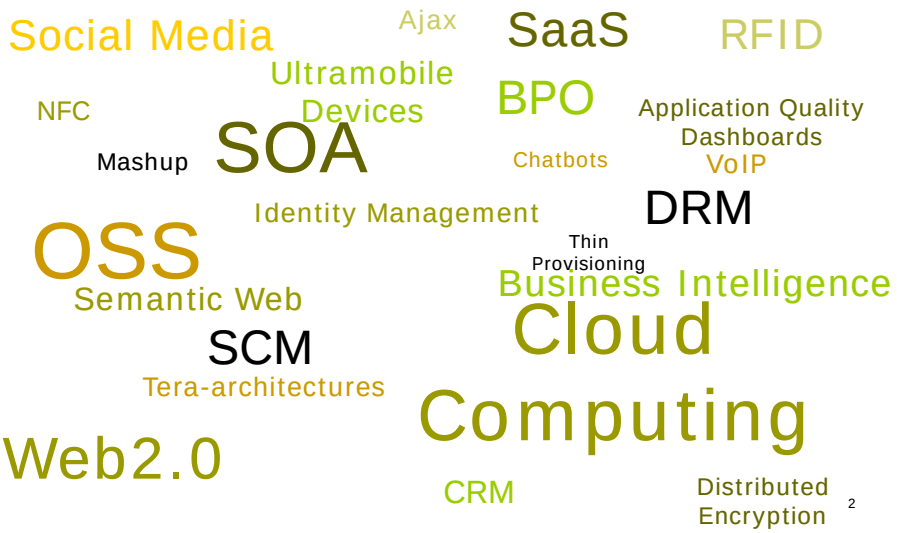
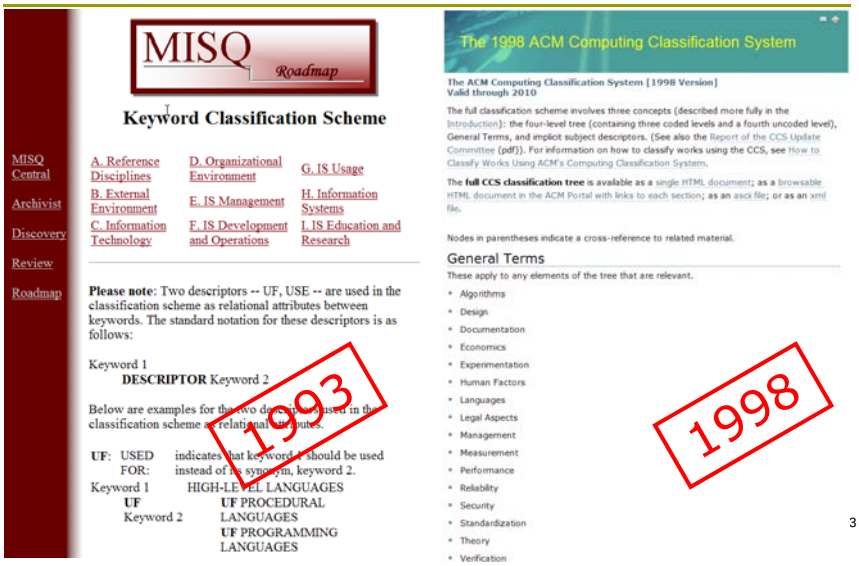
Evaluating Two Automatic Methods for Classifying IT Concepts

Chia-jung Tsui
Ping Wang
Kenneth R. Fleischmann

Lots of IT Concepts, But …



… Little and Dated Understanding



Present Approach to IT Classification

- Compile list of IT concepts from surveys
- Experts classify ITs according to their assessments of functions or features of the technologies.
- Challenges
  - Arbitrary and limited choices of features and functions, lacking objectivity
  - Not scalable: more ITs → lower reliability
  - For special purposes, not general purpose

4

Our Data

- Downloaded full-text articles published in 1998-2007 from six magazines:
  - *ComputerWorld & InformationWeek*
  - *BusinessWeek & The Economist*
  - *Newsweek & US News and World Report*
- Extracted ~220,000 paragraphs containing 35 IT concepts

5

Concepts in ProQuest Classification

1 air traffic control, aviation, global positioning system
2 application service provider, enterprise resource planning
3 artificial intelligence, robot
4 blog, rss technology, social networking
5 business intelligence software, customer relationship management, data mining, salesforce automation
6 chat room, instant messaging
7 data warehouse, online analytical processing
8 digital music, mp3 player
9 electronic commerce, online advertising, online sales, public key infrastructure, world wide web
10 inventory management, supply chain management
11 linux, open source software
12 quality control, six sigma
13 service oriented architecture, web service
14 virtual private network, virtualization

6

Co-Occurrence of IT Concepts

"Over the past few years, we have seen the **ERP** vendors-led by SAP-move into different business areas," says Byron Miller, an analyst with the Giga Information Group. "The competitive advantage of just having **ERP** has diminished. The next big thing beyond **ERP** is **supply-chain management**."

