NSLAB TSINGHUA UNIVERSITY

概率主题模型及其应用

Present by *Ming Xu*

November 9, 2011

内容

- 背景
- 文本表示方法
- 主题概率模型
- 应用

信息爆炸的时代

- 2002 Google -搜索20亿张网页
- 2004 Google -搜索43亿张网页
- 2006 Google -搜索超过100亿张网页

信息检索

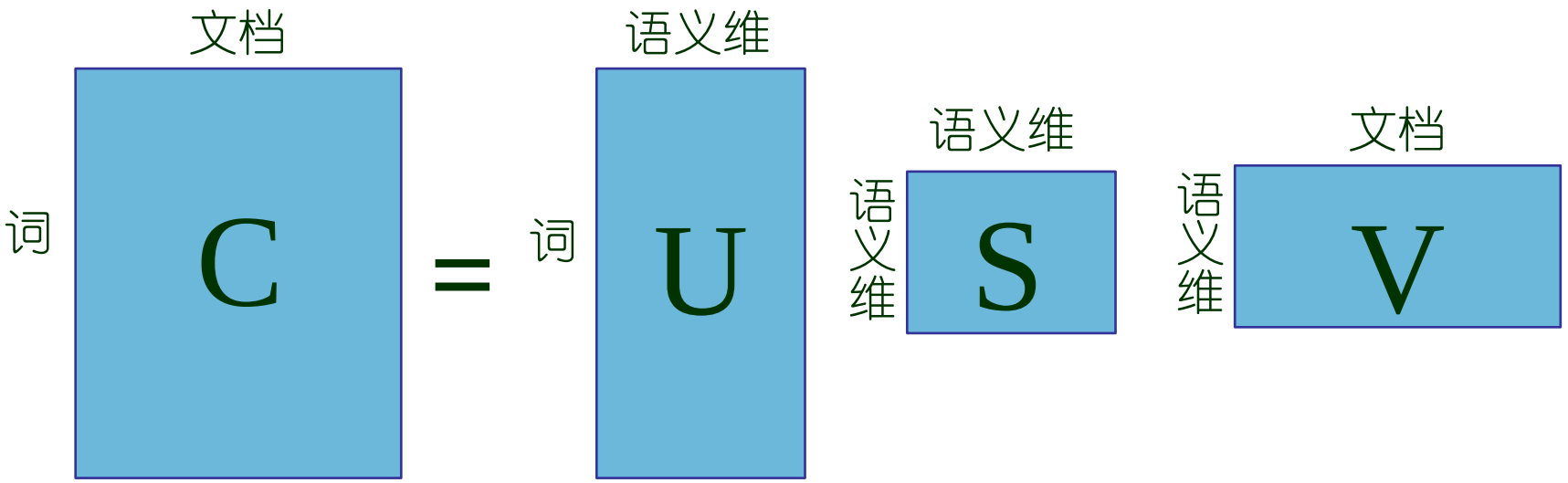
- 传统检索：关键字
- 智能检索：同音、同义，歧义信息处理
- 结合推荐系统：存在富者愈富的问题。

文本表示方法——向量空间模型

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
human	1	0	0	1	0	0
interface	1	1	0	0	0	0
user	0	1	1	0	1	0
system	0	1	1	1	0	0
response	0	1	0	0	1	0
time	0	1	0	0	1	0
EPS	0	0	1	1	0	0
survey	0	1	0	0	0	0
trees	0	0	0	0	0	1
graph	0	0	0	0	0	0
minors	0	0	0	0	0	0
…						

文本表示方法——潜在语义分析

- LSA(Latent semantic analysis):每个语义对应一个特征向量
- PLSA： P(w|t), P(t|d)



