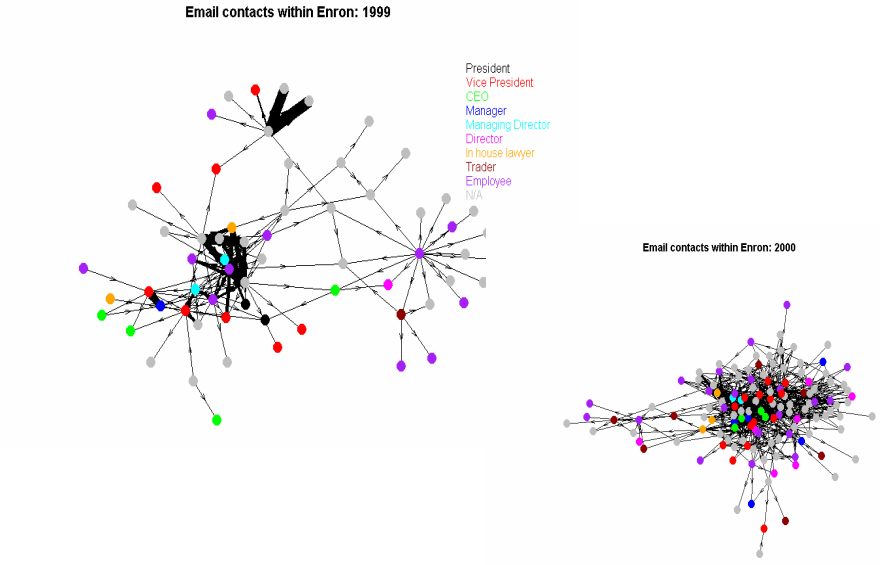
- Summarize important events from different timelines
- The events with most occurrences from multiple media's timelines
  - 14 August 2001 -- Jeffrey Skilling resigns after just six months; Mr Lay returns to day-to-day management of the company.
  - 20 August 2001 -- Mr Lay exercises Enron share options worth $519,000.
  - 12 October 2001 -- Accounting firm Andersen begins destroying documents relating to the Enron audits.The destruction continues until November when the company receives a subpoena from the Securities and Exchange Commission.
  - 16 October 2001-- Enron reports losses of $638m run up between July and September and announces a $1.2 billion reduction in shareholder equity. The reduction in company value relates to partnerships set up and run by chief financial officer Andrew Fastow.
  - …

Social Network Analysis:

- Potential of social network analysis
  - every person in the world is only six edges away from every other, if an edge between *i* and *j* means "*i* knows *j*" [Milgram 1967]
- Static social network analysis
  - In social network analysis: Exponential Random Graph Models [Wasserman and Pattison, 1996]
  - In information mining area:
  - Mine social relationships from email logs by using a set of heuristic graph algorithms [Schwartz and Wood 1993]
  - Mine a social network from a wide variety of publicly-available online information to help individuals find experts who could answer their questions [Kautz *et al.* 1997]
  - *M*ine communities from the Web (defined as sets of sites that have more links to each other than to non-members) [Flake *et al. 2002*]
  - Use a betweeness centrality algorithm for the automatic identification of communities of practice from email logs within an organization [Joshua T. *et al. 2003]*
  - The Google search engine and HITS algorithm [Page *et al.* et al. 1998] [Kleinberg 1998]
- Dynamical social networks
  - In social network analysis: Dynamic actor-oriented social network [Snijder 2002]
  - Changes in the network are modeled as the stochastic result of network effects (density, reciprocity, etc.)
  - Network evolution is modeled by continuous time Markov chain models
  - In information mining area:
  - link prediction problem - Infer which new interactions among its members are likely to occur in the near future [Liben-Nowell 2003]
  - Track changes in large-scale data by periodically creating an agglomerative clustering and examining the evolution of clusters over time [Kubica *et al. 2002*]

➔ However, all of them are only based on pure network properties, without knowing what people are talking about and why they have close relationship

Dynamic social networks



Using Traditional SNA -- People with top 10 centralities in Enron

Centrality: Actor has high involvement in many relations, regardless of send/receive directionality (volume of activity)

Centrality	1999		2000		2001		2002	
	Name	Position	Name	Position	Name	Position	Name	Position
1	Mark_Taylor	Employee	David_Delainey	CEO	Steven_Kean	Vice_President	Kevin_Presto	Vice_President
2	Tana_Jones	N/A	Steven_Kean	Vice_President	John_Lavorato	CEO	Louise_Kitchen	President
3	Sara_Shackleton	N/A	John_Lavorato	CEO	Jeff_Dasovich	Employee	John_Lavorato	CEO
4	Richard_Sanders	Vice_President	Vince_Kaminski	Manager	Vince_Kaminski	Manager	Hunter_Shively	Vice_President
5	Elizabeth_Sager	Employee	Jeff_Skilling	CEO	Louise_Kitchen	President	James_Steffes	Vice_President
6	Mark_Haedicke	Managing_Director	Mike_McConnell	N/A	David_Delainey	CEO	Greg_Whalley	President
7	John_Hodge	Managing_Director	Greg_Whalley	President	Greg_Whalley	President	Fletcher_Sturm	Vice_President
8	Steven_Kean	Vice_President	Sally_Beck	Employee	Mark_Haedicke	Managing_Director	Doug_Gilbert-Smith	N/A
9	Dan_Hyvl	Employee	Jeffrey_A_Shankman	N/A	Phillip_Allen	N/A	Dana_Davis	N/A
10	Carol_Clair	In_house_lawyer	John_Arnold	Vice_President	Mary_Hain	In_house_lawyer	Mark_Haedicke	Managing_Director

Using Traditional SNA -- People with top-10 prestige in Enron

Prestige: Actor is the recipient of many directed ties

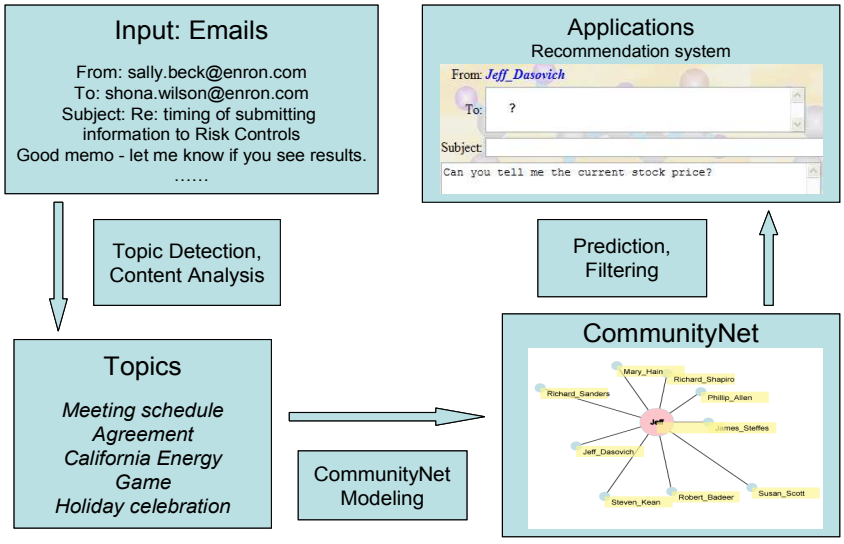
Prestige	1999		2000		2001		2002	
	Name	Position	Name	Position	Name	Position	Name	Position
1	John_Hodge	Managing_Director	John_Lavorato	CEO	John_Lavorato	CEO	Darron_Giron	N/A
2	Steven_Kean	Vice_President	Greg_Whalley	President	Louise_Kitchen	President	Phillip_Love	N/A
3	Vince_Kaminski	Manager	David_Delainey	CEO	Phillip_Allen	N/A	Kam_Keiser	Employee
4	Mark_Haedicke	Managing_Director	Steven_Kean	Vice_President	Greg_Whalley	President	Errol_McLaughlin	N/A
5	Elizabeth_Sager	Employee	Vince_Kaminski	Manager	Kevin_Presto	Vice_President	Stacey_White	N/A
6	Richard_Sanders	Vice_President	Rick_Buy	Manager	Barry_Tycholiz	Vice_President	Fletcher_Sturm	Vice_President
7	Kevin_Presto	Vice_President	Kevin_Presto	Vice_President	Steven_Kean	Vice_President	NA	NA
8	Mark_Taylor	Employee	Jeffrey_A_Shankman	N/A	Mike_Grigsby	N/A	NA	NA
9	Michelle_Cash	N/A	Phillip_Allen	N/A	David_Delainey	CEO	NA	NA
10	Stacy_Dickson	Employee	Jeff_Skilling	CEO	Hunter_Shively	Vice_President	NA	NA

Most of them have relatively high position in Enron, which reveal the roles in the social network actually are almost corresponding to the roles in the real life

Our Contributions

- A novel way to automatically model and predict human behavior of receiving and disseminating information.
- Generate an application model (CommunityNet) which describes personal dynamic community network.
- Develop a new algorithm incorporating content, time and social networks simultaneously.
- Experiments results show that personal behavior and intention are somewhat predictable – e.g., to whom a person is going to send a specific mail.
- The performance of the proposed adaptive algorithm is 58% better than the model only based on social network, and is 75% better than an aggregated model based on the state-of-the-art content analysis model with social network enhancement.
- Developed prototypes showing how this model can be applied to organization management and social capital management