Links between **groupware** and **ERP** applications speed users' access from within a groupware application to key business data, such as purchase orders, inventory, customer histories, and other supply-chain information.

7

Partial Co-Occurrence Matrix

	AI	ASP	ATC	CRM	DM	EC	ERP	GPS	PKI	WWW	...
AI	571*										
ASP	0	1,060									
ATC	0	0	326								
CRM	2	39	0	2,014							
DM	13	2	0	44	1,336						
EC	10	114	3	201	55	12,311					
ERP	1	47	0	264	32	245	2,318				
GPS	1	0	8	0	0	7	1	504			
PKI	0	2	0	0	0	37	2	0	283		
WWW	6	3	2	2	9	141	29	1	4	1,411	
...											

8

Kullback-Leibler (KL) Divergence

- KL divergence measures difference between two probability distributions.

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log(P(i) / Q(i))$$

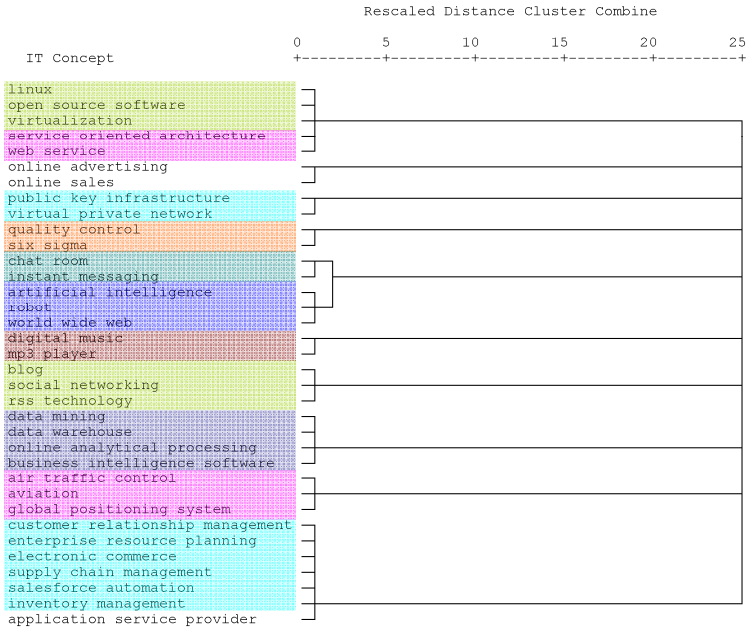
- Symmetrized KL divergence matrix by averaging divergence values in each direction.

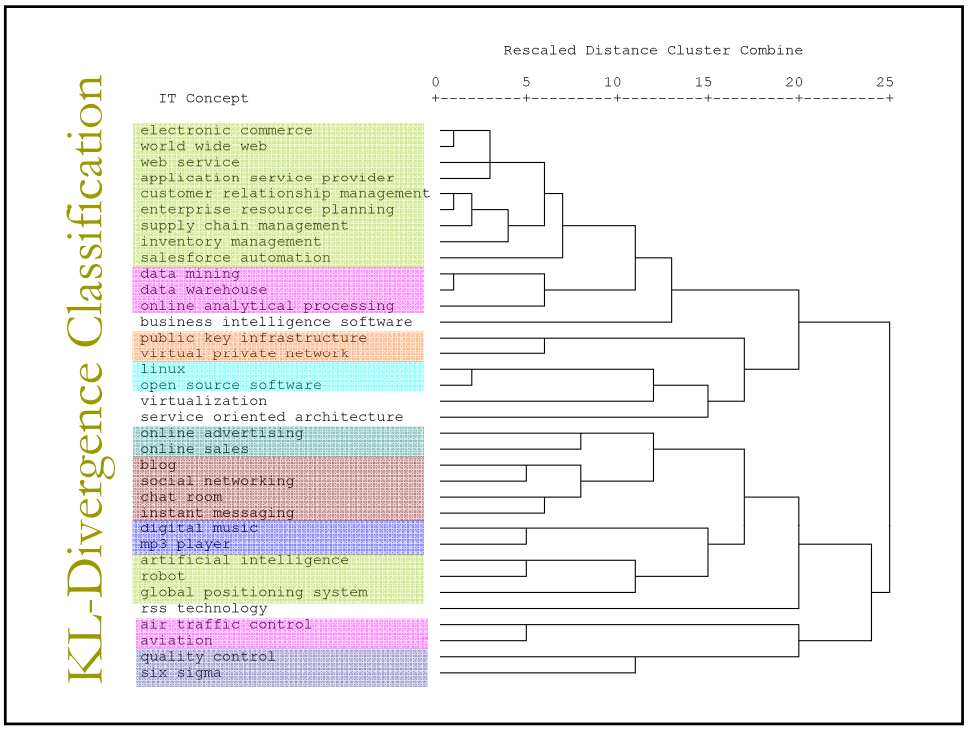
$$D_{KL}(P \parallel Q) = \frac{\sum_i P(i) \log(P(i) / Q(i)) + \sum_i Q(i) \log(Q(i) / P(i))}{2}$$