主题概率模型

Seeking Life's Bare (Genetic) Necessities

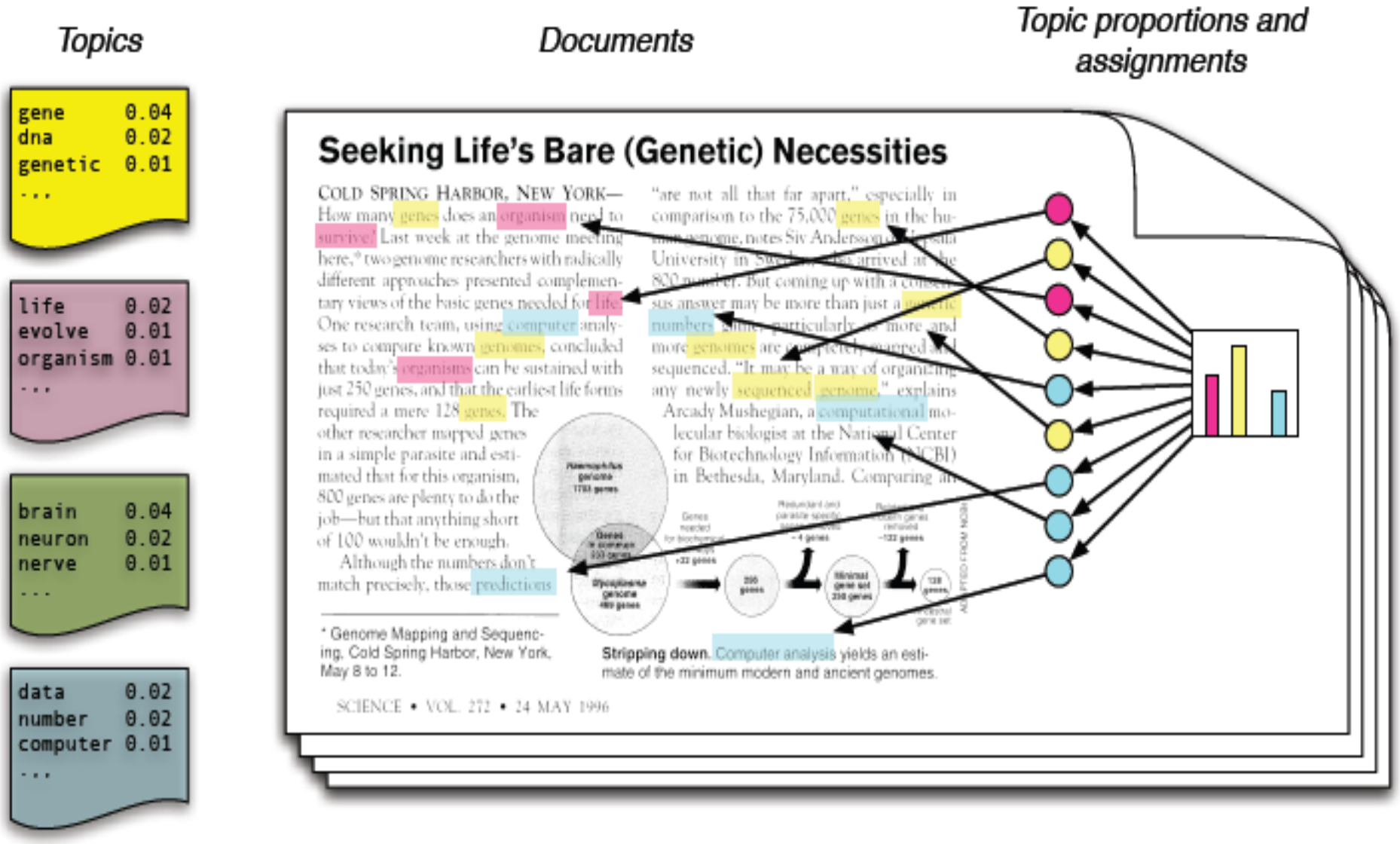
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

主题概率模型

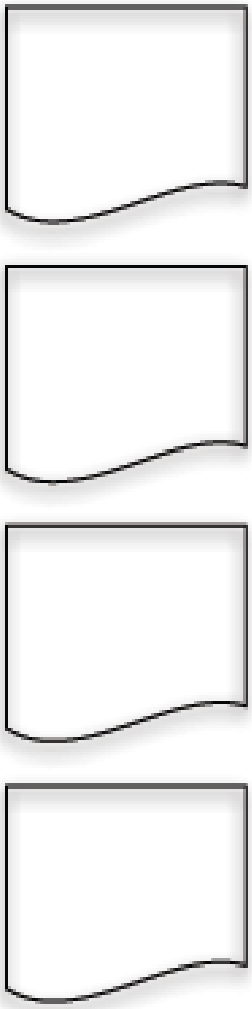

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—
How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.
Although the numbers don't match precisely, those predictions
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