An Overview of CommunityNet



Content Analysis Prior Art I -- Latent Semantic Analysis

- Latent Semantic Analysis (LSA) [Landauer, Dumais 1997]
  - Descriptions:
    - Capture the semantic concepts of documents by mapping words into the latent semantic space which captures the possible synonym and polysemy of words
    - Training based on different level of documents. Experiments show the synergy of the # of training documents and the psychological studies of students at 4th, 10th, and college level. Used as an alternative to TOEFL test.
  - Based on truncated SVD of document-term matrix: optimal least-square projection to reduce dimensionality
  - Capture the concepts instead of words
    - Synonym
    - Polysemy

LSA

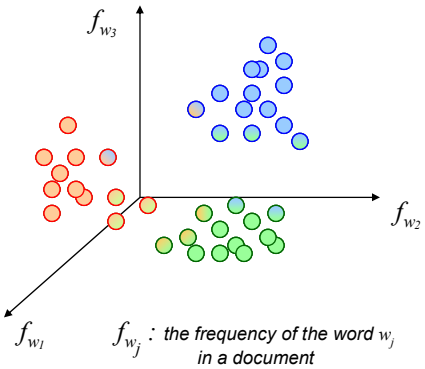
documents

terms

$$X \cong T_0 \cdot S_0 \cdot D_0'$$

$N \times M \quad N \times K \quad K \times K \quad K \times M$

Traditional Content Clustering



f_{w_j} : *the frequency of the word* w_j *in a document*

Clustering:

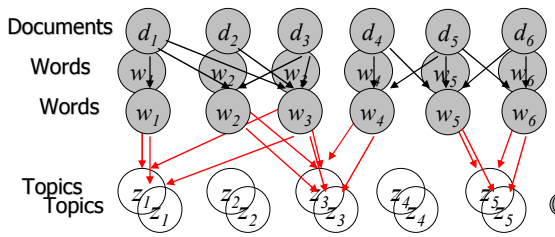
Partition the feature space into *segments* based on training documents. Each segment represents a topic / category. (← Topic Detection)

Hard clustering: e.g., K-mean clustering

$$d = \{f_{w_1}, f_{w_2}, ..., f_{w_N}\} \rightarrow z$$

Soft clustering: e.g., Fuzzy C-mean clustering

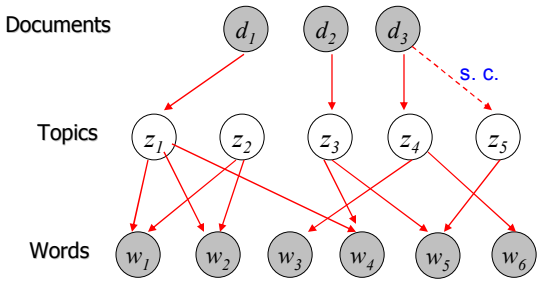
$$P(Z \mid \mathbf{W} = \mathbf{f}_w)$$



Another representation of clustering (we showing the deterministic part)

: observations

Content Clustering based on Bayesian Network



$h(D)$ $P(Z \mid D)$

hard clustering soft clustering

$P(W \mid Z)$

: observations

Bayesian Network:

- Causality Network – models the causal relationship of attributes / nodes
- Allows hidden / latent nodes

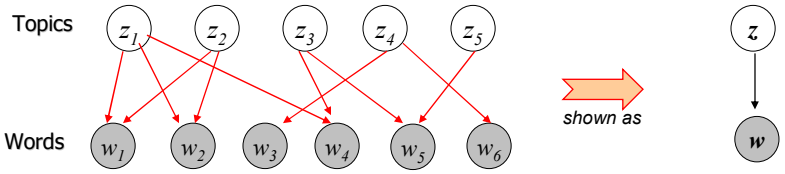
Hard clustering:

$$h(D = d) = \arg\max_{z} P(\mathbf{W} = \mathbf{f}_w \mid Z) \quad \text{<= MLE}$$

$$P(W \mid Z) = \frac{P(Z \mid W)P(W)}{P(Z)} \quad \text{<= Bayes Theorem}$$

29

Content Clustering based on Bayesian Network – Hard Clustering



: observations

$$P(W \mid Z) = \frac{P(Z \mid W)P(W)}{P(Z)} = \frac{P(Z \mid W)P(W)}{\int P(Z \mid W)dW}$$

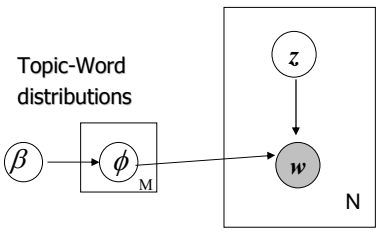
N: the number of words
(The number of topics (M) are pre-determined)

Major Solution 1 -- Dirichlet Process:

- Models P(W | Z) as mixtures of Dirichlet probabilities
- Before training, the prior of P(W|Z) can be a easy Dirichlet (uniform distribution). After training, P(W|Z) will still be Dirichlet. (← *The reason of using Dirichlet*)

Major Solution 2 -- Gibbs Sampling:

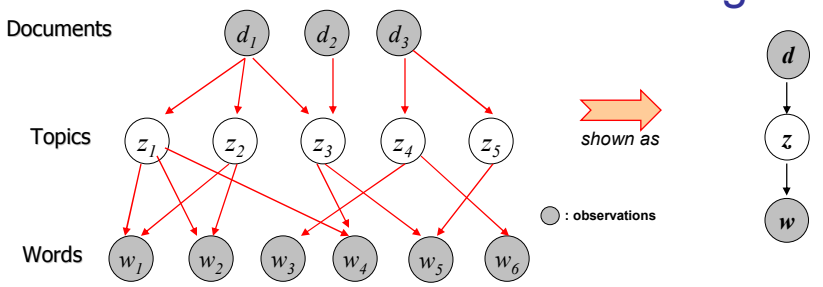
- A Markov chain Monte Carlo (MCMC) method for integration of large samples → calculate P(Z)

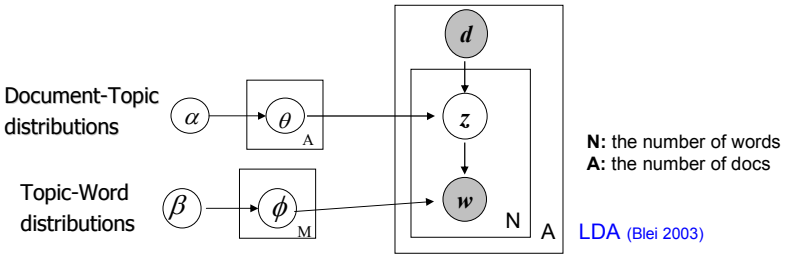


Latent Dirichlet Allocation (LDA) (Blei 2003)

30

Content Clustering based on Bayesian Network – Soft Clustering





N: the number of words
A: the number of docs

LDA (Blei 2003)

31

Some Insight on BN-based Content Clustering

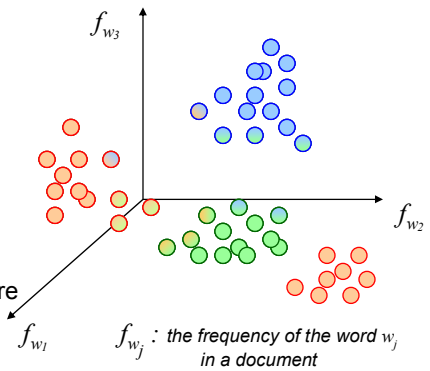
Bayesian Network:

• Models the *practical* causal relationships..

Content Clustering:

• Because documents and words are dependent,

➔ only close documents in the feature space can be clustered together as one topic.



⇒**Incorporating human factors can possibly *link* multiple clusters together.**

32

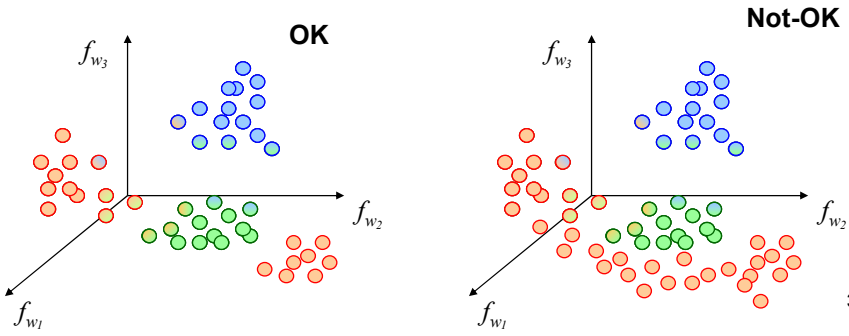
Why We Need Simultaneous Multimodality Clustering?

Multiple-Step Clustering:

• e.g., Naïve way to combine content filtering and collaborative filtering

➔ Independently cluster first. Combine later.

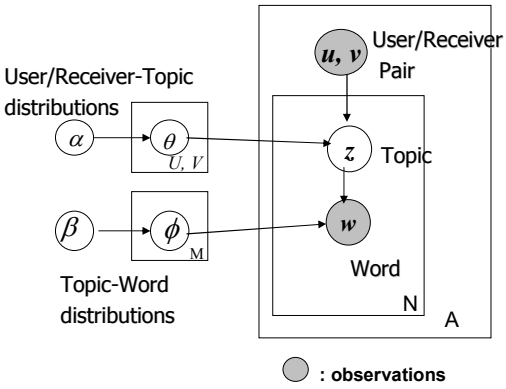
➔ ***Simultaneous Multimodality Clustering is important.***



33

How to Incorporate Human Factors?

