

NUS - Extreme - Tsinghua

Challenges in Real-time Social Media Analytics: What has changed over the last 5 years?

CHUA Tat-Seng

National University of Singapore

OUTLINE

- The Live Social Observatory
- The Technologies behind the System
- Next Generation Systems and Challenges
- Other Related Research
- Summary



Users at Center of Social Media Environment



Functions of Social Media Platforms

From users' perspective

- Communications: sharing, interacting, keeping up-to-date.. with friends
- Expression: air/project views
- Self-preservation: Wellness, exercise, self-improvements
- Local communities
- From platform providers' perspective
 - Attract and retain users; enrich contents; monetization
 - Offer innovative/fun services; improve user engagement
 - Co-creation of contents with users



Key Difference between Social Media Platform and Broadcast Media



Ex	pressing			Networkin	g
Publication	Discussion Distinct Dist	n Unovo Unovo ISQUS Uscovo ISQUS Uscov ISQUS Uscov ISQUS Uscov	Search Classmates DO Classmates DO Classmates DO Classmates DO Classmates Cla	Niche BtoB Linked In P plaxo XING X viadeo	Mobile Tools (foor) COO Kitting Kitting Coo Coo Kitting Coo Coo Kitting Coo Coo Kitting Coo Coo Kitting Coo Coo Kitting Coo Coo Kitting Coo Coo Kitting Coo Coo Kitting Coo Coo Kitting Coo Coo Kitting Coo Coo Coo Coo Coo Coo Coo Co
		Social P	lattorms		
View Flickr Content vimeo Summa ILike Video Proto Music	magnelia OScribd Preddit Skideo (Construction)	facebook Ofriends SKYROCK Immediate National Immediate Hyvesset BUZZNO	ster Octo orkut hij Prwndows Uver www.consultaria standaria Zorpia.com	Social Games Zynga Store Cesulcit oweves Cesulcit	Casual Games Correction Pogo Correction Region Correction Region Correct
Crowdstow feedback juin @www.second juin juin stylehive @www.second juin Recommand* Suggestions SI	are internet	tful 2007PLR Trip	Casual MMO		MMORPG MMORPG
	Sharing			Gaming	

Broadcast Media Filter->Publish

Offers curated content

Social Media Publish->Filter

Offers no content; Contents all created by Users

Functions of Social Media Platforms

- Lead to 3 key sources of live info streams:
 - Spontaneous User-Generated Contents (UGC)
 - Device-Generated Contents (DGC)
 - Structured Database Contents (SDC)

What Happens in an Internet Minute?





NExT Research Center (http://next.comp.nus.edu.sg)

- NExT: NUS-Tsinghua Center on Extreme Search
 - Initiated in May 2010 with a 5-Year Grant of S\$10 million from MDA/NRF of Singapore
 - Among the first to examine big data analytics
- Vision: To access live information that is not easily available on the current Web

00:60 Social media
media
Pinterest 1090
Солледиало 2000 CHEEKING 👟 👟
flickr 3125 Protect == == ==
TAGGED 3500
Linked in 7,610 @ @ @ @ @ @ @
Stumble loop 7.630
STUMBLES
facebook 700K00000000000
MESSAGES SENT
00000000
000000000
You Tube ZIVIIVI aaaaaaaaa

66666666

SOURCES: Platerest (http://techsnuch.onu/2017/02/07/pinterest-monthly-unliques) Foursquare (http://techsuches.onu/2017-06-02/tech/30097137_1_foursquare.uses-merchants-ins) Flickr (http://www.pinterest-moncerv.2017/07/interest-2017) Flickr (http://www.pinterest-analytic.uses-merchants-ins)
Tagged (http://bdo.ut.tagged.com) Linkedin (http://bdo.ut.tagged.com) StumbleUgen (http://www.stumbleupon.com/bbs//bdo.ut/beckie.ord-a-web-page-on-stumbleupon) Twitter (http://bdog.twitter.com/2010/06/20-onition-twest-par-day/kmit)
Facebook (http://www.tacebook.com/inde.php?hote_id=9155668919) Youtube (http://www.youtube.com/i/press_statistics)
SOCI@L JUMPSTART
Designed by David Fung (#cobradeve) for Social Jumpstart