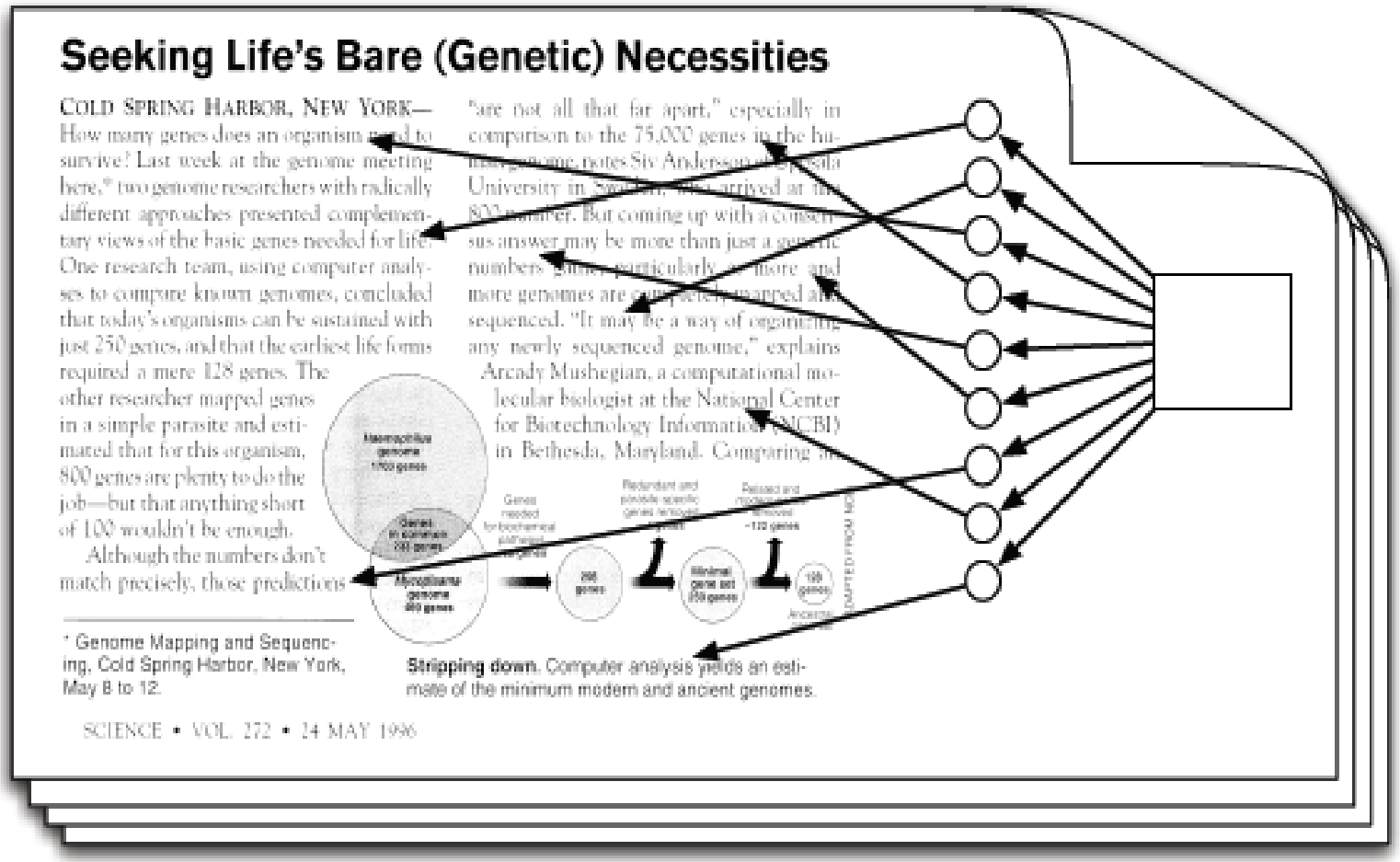
主题概率模型

Topics

Documents

Topic proportions and assignments





Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a question of numbers games, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing the

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

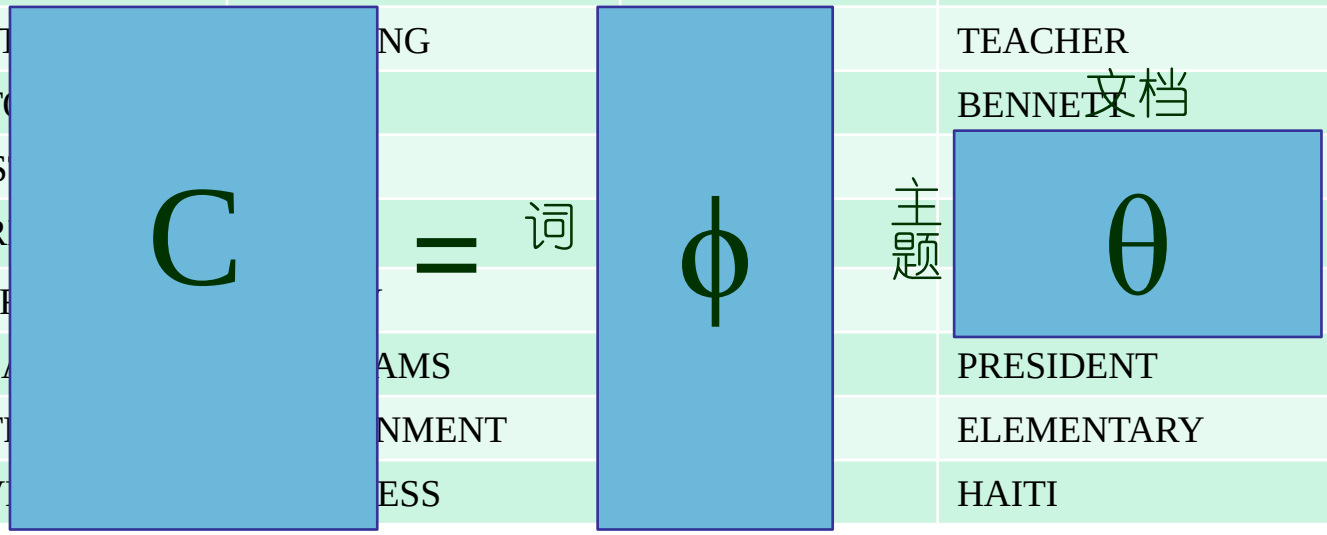
SCIENCE • VOL. 272 • 24 MAY 1996

主题概率模型

- 主题：主题是语料集合上语义的高度抽象、压缩表示。

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
Arts						
Budgets						
Children						
Education						

PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	NG		TEACHER
ACT			BENNETT
FIRS			
YOR			
OPE			
THE	AMS		PRESIDENT
ACT	NMENT		ELEMENTARY
LOV	ESS		HAITI