2-Stage Bayesian Network Modeling for Emails [Song et. al. KDD 2005]

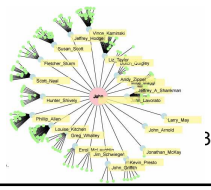


CTR Modeling :

• Objective: Build 2-stage Bayesian Networks to represent the causal relations of user, receiver, topic, and words of emails.

• Technical Achievement: Show that multiple Gibbs Sampling can be applied to multistage BNs, if nodes are only 1-stage dependent.

• Applications: Email Topic/Thread Classification and Information Flow Predication



z : latent topics
θ: User/Receiver-Topic distributions
ϕ: Topic-Word distributions
α, β: Dirichlet parameters

A: the number of emails
M: the number of latent topics
N: the number of words in a document
U, V: the number of sender / receiver

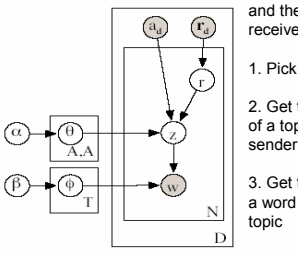
34

Novel Content-Time-Relation Algorithm -- II

- Content-Time-Relation (CTR) [Song, Lin, Tseng, Sun, KDD-submission Feb. 2005]:
  - Incremental Latent Dirichlet Allocations
  - Capture evolutionary information
  - Integrate social network model

  ➔*Combine content, time and social relation information with Dirichlet allocations, a causal Bayesian network and an Exponential Random Graph Social Network Model.*
  - *Besides, for the time windowing, one can use Poisson distribution to replace the Dirichlet allocation, where* $\varphi = \gamma^{\wedge} | t - t_0 |$.

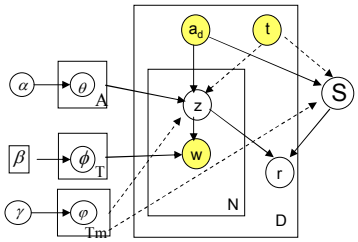
Author-Recipient-Topic Model
(ART)
[McCallum et. al, 2005]



Given the sender and the set of receivers of an email:

1. Pick a receiver
2. Get the probability of a topic given the sender and receiver
3. Get the probability a word given the topic

CTR model

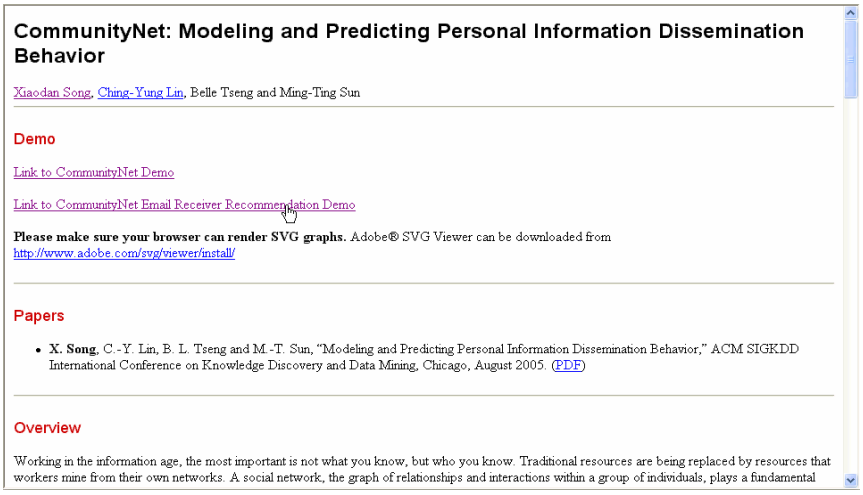


: observations

Given the sender and the time of an email:

1. Get the probability of a topic given the sender
2. Get the probability of the receiver given the sender and the topic
3. Get the probability of a word given the topic

Demo – The Email Receiver Recommendation System



url: http://nansen.ee.washington.edu/communitynet

Topic Analysis Results - Hot and cold topics in Enron Email Corpus

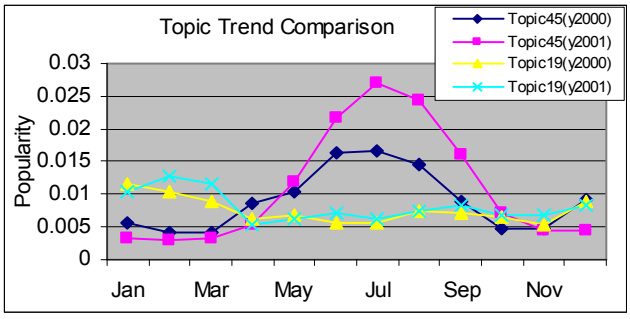
Table 1. Hot Topics

Meeting	Deal	Petroleum	Texas	Document
meeting	deal	Petroleum	Houston	letter
plan	desk	research	Texas	draft
conference	book	dear	Enron	attach
balance	bill	photo	north	comment
presentation	group	Enron	America	review
discussion	explore	station	street	mark

Table 2. Cold Topics

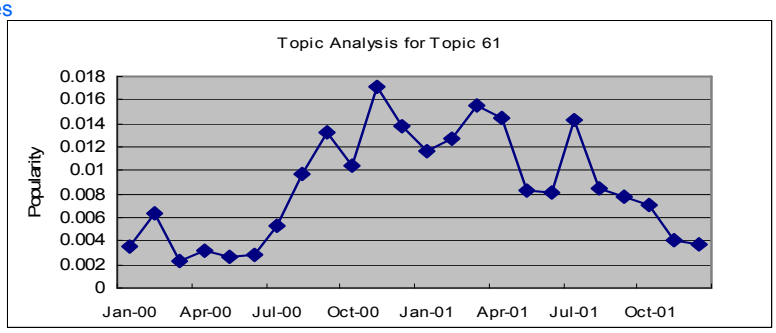
Trade	Stock	Network	Project	Market
trade	Stock	network	Court	call
London	earn	world	state	market
bank	company	user	India	week
name	share	save	server	trade
Mexico	price	secure	project	description
conserve	new	system	govern	respond

Corporate Topic Trend Analysis Example: Yearly repeating events



Topic 45, which is talking about a schedule issue, reaches a peak during June to September. For topic 19, it is talking about a meeting issue. The trend repeats year to year.

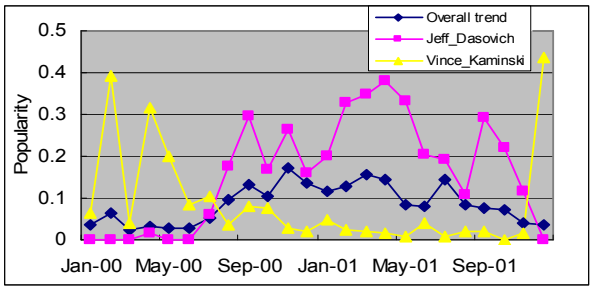
Topic Detection and Key People Detection of "California Power" Match Their Real-Life Roles



Key Words	power 0.089361 California 0.088160 electrical 0.087345 price 0.055940 energy 0.048817 generator 0.035345 market 0.033314 until 0.030681
Key People	Jeff_Dasovich 0.249863 James_Steffes 0.139212 Richard_Shapiro 0.096179 Mary_Hain 0.078131 Richard_Sanders 0.052866 Steven_Kean 0.044745 Vince_Kaminski 0.035953

Event "California Energy Crisis" occurred at exactly this time period. Key people are active in this event except Vince_Kaminski …

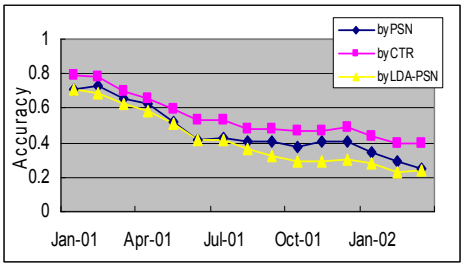
Topic Trends of Key Person's Communication Records Match the Trends in the Topic-Community



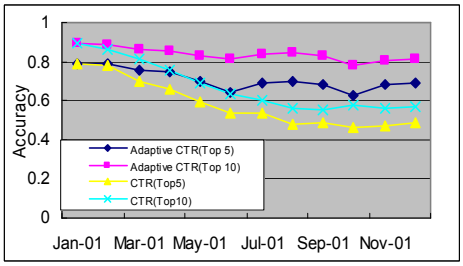
Name	Corr. Coef.	Position
Jeff_Dasovich	0.7965	*Enron government relations executive*
James_Steffes	0.6501	*Vice President of Government Affairs*
Richard_Shapiro	0.5604	*Vice President of Regulatory Affairs*
Mary_Hain	0.5994	*lawyer at Enron's West Coast trading hub.*
Richard_Sanders	0.2745	*Vice President and Assistant General Counsel*
Steven_Kean	0.3585	*Executive Vice President and Chief of Staff*
Vince_Kaminski	-0.4617	*economist and Head of Research for Enron Corp*

All of the key people were sort of politicians except Vince Kaminski is a researcher.