	AI	ASP	ATC	CRM	DM	EC	ERP	GPS	PKI	WWW	...
AI	0.00										
ASP	0.60	0.00									
ATC	0.84	0.82	0.00								
CRM	0.60	0.24	0.88	0.00							
DM	0.47	0.39	0.85	0.30	0.00						
EC	0.44	0.19	0.68	0.21	0.29	0.00					
ERP	0.58	0.24	0.80	0.16	0.31	0.20	0.00				
GPS	0.61	0.63	0.64	0.70	0.65	0.50	0.65	0.00			
PKI	0.79	0.50	0.97	0.66	0.66	0.45	0.60	0.81	0.00		
WWW	0.41	0.26	0.68	0.32	0.31	0.15	0.25	0.48	0.47	0.00	
...											

9

Co-occurrence Classification





Concepts Consistently Clustered

1. air traffic control, aviation
2. artificial intelligence, robot
3. blog, social networking
4. chat room, instant messaging
5. data warehouse, online analytical processing
6. digital music, mp3 player
7. inventory management, supply chain management
8. linux, open source software
9. quality control, six sigma

12

Quantitative Evaluation: F-Measure

		Correct classification	
		Positive	Negative
Obtained classification	Positive	true positive (*tp*)	false positive (*fp*)
	Negative	false negative (*fn*)	true negative (*tn*)

- Precision = tp/(tp + fp) Recall = tp/(tp + fn)
- F-measure $= \frac{2 \times precision \times recall}{precision + recall}$
- Co-occurrence classification F:0.729
- KL-divergence classification F:0.615

13

Qualitative Evaluation: Survey

Classification A
application service provider
artificial intelligence, robot, world wide web
business intelligence software, data mining, data warehouse, online analytical processing
customer relationship management, electronic commerce, enterprise resource planning, inventory management, salesforce automation, supply chain management
linux, open source software, virtualization
online advertising
online sales
public key infrastructure, virtual private network

Classification B
application service provider, enterprise resource planning
artificial intelligence, robot
business intelligence software, customer relationship management, data mining, salesforce automation
data warehouse, online analytical processing
electronic commerce, online advertising, online sales, public key infrastructure, world wide web
inventory management, supply chain management
linux, open source software
virtual private network, virtualization

1. In your opinion, which classification, A or B, is a better one? Please explain.

14

Rankings by Survey Respondents

IT Classification	1	1.5	2	2.5	3	Mean Rank
ProQuest classification	11	1	6	2	1	1.55
Co-occurrence-based classification	4	1	9	5	2	2.00
KL divergence-based classification	2	0	5	5	9	2.45

15

Benefits of This Approach

- Objective
  - Produce classification from IT discourse itself
  - Classification does not rely on experts
- Scalable
  - More periods, sources, and concepts
- Generalizable
  - General purpose classification
  - Applicable to broad range of contexts
  - Co-occurrence and KL divergence are generalizable.

16

Limitations

- Specific data sources
- Specific period (1998-2007)
- Specific IT concepts
- Applied to IT; what about other domains?
- Small respondent pool for evaluation survey

17

Takeaways

- Co-occurrence, KL divergence, hierarchical clustering can help explore IT concept relationships in objective, scalable, and generalizable way.
- Success does not always depend on complexity of methods.
- Automated approach can help human experts, rather than replace them.

18

Thanks & Contact Info

UNIVERSITY OF MARYLAND

- We thank National Science Foundation for grants IIS-0729459 and SBE-0915645
- Many thanks to Doug Oard, Amy Weinberg, Asad Sayeed, Lidan Wang, and our survey respondents!

Ping Wang pwang@umd.edu
University of Maryland
Room 4105 Hornbake Bldg
College Park, MD 20742-4325
http://terpconnect.umd.edu/~pwang
新浪微博 @DrWangPing

19