主题概率模型——产生式模型

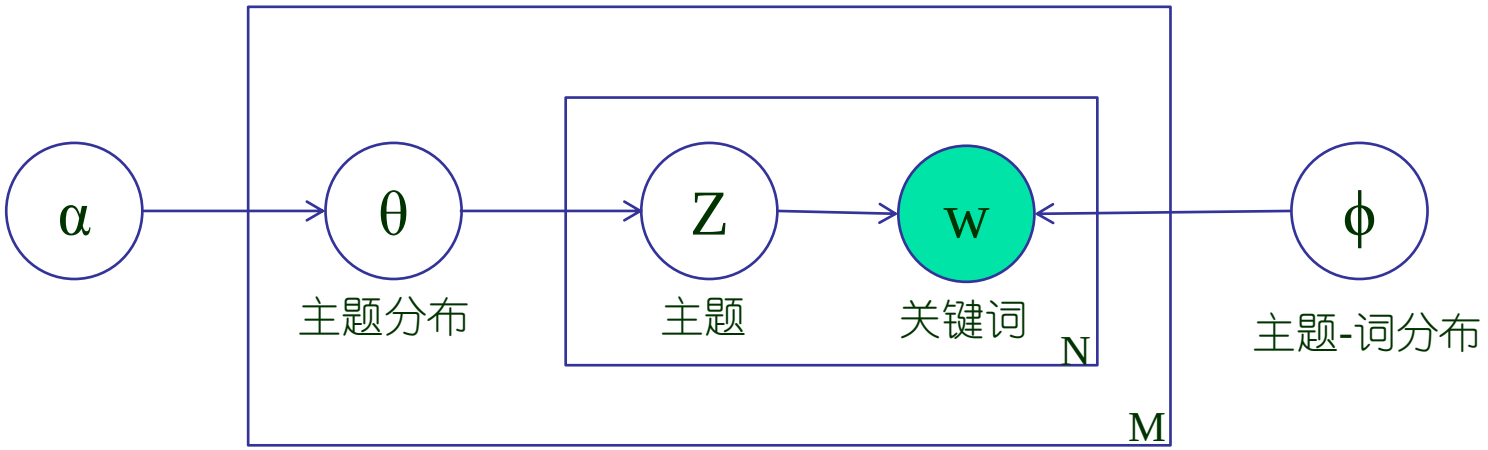
- Latent Dirichlet Allocation

Choose $\theta_d \sim Dir(\alpha)$

For each of the N words w_n :

--Choose a topic $Z_n \sim \text{Multi}(\theta_d)$

--Choose a word $w_n \sim \text{Multi}(\phi)$



Dirichlet 分布

- 掷骰子游戏

Belief	Face	1	2	3	4	5	6
0.5	Probability	1/7	1/7	1/7	1/7	1/7	2/7
0.25	Probability	1/8	1/8	1/8	1/8	1/8	3/8
0.25	Probability	1/6	1/6	1/6	1/6	1/6	1/6