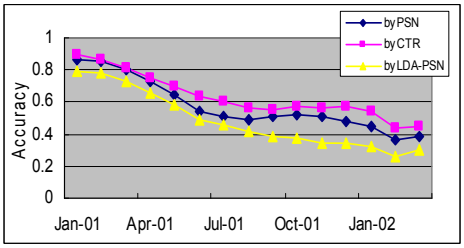
Receivers are Predictable Given the Content and the Composer of an Email



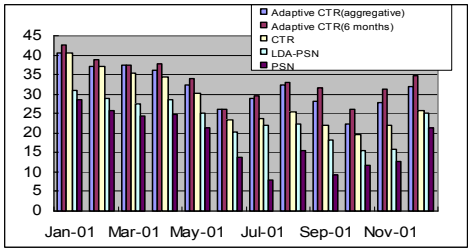
(a) Accuracy based on the top 5 most likely people



(c). Comparison between Adaptive CTR and CTR models



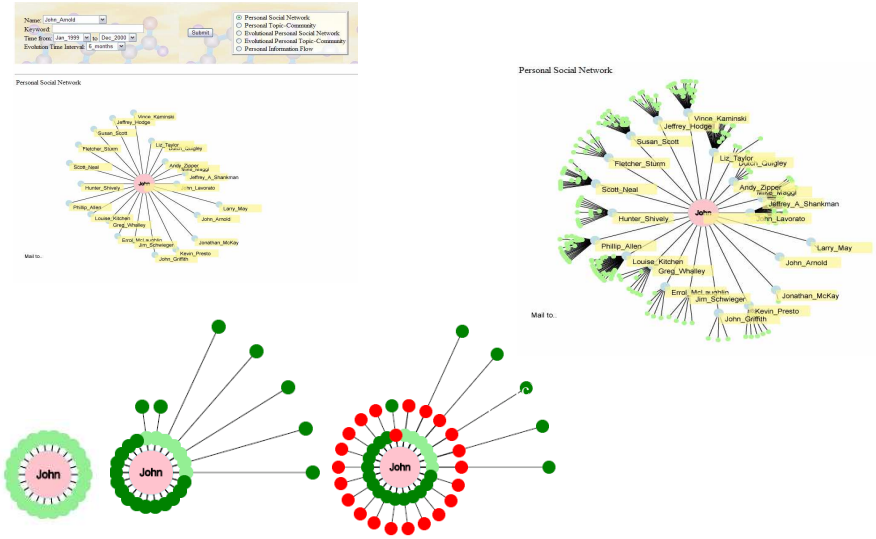
(b) Accuracy based on the top 10 most likely people



(d) Comparison using Breese evaluation metrics

Accuracy is measured by testing whether the "real" receiver is among the prediction list of the top 5 or 10 most likely people

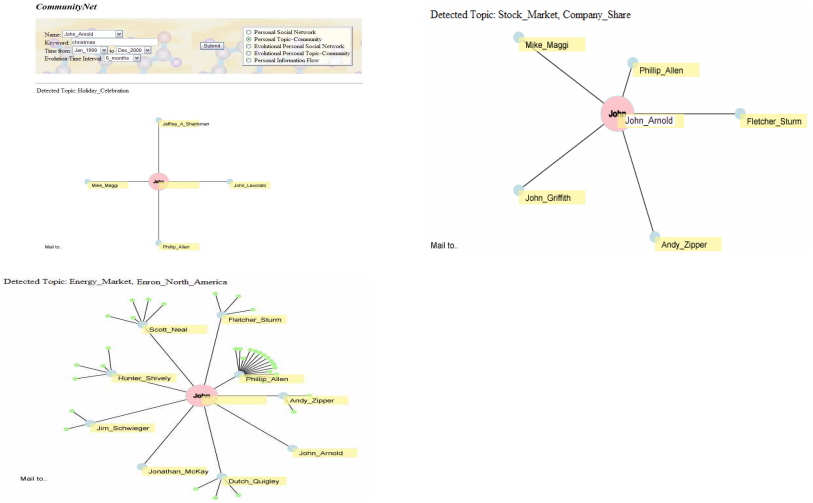
Personal Social Network - who a user contacts with during a time period



(b1) Jan-'99 to Dec-'99 (b2) Jan-'00 to Jun-'00 (b3) Jul-'00 to Dec-'00

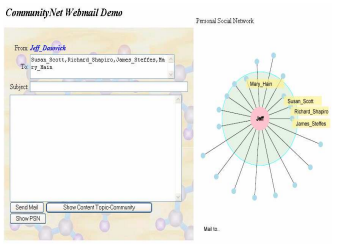
(b) Evolutionary Personal Social Network

Personal Topic-Community Networks- who a user will contact with under different topics

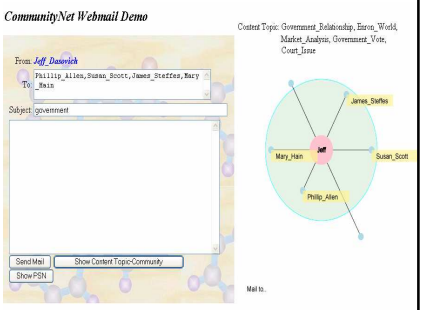


(b) Personal Topic-Community Information flow for keyword "energy" with two-level personal Topic-Community network

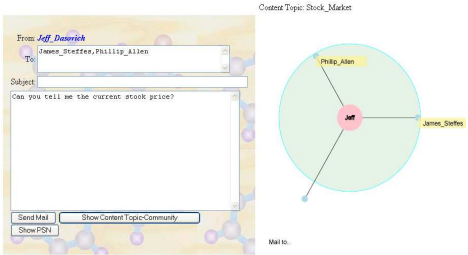
Personal Social Capital Management -Receiver Recommendation Demo



(a) Receiver recommendation based on Personal Social Network



(b) Receiver recommendation for "Government"

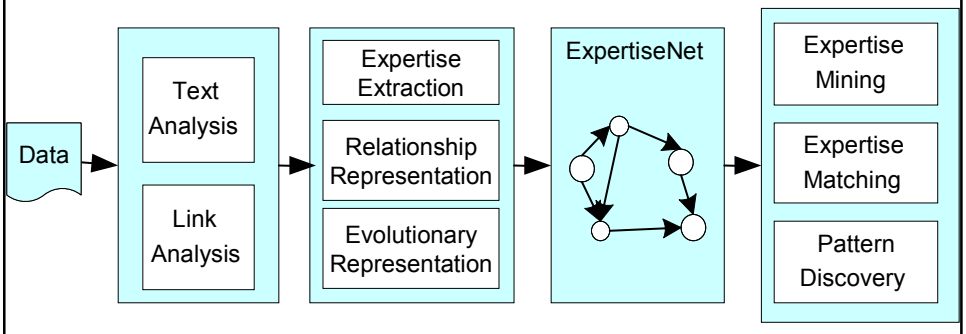


(c) Receiver recommendation for "Can you tell me the current stock price?"

Outline

- Motivation
- Social Network Analysis and Modeling
- Expertise Modeling
- Personal / Community Interest Modeling
- Sleep Quality Inference
- Conclusion

ExpertiseNet



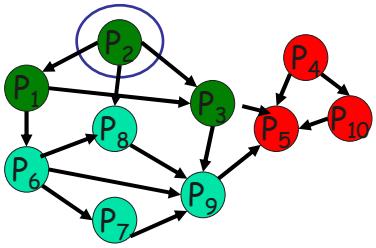
Build graphical representations of a researcher's expertise based on her/his publications and citation networks

46

Expertise Extraction

- Adaboost based algorithm
  - Weighted combination of weak learners
  - Not prone to overfitting
- Features
  - Bag of words with tf as weight
    - Title
    - Abstract
    - Title of the reference
  - Expertise categories from references and the associated ratio (ref. prob.)

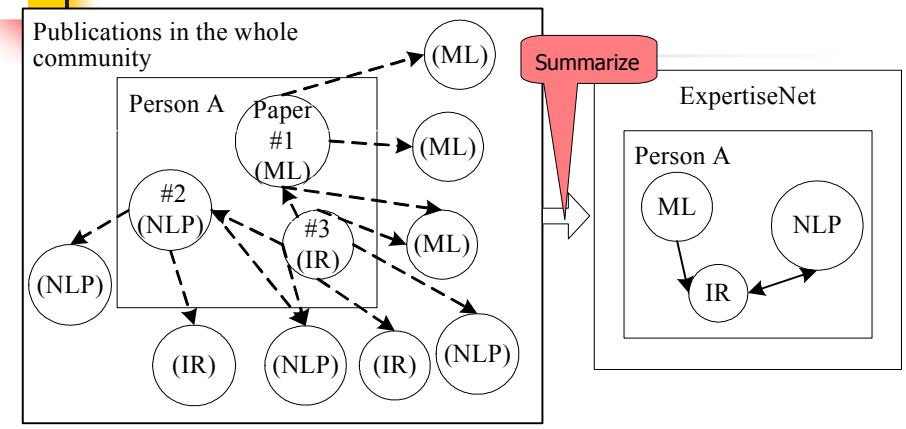
Citation network



expertise category set { }

47

Building Relational ExpertiseNet

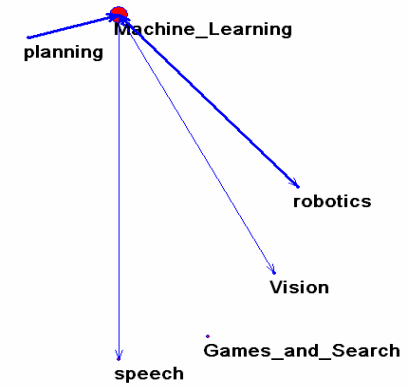


➔ *Capture one's research expertise + how one expertise influences and is influenced by others*

48

Relational ExpertiseNet

- Graph G=(V,E)
  - V = set of nodes
  - E = set of edges
- Node: Expertise
  - The strength of the expertise
- Edge: Relationship between two expertise
  - The strength of the correlation between two expertise



An example of Relational ExpertiseNet

49

Model the ExpertiseNet

- Capture expertise as well as the structure influences
  - The dependency relationship
  - The hierarchical structure

