



Evaluating Two Automatic Methods for Classifying IT Concepts

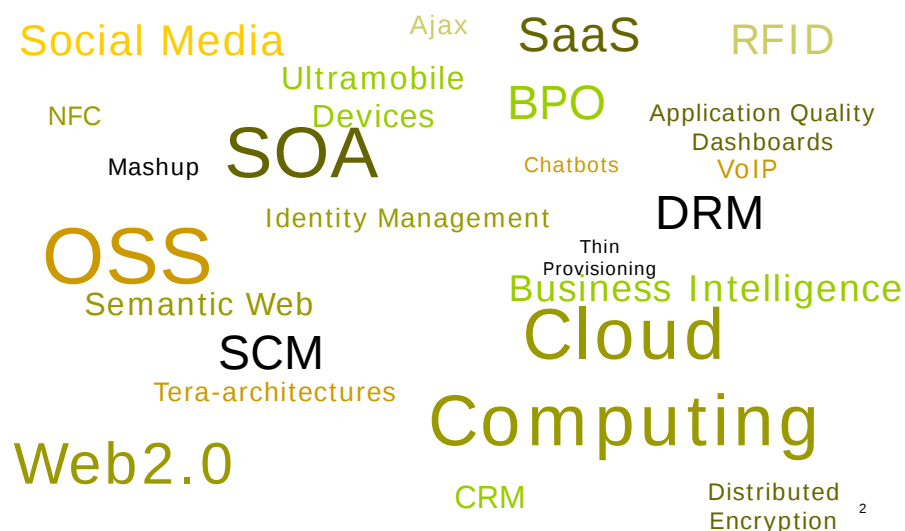
Chia-jung Tsui

Ping Wang

Kenneth R. Fleischmann



Lots of IT Concepts, But ...



... Little and Dated Understanding

MISQ Roadmap

Keyword Classification Scheme

A. Reference Disciplines	D. Organizational Environment	G. IS Usage
B. External Environment	E. IS Management	H. Information Systems
C. Information Technology	F. IS Development and Operations	I. IS Education and Research

Please note: Two descriptors -- UF, USE -- are used in the classification scheme as relational attributes between keywords. The standard notation for these descriptors is as follows:

Keyword 1
DESCRIPTOR Keyword 2

Below are examples for the two descriptors used in the classification scheme as relational attributes:

UF: USED indicates that keyword 1 should be used
 FOR: instead of the synonym, keyword 2.

Keyword 1 HIGH-LEVEL LANGUAGES
 UF
 Keyword 2 UF PROCEDURAL LANGUAGES
 UF PROGRAMMING LANGUAGES

The 1998 ACM Computing Classification System

The ACM Computing Classification System [1998 Version]
 Valid through 2010

The full classification scheme involves three concepts (described more fully in the Introduction): the four-level tree (containing three coded levels and a fourth uncoded level), General Terms, and implicit subject descriptors. (See also the Report of the CCS Update Committee (pdf)). For information on how to classify works using the CCS, see How to Classify Works Using ACM's Computing Classification System.

The full CCS classification tree is available as a single HTML document; as a browsable HTML document in the ACM Portal with links to each section; as an ascii file; or as an xml file.

Nodes in parentheses indicate a cross-reference to related material.

General Terms

These apply to any elements of the tree that are relevant.

- Algorithms
- Design
- Documentation
- Economics
- Experimentation
- Human Factors
- Languages
- Legal Aspects
- Management
- Measurement
- Performance
- Reliability
- Security
- Standardization
- Theory
- Verification

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

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.

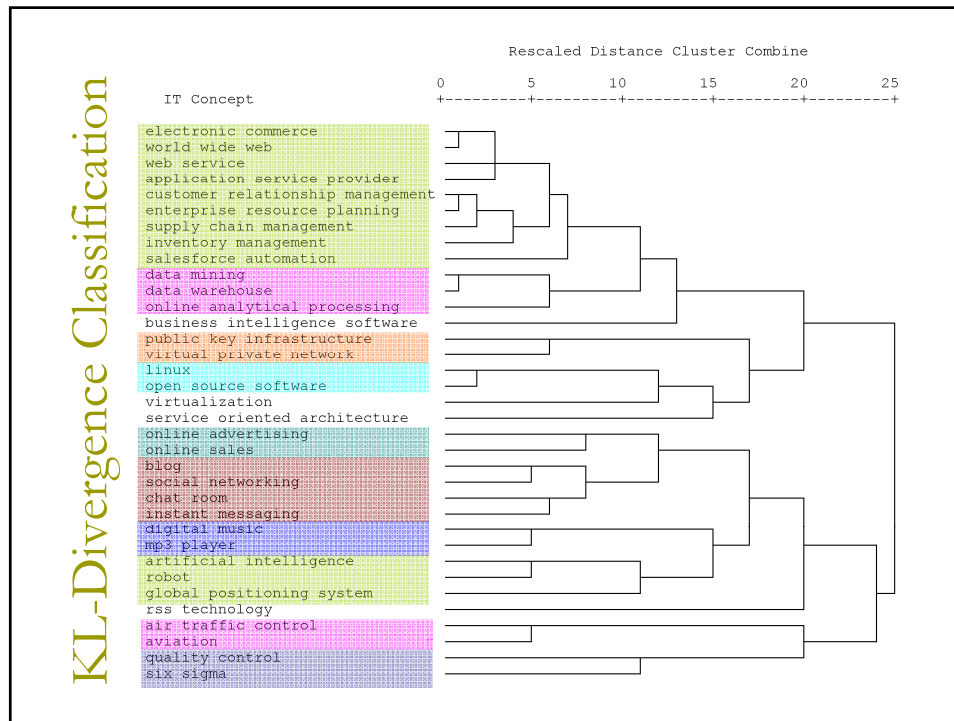
$$D_{KL}(P \parallel Q) = \frac{\sum_i P(i) \log(P(i) / Q(i)) + \sum_i Q(i) \log(Q(i) / P(i))}{2}$$

	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

Quantitative Evaluation: F-Measure

		Correct classification	
		Positive	Negative
Obtained classification	Positive	true positive (<i>tp</i>)	false positive (<i>fp</i>)
	Negative	false negative (<i>fn</i>)	true negative (<i>tn</i>)

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



□ 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