LDA参数估计

- 基于变分法的EM
- Gibbs算法

主题数的确定

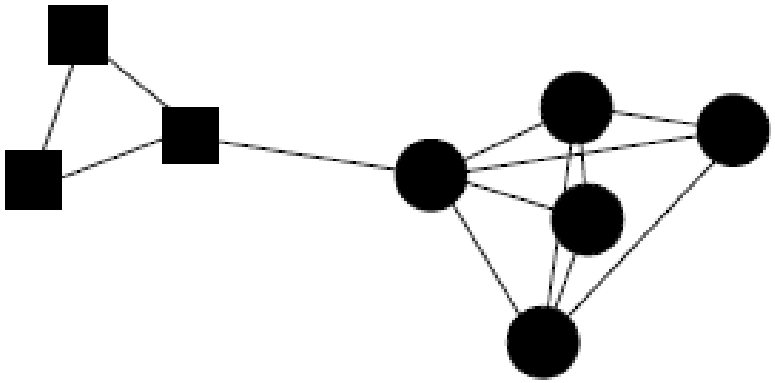
- 经验设定，大部分研究工作默认的方法
- 基于Perplexity的确定方法

$$perplexity(Dtest) = \exp\{\frac{-\sum_d \log(P(w_d))}{\sum_d N_d}\}$$

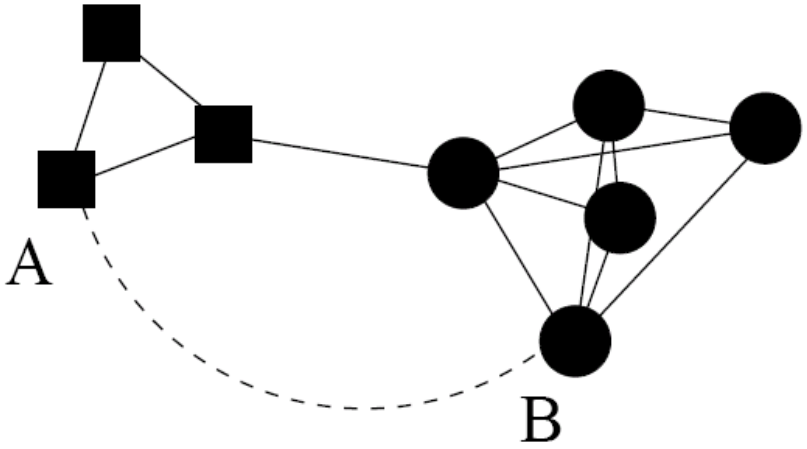
- 非参数的贝叶斯方法对主题模型进行扩展，可以自动学习出主题的数目。

应用——社会网络社团发现

- Probabilistic Models for Discovering E-Communities.www06

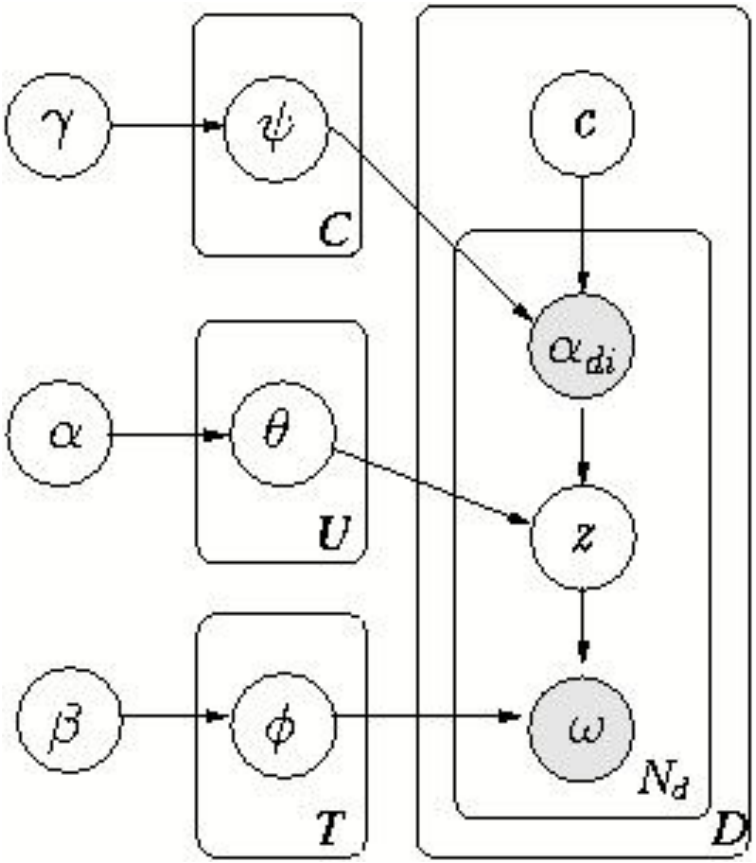


Communication frequency-based community



Communication semantic based community

应用——社会网络社团发现



CUT1: When community affects user communication only.

应用——社会网络社区发现

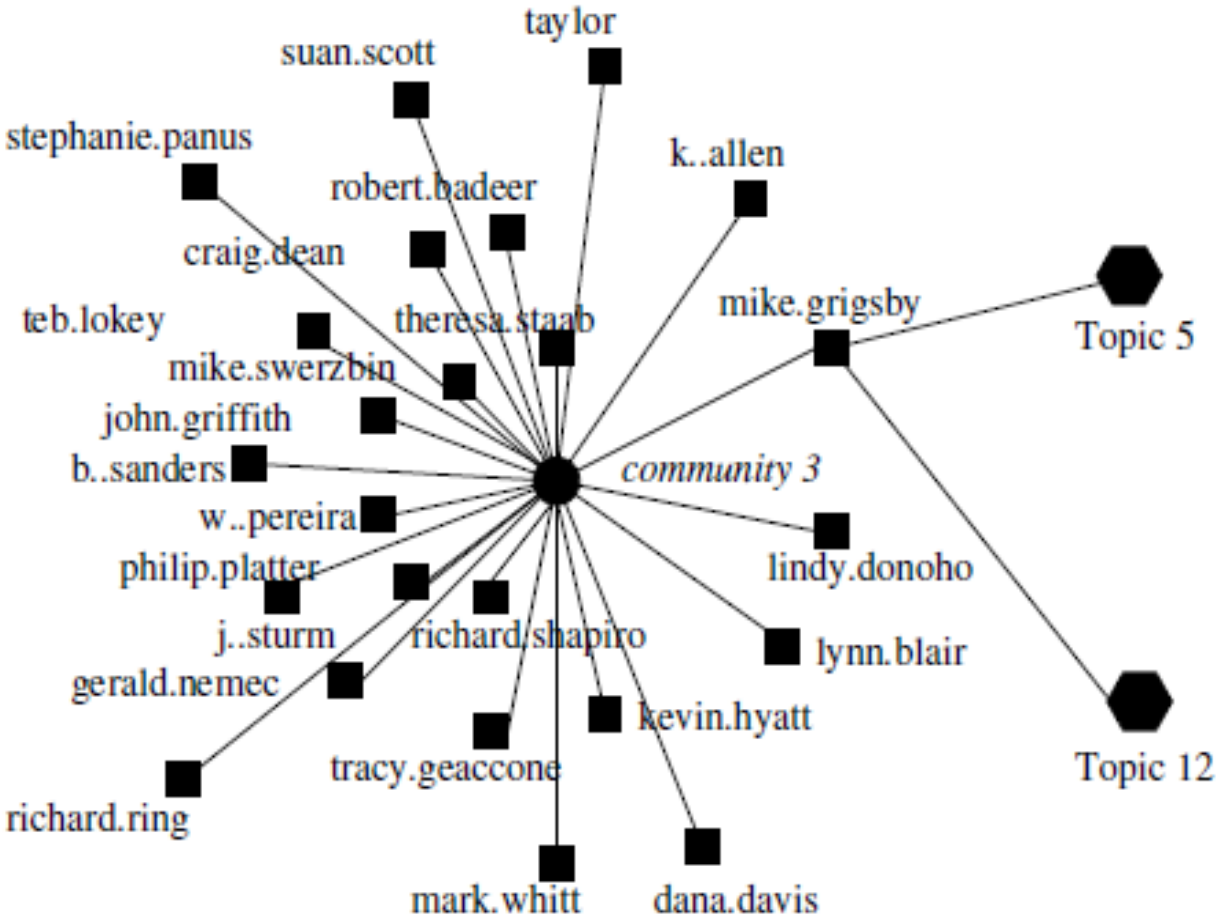
经验设定T=20，C=6

采样4个主题

Topic 3	Topic 5	Topic 12	Topic 14
rate	dynegy	budget	contract
cash	gas	plan	monitor
balance	transmission	chart	litigation
number	energy	deal	agreement
price	transco	project	trade
analysis	calpx	report	cpuc
database	power	group	pressure
deals	california	meeting	utility
letter	reliant	draft	materials
fax	electric	discussion	citizen