→ ERGM

$$P_\theta(Y = y) = \frac{\exp\left(\theta^T s(y)\right)}{c(\theta)}$$

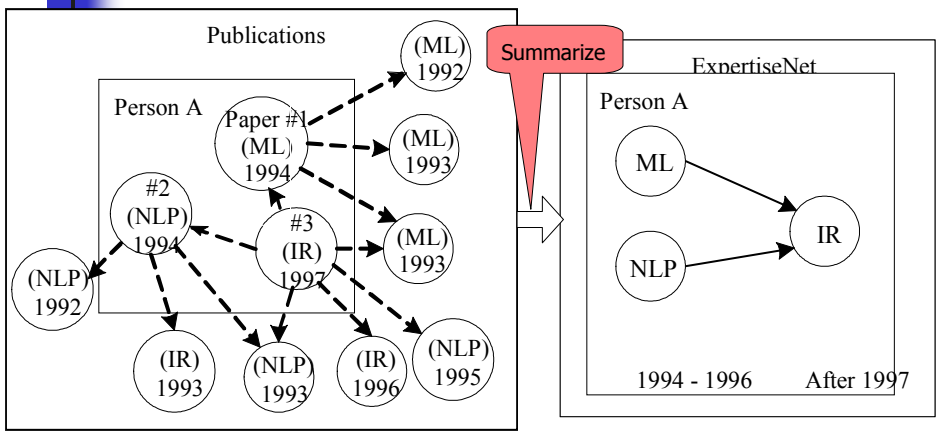
  - Y : a random graph on a set of n nodes
  - y : an observation
  - $s(y)$: graph statistics on y
    Density, reciprocity, transitivity, strengths of edges
  - θ : parameters to model the importance of the statistics on the graph
- The structure distance of two Relational ExpertiseNets

$$D(E_i, E_j) = \sum_{k=1}^{M} \left| \theta_{i,k} - \theta_{j,k} \right|$$

  - M: the total number of the parameters

50

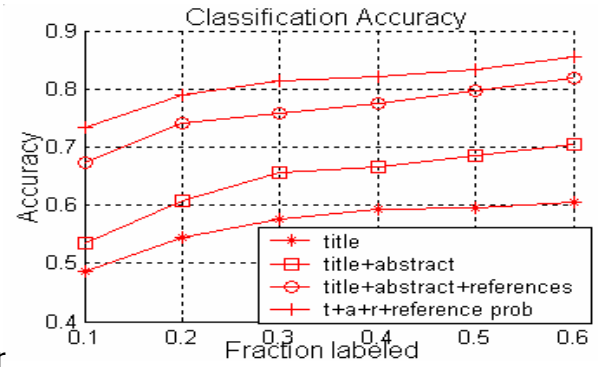
Building Evolutionary ExpertiseNet



➔ Capture *one's research expertise* + *how they evolved*

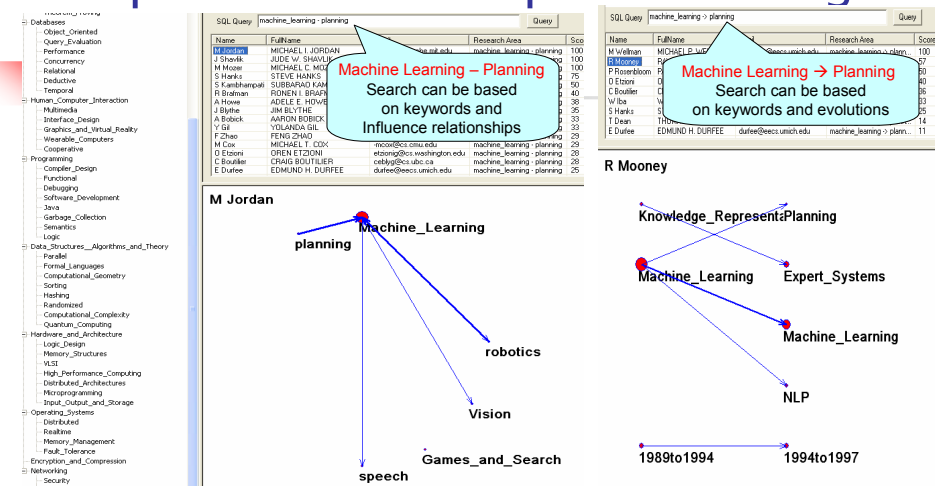
Expertise Extraction

Dataset:
Cora Research Publication Corpus
• Over *50,000* papers
• About *715,000* citation links
• Labeled into a topic hierarchy with 69 leaves
• Provide bibliographic information for each paper



➔ Incorporation of ref. prob. as a feature boosts classification accuracy

ExpertiseNet for Expertise Mining



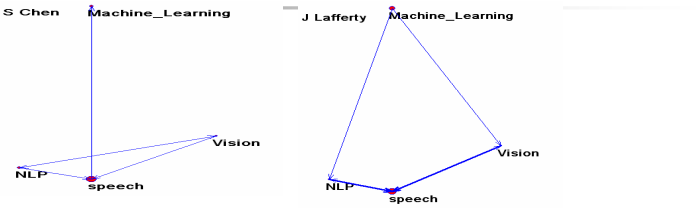
(a) *Relational mining* (b) *Evolutionary mining*

➔ Rich graphical user profiling provides a more discriminative system on mining and matching experts.

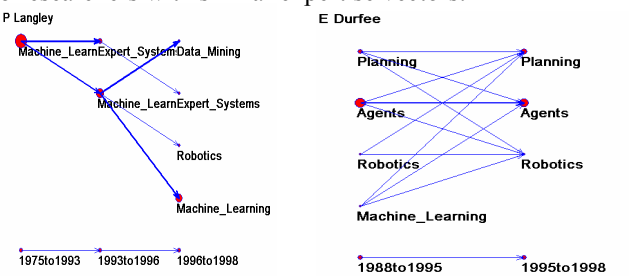
53

Examples: enhanced discriminative capability of ExpertiseNets

Two researchers with the same expertise vector:



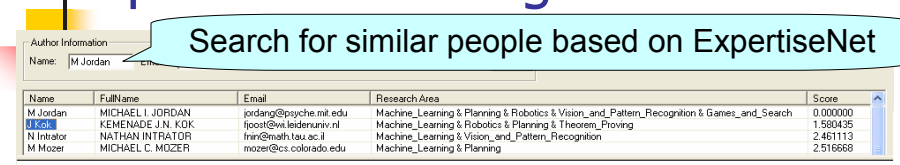
Another two researchers with similar expertise vectors:



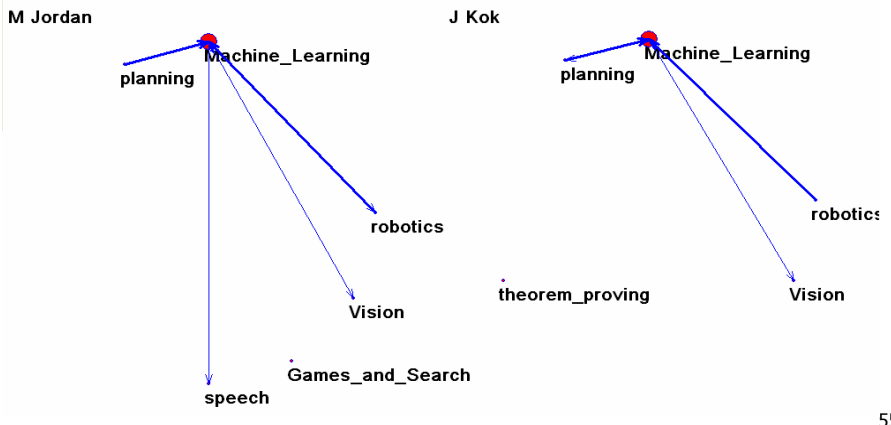
54

Expertise Matching

Search for similar people based on ExpertiseNet



Name	FullName	Email	Research Area	Score
M Jordan	MICHAEL I. JORDAN	jordan@psyche.mit.edu	Machine_Learning & Planning & Robotics & Vision_and_Pattern_Recognition & Games_and_Search	0.000000
J Kok	KEMENADE J.N. KOK	kok@wi.leidenuniv.nl	Machine_Learning & Robotics & Planning & Theorem_Proving	1.580435
N Intrator	NATHAN INTRATOR	nin@math.tau.ac.il	Machine_Learning & Vision_and_Pattern_Recognition	2.461113
M Mozer	MICHAEL C. MOZER	mozer@cs.colorado.edu	Machine_Learning & Planning	2.516668



55

Outline

- Motivation
- Social Network Analysis and Modeling
- Expertise Modeling
- Personal / Community Interest Modeling
- Sleep Quality Inference
- Conclusion

Modeling Interests for Recommendation

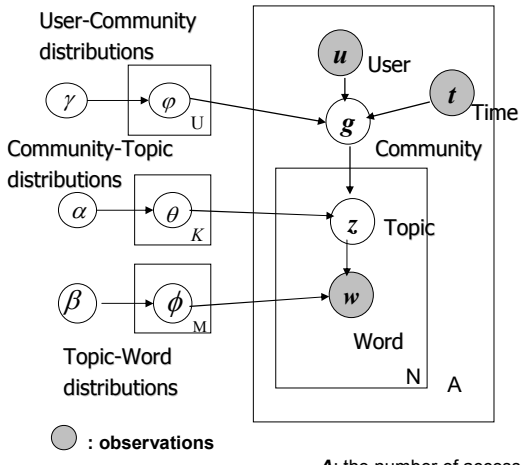
Traditionally, how to make recommendations?