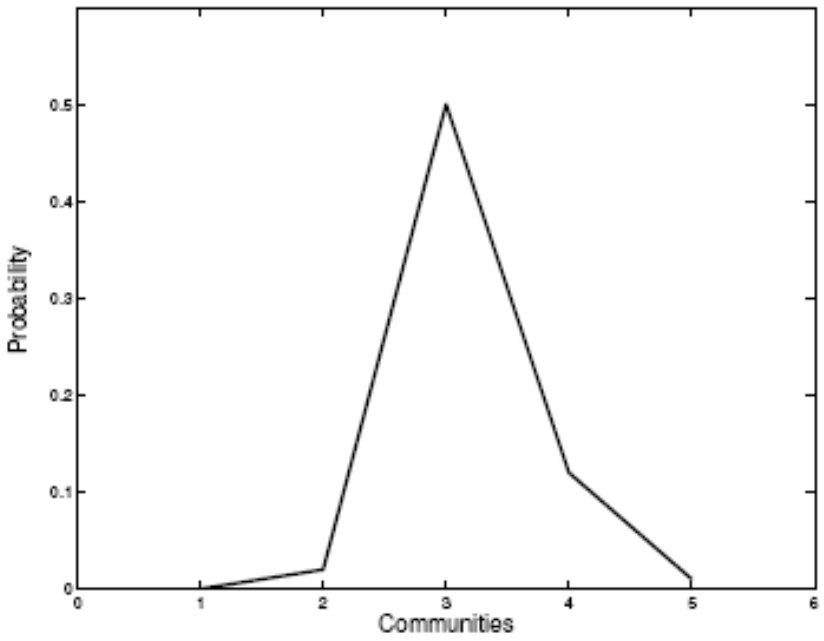
abbreviations	organizations
dynegy	An electricity, natural gas provider
transco	A gas transportation company
calpx	California Power Exchange Corp.
cpuc	California Public Utilities Commission
ferc	Federal Energy Regulatory Commission
epsa	Electric Power Supply Association
naruc	National Association of Regulatory Utility Commissioners

应用——社会网络社区发现

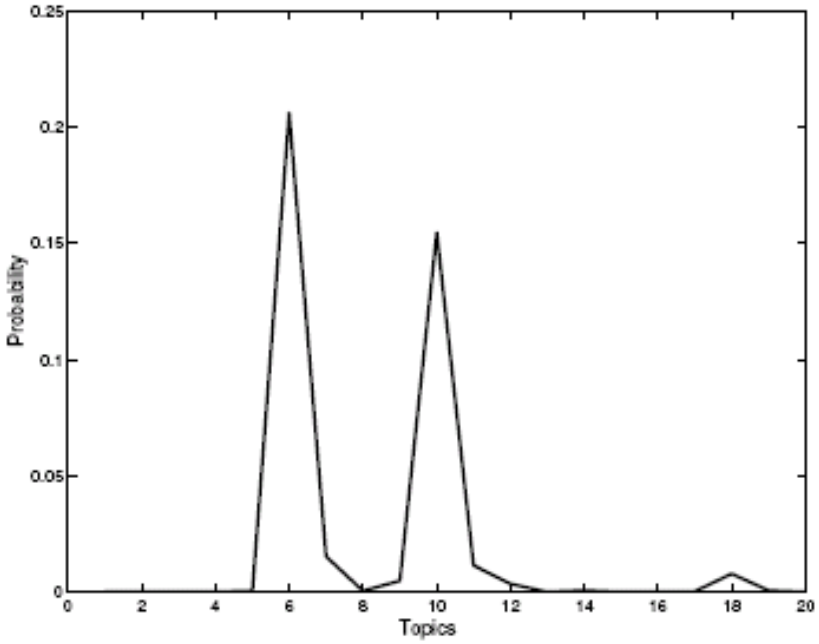


Topic 5	Topic 12
dynegy	budget
gas	plan
transmission	chart
energy	deal
transco	project
calpx	report
power	group
california	meeting
reliant	draft
electric	discussion