- Administrator looks at the new content – decide whether the content is interesting
- If we want to use automatic recommendation
  - **Content Filtering**
    categorize the new content, then decide whether this content is similar to the ones that users accessed before.

  - **Collaborative Filtering**
    Wait a few days – if a content becomes popular (access by many people similar to the user), then it must be interesting

57

Finding Communities and Topics Modeling for Document Recommendations [Song et. al. SDM 2006]



: observations

g, z : latent variables
θ: Community-Topic distributions
ϕ: Topic-Word distributions
φ: Community distributions
α, β, γ: Dirichlet parameters

A: the number of access records
K: the number of latent communities
M: the number of latent topics
N: the number of words in a document
U: the number of users

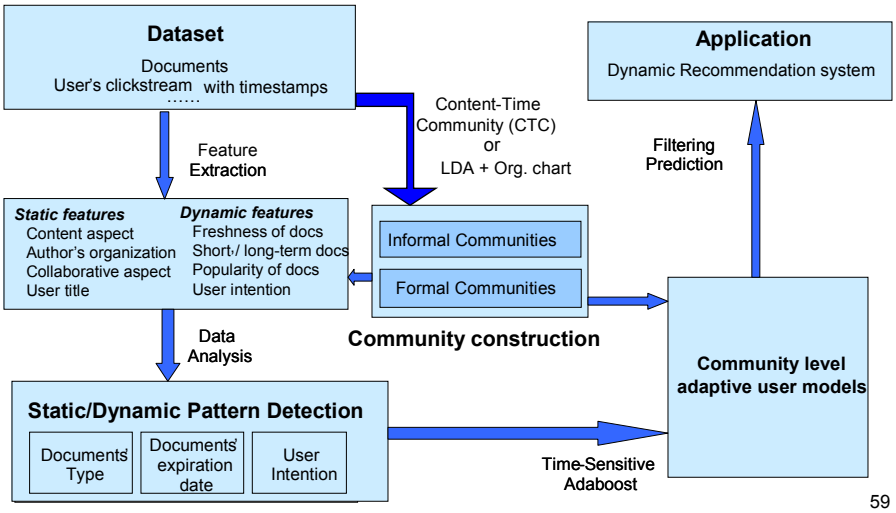
CTC Modeling:

- Build 3-stage BNs.
- Showing that models converge even for 2 latent layers
- Time determines a decaying factor of documents for measuring the similarity of users for community clustering.

➔ Obtains *communities* and *content topics* based on the observed time, users and the words in the accessed documents.

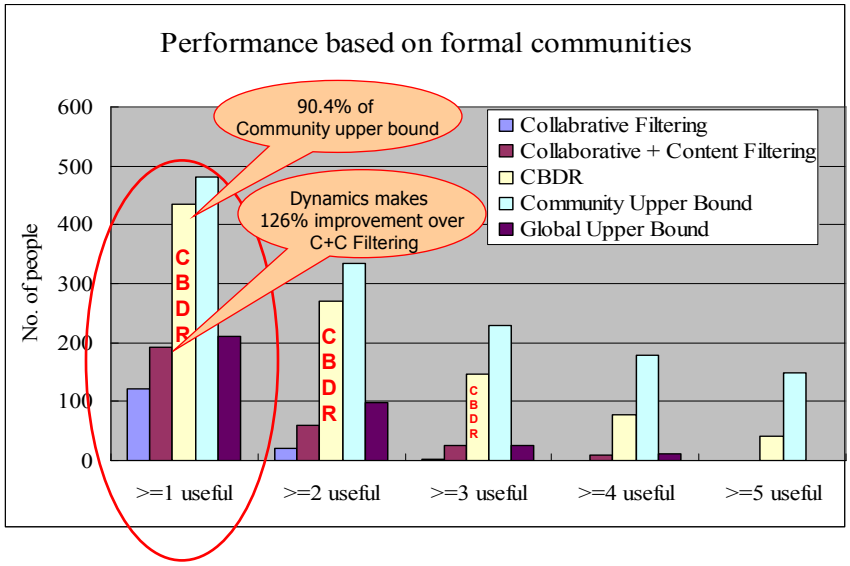
58

AIKO Community-Based Dynamic Recommendation (CBDR)



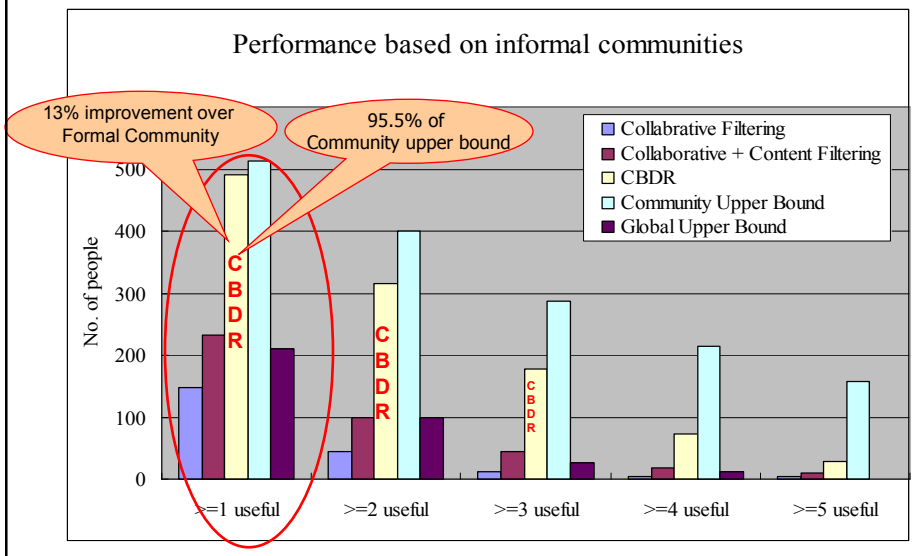
59

Formal Community Recommendations



60

Informal Community Recommendations



61

Summary of Personal Interest Modeling and Personalized Recommendation Work

- People's dynamic interests and documents' dynamic factors are analyzed to find out similar people and similar documents – CTC Model
- People's dynamic interests and documents' dynamic factors are considered for deciding whether a document is interesting to a person – Time-sensitive Adaboost.

AIKO Community Recommendation

62

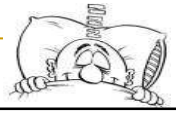
(1) Incorporate dynamic interests, document properties and user modeling (mode

Outline

- Motivation
- Social Network Analysis and Modeling
- Expertise Modeling
- Personal / Community Interest Modeling
- Sleep Quality Inference
- Conclusion

Sleep Monitoring

- Knowing a person's long-term sleep pattern is important.
- Current sleep quality monitoring is usually conducted at:
  - Clinics with specific complicated devices such as PSG.
  - Home using accelerometers (Actigraph) to record limb movements.
- Major drawbacks:
  - Because the sleeping environment is different, a subject's clinical sleep quality measurements may be affected by other factors that decrease the reliability.
  - Long term measurement of sleep quality is difficult through clinical measurement.
  - Actigraph only provide a single modality measurement. Wearing a specific device may be considered intrusive.
  - Subjective reports (e.g., sleep diary or PSQI) may not be reliable.



Our Goals

- Objective measurements:
  - Develop simple (wireless) multimodality sensors at home for long-term sleep logging.
- Early diagnoses based on machine cognition:
  - Instead of simply recording the signals, we are interested at developing inference techniques for:
    - Sleep pattern detection
    - Sleep quality detection
    - Sleep disorder detection
    - Sleep-related diseases detection

65

Many questionnaire items can be answered by automatic multimodality sensing

- Pittsburgh Sleep Quality Index (PSQI): a self-rated questionnaire (1 week/ 1 month)
- 19 individual items generate 7 component scores; their sum yields one global score.
- Example items:
  -- How many hours of actual sleep did you get at night?
  -- How often have you had trouble sleeping because you…
  Have to get up to use the bathroom?
  Cough or snore loudly?
  Had bad dreams? …
  …

First Page of PSQI Questionnaire

Appendix. Pittsburgh Sleep Quality Index (PSQI)

Name _____ ID # _____ Date _____ Age _____

Instructions:
The following questions relate to your usual sleep habits during the past month *only*. Your answers should indicate the most accurate reply for the *majority* of days and nights in the past month. Please answer all questions.

1. During the past month, when have you usually gone to bed at night?
   USUAL BED TIME _____
2. During the past month, how long (in minutes) has it usually take you to fall asleep each night?
   NUMBER OF MINUTES _____
3. During the past month, when have you usually gotten up in the morning?
   USUAL GETTING UP TIME _____
4. During the past month, how many hours of *actual sleep* did you get at night? (This may be different than the number of hours you spend in bed.)
   HOURS OF SLEEP PER NIGHT _____

For each of the remaining questions, check the one best response. Please answer *all* questions.