应用——社会网络社区发现



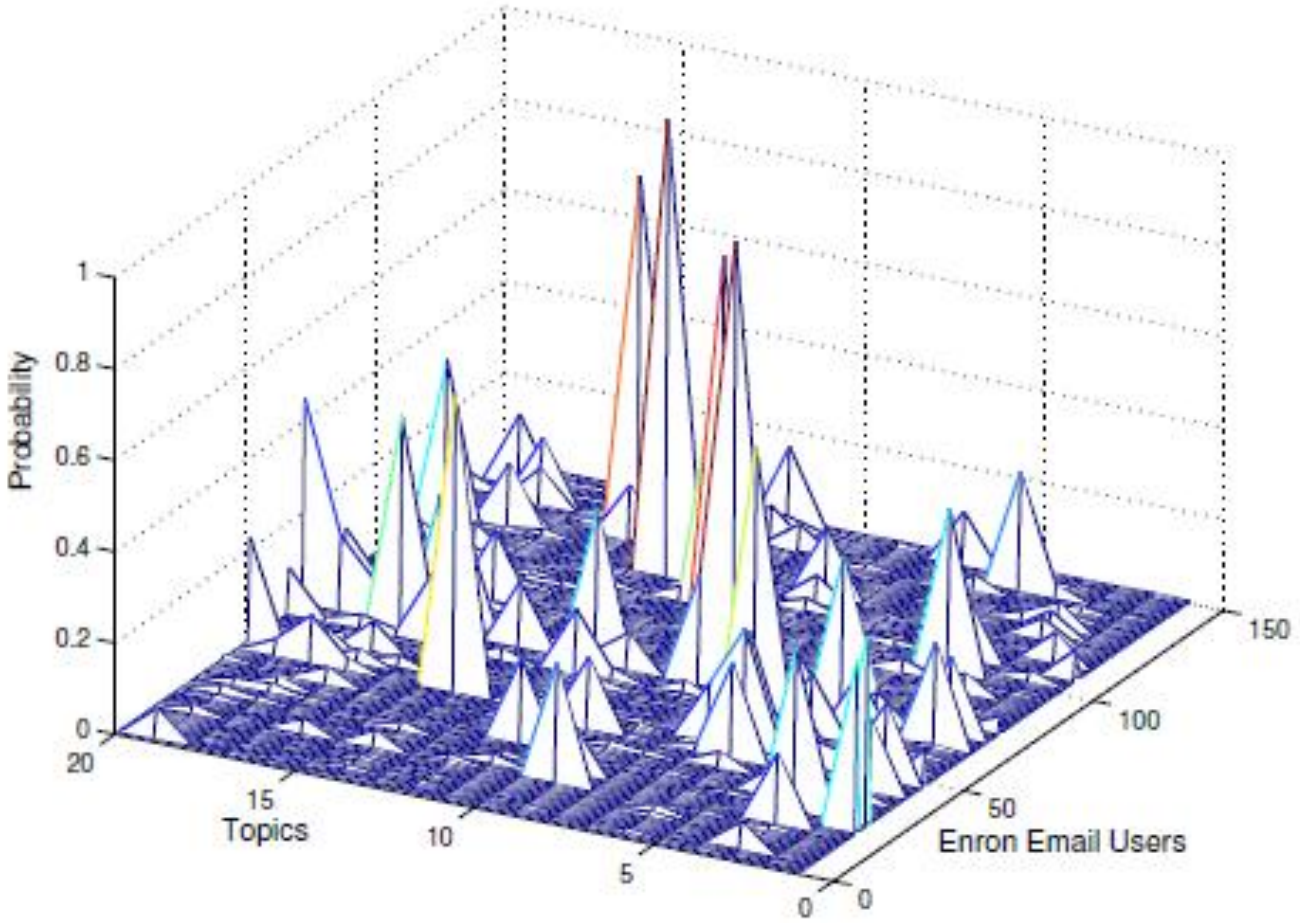
(a) Over communities



(b) Over topics

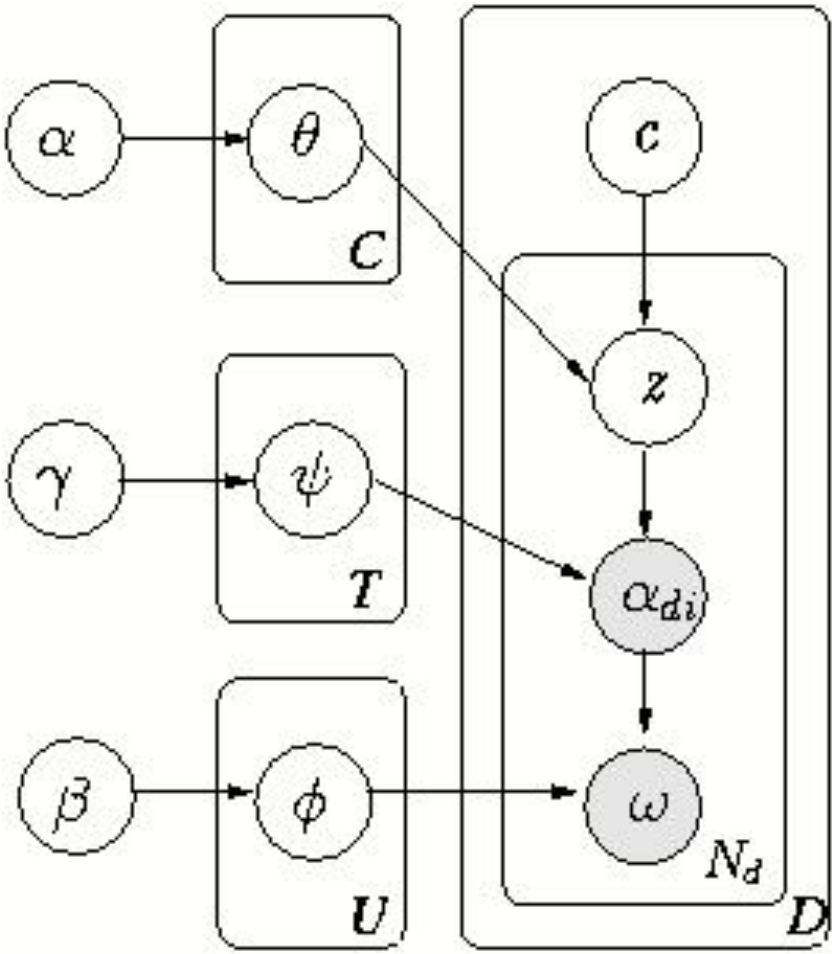
CUT2: When community affects topic generation only.