5. During the past month, how often have you had trouble sleeping because you...

	Not during the past month	Less than once a week	Once or twice a week	Three or more times a week
(a) Cannot get to sleep within 30 minutes				
(b) Wake up in the middle of the night or early morning				
(c) Have to get up to use the bathroom				
(d) Cannot breathe comfortably				
(e) Cough or snore loudly				
(f) Feel too cold				
(g) Feel too hot				
(h) Had bad dreams				

Some Sleep Activity Measure Metrics May Be Inferred by Multimodality Sensors

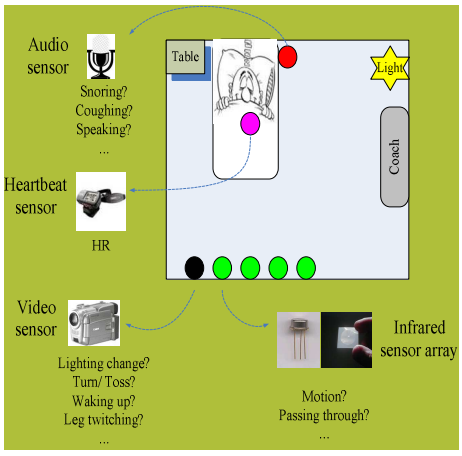
- Many of the 7 component scores in PSQI can be automatically filled up via audio-visual monitoring:
  - Subject Sleep Quality
  - Sleep Latency
  - Sleep Duration
  - Habitual Sleep Efficiency
  - Sleep Disturbances
  - Use of Sleep Medication -- may be inferred by observing abnormal patterns
  - Daytime Dysfunction – need additional wearable sensors
- Sleep-related diseases

  --- *sleep apnea, restless legs syndrome…;* they often show several syndromes during sleep. These syndromes may be observable through audio-visual signals.

67

Our Current Status

- What we have done:
  - Developed visual, audio, heartbeat and infrared sensors for sleep monitoring
  - Inference a person's sleep pattern by sleep/awake detection
  - Preliminary inference of sleep quality
  - Logging of sleep situation
- What we may do next:
  - Early detection and long term monitoring of sleep related diseases
  - Validation of the effectiveness of simple multimodality sensors with rigorous field study
  - Daytime wearable sensors to monitor dysfunctions caused by sleep disorder
  - Others…



68

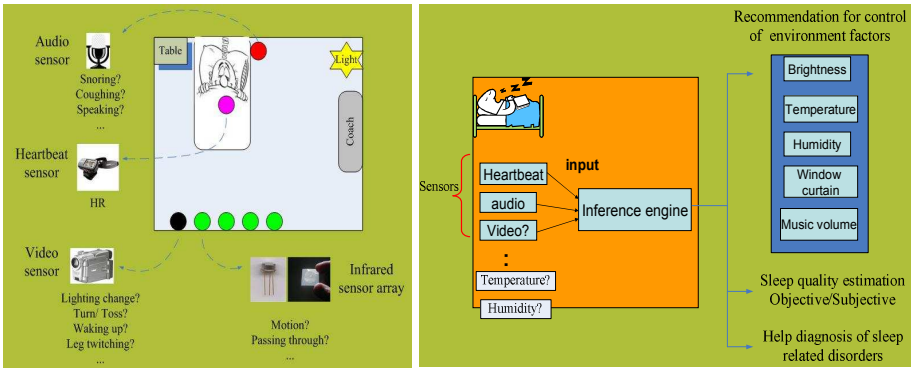
Outline

- Introduction
- Using simple-multimodality sensors to infer sleep condition (sleep vs. awake) and preliminary sleep quality measurement
- Experiments and results
- Conclusion and future works

69

Using simple-multimodality sensors for sleep monitoring

- Overall outline



→Focus on sleep/awake detection and extend the result to preliminary sleep quality inference first

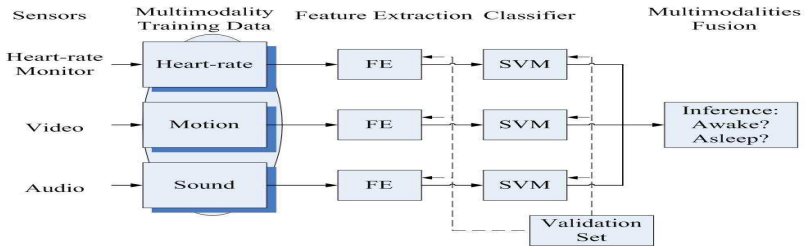
70

Using simple-multimodality sensors for sleep condition inference

- Data modalities:

  -- physiological (heart-rate), motion, sound

- System for asleep-awake detection



71

Data modalities and corresponding sensors

- Heart-rate

  -- sensor: Garmin Forerunner 301

  http://www.garmin.com

- Motion

  -- sensor: infrared webcam

  http://shop.store.yahoo.com/insidecomputer/6inniusb35we.html

- Sound

  -- sensor: laptop+ audio-recording software

72

Feature extraction (heart-rate part)

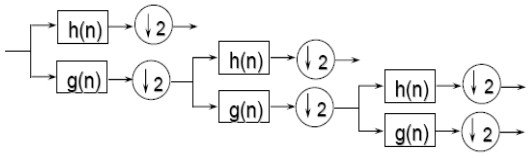
- Extracted feature

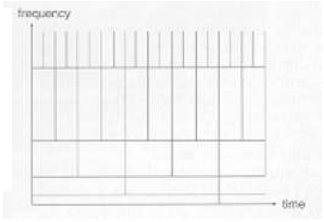
  -- power spectrum

$$F(f(x)) = F(u) = \int_{-\infty}^{\infty} f(x)\exp[-j2\pi ux]dx = R(u) + jI(u)$$

$$|F(u)|^2 = F(u)F^*(u) = R^2(u) + I^2(u)$$

  -- wavelet coefficients





Feature extraction (motion part)

- Normalized motion amplitude histogram (ME of consecutive I frames)

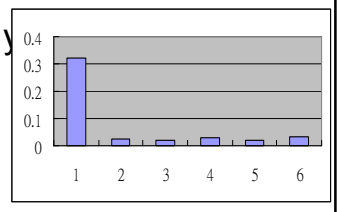
  -- Block size=8x8, SR=5 in both x, y

  -- define 6 bins as

  max(abs(Δx) ,abs(Δy))<1

  max(abs(Δx) ,abs(Δy))<2

  ⋮



- Non-motion ratio: 1st bin in the normalized motion amplitude histogram

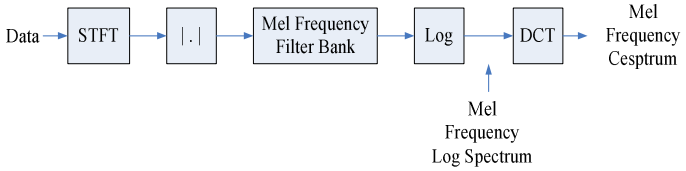
- Extracted feature

  -- Fourier transform coefficients of non-motion ratio

74

Feature extraction (audio part)

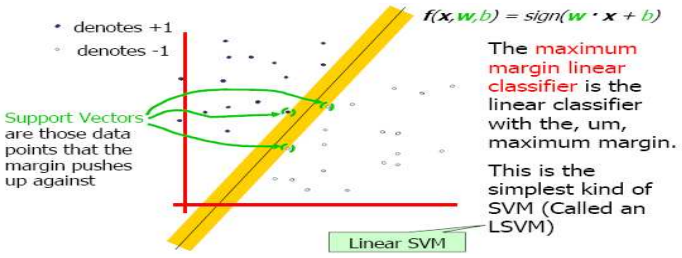
- Extracted feature
  -- Amplitude
  -- Fourier transform coefficients
  -- Mel-frequency cepstrum coefficients (MFCC)



- For each data modality, different lengths of windows with different overlaps are applied to extract the data for analysis

Classifier (1)

- SVM (support vector machine)
  --- linearly separable patterns

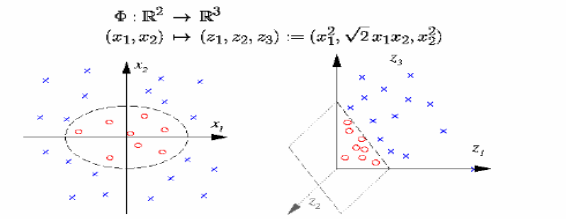


→ discriminant function $f(x) = w^T x + b$

Classifier (2)

- SVM

  --- nonlinearly separable patterns



- Kernel vs. nonlinear transform

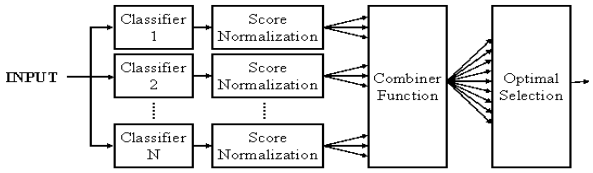
$$K(x, x_i) = \phi^T(x)\phi(x)$$

- Discriminant function

$$f(x) = \sum_{i=1}^{S} a_i K(x, x_i) + b$$

Multimodalities fusion

- Ensemble fusion



- Gaussian normalization:

$$f(x) = \frac{(x - \mu_x)}{\sigma_x}$$

- Combiner function:

  --- maxima

  --- average

$$f(x) = \max(x_1, x_2, \ldots)$$

$$f(x) = \sum_{i=1}^{N} w_i x_i$$

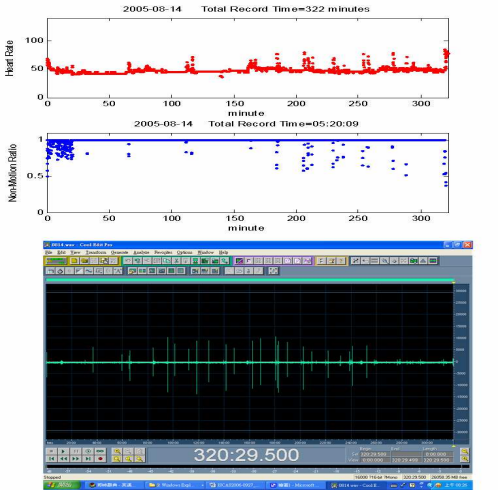
Outline

- Introduction
- Using simple-multimodality sensors for sleep condition (sleep vs. awake) inference and preliminary sleep quality measurement
- Experiments and results
- Conclusion and future works

79

Data Collection

- 28 days of HR, motion, sound data along with filled-up PSQI questionnaire
- Example data



80

Inference results of sleep-awake detection (1)

- 6 days for validation set and 20 days left for randomly partitioned training and testing sets
- Exclude audio data here…

Average performance	Classifier			
	Ensemble fusion (average)	Ensemble fusion (maxima)	Motion only	Heart-rate only
FA rate	0.2283	0.5905	0.0312	0.5484
Miss rate	0.0357	0.0174	0.0694	0.3644
Accuracy	0.9359	0.9169	0.9764	0.6173

- Motion data is a strong and dominant indicator for the sleep/awake detection…

81

Inference results of sleep-awake detection (2)

- Modified experiments that the subject is inactive during the awake time

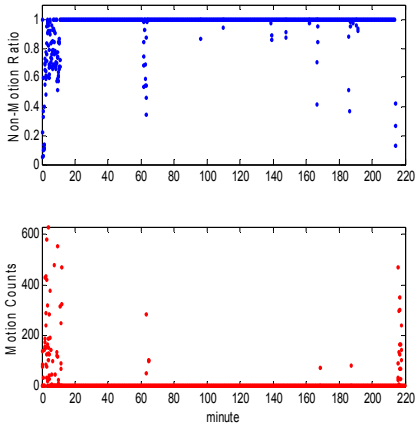
Average performance	Classifier			
	Ensemble fusion (average)	Ensemble fusion (maxima)	Motion only	Heart-rate only
FA rate	0.4143	0.475	0.9722	0
Miss rate	0.1317	0.0360	0.0053	0.3076
Accuracy	0.8410	0.9309	0.9294	0.7

- Multi-modality fusion actually can improve the classification results under some situations

82

Comparison of video and Actigraph for sleep-awake detection

- Example



- Performance with the same method applied

Average performance (5 days)	Device	
	Video sensor	Actigraph
FA rate	0.2928	0.1956
Miss rate	0.0331	0.0622
Accuracy	0.9383	0.9244

83

Preliminary results of automatic sleep quality indexing (1)

- 3 objective component scores in PSQI

  --- *sleep latency*: the time you spend before falling asleep

  --- *sleep duration*: total time you spend on the bed

  --- *habitual sleep efficiency*: asleep time/total bed time

- Using our sleep-awake inference results, only count detected awake time before detected sleep situation

84

Preliminary results of automatic sleep quality indexing (2)

- Example results

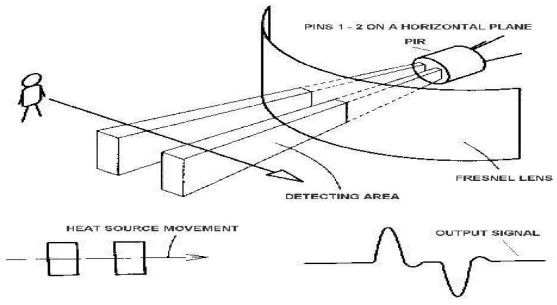
Inference Items (vs. subjective ground truth)	Example Testing Day	
	Aug. 24	Sep. 10
Awake time before sleep (vs. subjective sleep latency)	17 minutes (20 minutes)	20 minutes (22 minutes)
Sleep duration (vs. self-recorded sleep duration)	304 minutes (304 minutes)	217 minutes (217 minutes)
Habitual sleep efficiency (vs. subjective sleep efficiency)	94.4% (93.4%)	90.8% (89.9%)
Estimated PSQI (vs. subjective PSQI)	>=3 (3)	>=4 (5)

- Provide a preliminary, automatic score range for subjective sleep quality measurement

85

Extended work (1)

- For privacy concern, people may be unwilling to use video
- Using economic PIR (passive infrared sensor) to detect the motion



86

Extended Work (2)

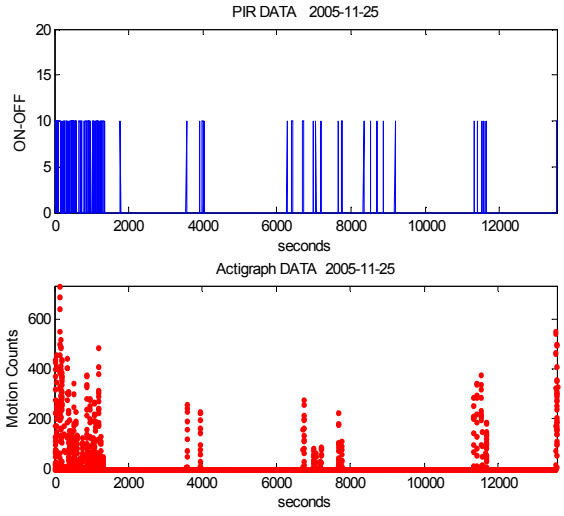
- PIR sensor, wireless TX and RX (*zigbee* communication)



87

Extended Work (3)

- Visualization of example data



88

Conclusion and future work (1)

- A novel, economical system (multimodality sensors with machine learning methods) for sleep-awake detection
- About 0.8~0.9 detection accuracy using HR and video sensors
- Explore the possibility of using simple video sensor rather than the costly Actigraph (>=$1000 USD)
- Apply the sleep-awake inference result to an automatic, preliminary indexing for subjective sleep quality assessment
- Replace video sensor with PIR sensor for motion information acquisition (data collection is going on…)

89

Conclusion and future work (2)

- Bottleneck
  - -- hardware limitation (ex. noisy HR data)
  - -- data collection (different subjects, better procedure…)
  - -- ground truth for more meaningful evaluation
  - -- better approach for sleep quality measurement (postsleep inventory?)
  - -- meaningful & valuable issues (ex. sleep log)?
- Near-Term Future work
  - -- distributed system (going on now…)
  - -- use of audio data for snoring detection, disturbance detection, etc.
  - -- behavior of HRV (heart-rate variation)
  - -- thorough measuring sleep quality via simple sensors

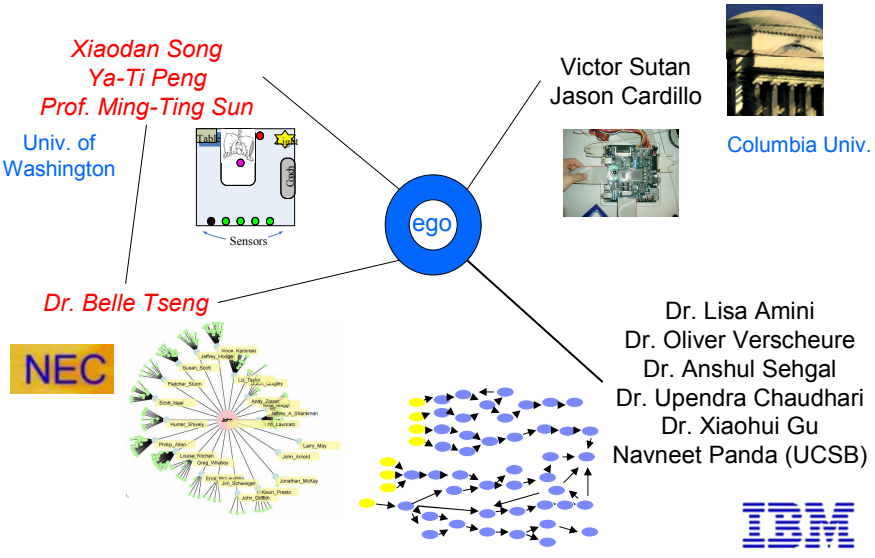
90

Collaborations with UW Sleep Lab

- Joint proposals to NIH or related interested companies.
- Set up a system including non-intrusive sensors in the Sleep Lab:
  - Compare signal quality with existing devices.
  - Test whether the captured signals can be used by experts for diagnostic purpose.
  - Compare the inference accuracy with expert opinions.
- Develop more prototype systems for home-based long term tests on sleep disordered subjects.
- Develop efficient visualization and mining/retrieval tools for long-term sleep logging and pattern analyses.

91

My Social Network – Current Collaborators



References

1. Xiaodan Song, Ching-Yung Lin, Belle L. Tseng and Ming-Ting Sun, "**Modeling and Predicting Personal Information Dissemination Behavior**," *ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, Chicago, August 2005. (KDD 2005 Student Travel Award)
2. Xiaodan Song, Belle L. Tseng, Ching-Yung Lin and Ming-Ting Sun, "**ExpertiseNet: Relational and Evolutionary Expert Modeling**," *Intl. Conf. on User Modeling*, Edinburgh, UK, July 2005. (US National Science Foundation UM05 Student Travel Award)
3. Xiaodan Song, Ching-Yung Lin, Belle L. Tseng and Ming-Ting Sun, "**Modeling Evolutionary and Relational Behaviors for Community-based Dynamic Recommendation**," *SIAM Data Mining Conference,* Bethesda, MD, April 2006.
4. Ya-Ti Peng, Ching-Yung Lin, Ming-Ting Sun and Ming-Whei Feng, "**Sleep Condition Inferencing Using Simple Multimodality Sensors**," *IEEE Intl. Symposium on Circuits and Systems*, Kos Island, Greece, May 2006.
5. Ya-Ti Peng, Ching-Yung Lin and Ming-Ting Sun, "**A Distributed Multimodality Sensor System for Home-Used Sleep Condition Inference and Monitoring**," *IEEE/AMA/BMES Transdisciplinary Conference on Distributed Diagnose and Home Healthcare*, Arlington, VA, April 2006.