应用——社会网络社区发现



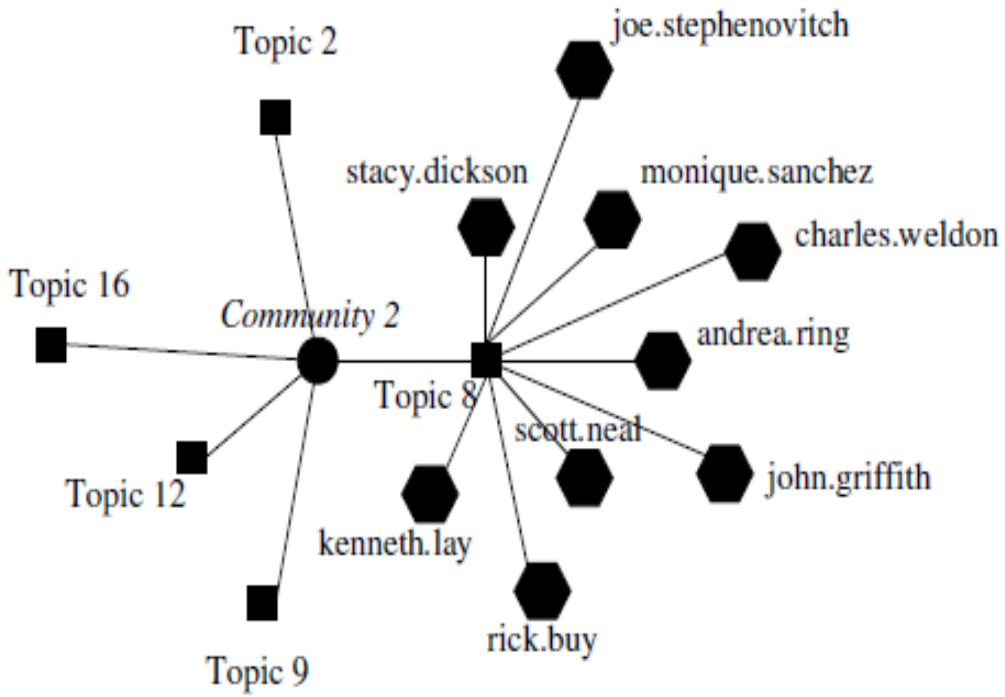
Distribution over topics for all users.

应用——社会网络社区发现



CUT2: When community affects topic generation only.

应用——社会网络社区发现



A Community Discovered by CUT_2.

d..steffes	cara.s	mike.grigsby	rick.buy
power	number	file	corp
transmission	cash	trader	loss
epsa	ferc	report	risk
ferc	database	price	activity
generator	peak	customer	validation
government	deal	meeting	off
california	bilat	market	business
cpuc	caps	sources	possible
electric	points	position	increase
naruc	analysis	project	natural

Distribution over words of some users.

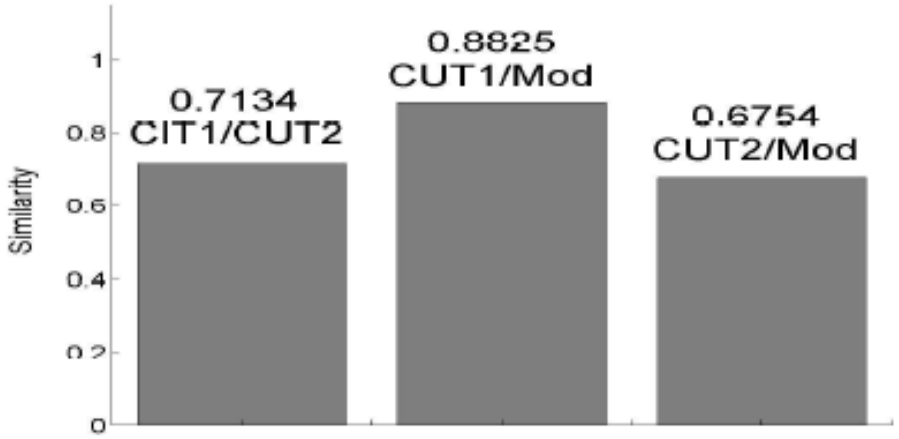
应用——社会网络社区发现

$$\lambda = \frac{N_{00} + N_{11}}{N*(N-1)/2}$$

where $0 \le \lambda \le 1$, 1 indicates the same

N_{00} is the number of objects in same cluster for both clustering

N_{11} is the number of objects in different clusters for both clustering



Community similarity comparisons.

Thank you