Overview of Research From Unstructured Info into Structured Knowledge



Platform as a Service: Indian Elections Sub-Events Around an Entity (Modi)

 vesenze	Narendra Modi 🌉 🔸 Week:201411 👻				👤 🝷 chua
Home	SUB TOPICS				
Timeline					
Sub Topics	Rin Seat Give Ticket Sir	Bip.Follow.Tweet.		Ur.1	Fake.Sir.
Entity Graph	Djp,Seat,Give,Ticket,Si	Twitter,Rss		Mo	dus,Frm
Trending Words	sir please also give ticket to BJP candidate from sitamarhi (Bihar) th birth place of maa ianaki Idon't aive this seat 2 ally	e there is not even a single symbol		thanx	sir fr ur areat
Sentiment		of BJP(lotus) in your twitter cover		word f	
Influence Users		follow "Vyakti pooja"		again:	
Hot Zone					
System Management 🔹				-371	
		Bjp,Varanasi,Contest,Poll,S eat	Kejriwal,Arvind, Gujarat,Question ,Development	Media,Aap,C ongress,Say, Kejriwal	Win,Electio Wish,Sir,Gu rat
	Sir,India,Pm,Wish,Hope	BJP's Narendra Modi to contest from MM Joshi's Varanasi seat in Uttar	A Gujarati Replies to the 16 Questions asked by Arvind	Who exposed Narendra Modi? Media -	happy holi to you sir and wi you all the be
	sir,I hope u r the next PM of India sir.	Bjp,Seat,Varanasi,Contest, Candidate	Kejriwal to Narendra Modi	No Congress - No AAP - YES	for upcoming lok sabha elections. Yo
		BJP's prime ministerial candidate Narendra Modi to contest from	rime ministerial candidate dra Modi to contest from 7,Address,Shri,Delhi,O a WATCH LIVE: Shri Narendra		Speech,Know, Support,Bjp,T hackeray
	Vote, Congress, Bjp, Party, Time MY vote for mr narendra modi bjp, aap has become paap, and congress is self made secular, rest partys another face of CONGRESS	Rally,Address,Shri,Delhi,O disha Today WATCH LIVE: Shri Narendra			Raj Thackeray supports Narendra Moc but his Sena m hurt BJP

Platform as a Service: Indian Elections Details of a Sub-Event (Modi)

ļ	iveSenze	Narendra Modi 🧟 🗸 Week:201411 👻	👤 👻 chua
		A Home /	
-	Home	TOPICS DETAIL	
Ξ	Timeline		
÷	Sub Topics	(i) govt.gujarat,state,need,india 2707	
ĝ	Entity Graph		
	Trending Words	= Topic Detail	Images
υŴ	Sentiment	NSUI UP West: @ narendramodi alloted 8 lakh sg. m land to L&T @ Re1 per sg. m whose actual price is Rs950 per sg. m. 🕥	
5	Influence Users	State loss 800000xRs949. Why?	
9	Hot Zone	Shrishail s k: @ narendramodi farmer dipend n agricultu it dipend n tree, trees r in forest if forest II burn, farmer II	
°	System Management 🛛 👻	bum. i nen india?	
		Narendra Modi For PM: RT @ narendramodi: SSC & HSC exams begin tomorrow in Guj. Best wishes to my young friends.	
		Waseem: 376561 sq mtr land to raheja@Rs457 per sq mtr while air force asked to pay 1100 /sq mtr # GujratDictatorshipModel	
		AMITKUMAR.BHIMANI: @ narendramodi Dear sir , kindly look into the matter regarding provide exemption form B.ed for Schuler Sch	
		AMITKUMAR.BHIMANI: @ narendramodi Dear sir , kindly look into the matter regarding provide exemption form B.ed for MscIT&CA students	
		AMITKUMAR.BHIMANI: @ narendramodi Dear sir , kindly look into the matter regarding provide exemption form B.ed for Scale	GUIARAT, OESTI INATION IN OSI
		MBS: @ narendramodi alloted 8 lakh sq. m land to L&T @ ₹1 per sq. m whose actual price is ₹950 per sq. m. State loss 800000x₹949. Why?	
		Cyber Activist: @ narendramodi alloted 8 lakh sq. m land to L&T @ ₹1 per sq. m whose actual price is ₹950 per sq. m. State loss 800000x₹949. Why?	

Platform as a Service: Indian Elections Relations between Entities (Modi)



Platform as a Service: Indian Elections From Data to Event to Reports







Insights Gained from Social Media Analytics

Events, topics, Users and their relationships



Example 2: Automobile Sales in China Sub-Events related to Camry

Li	veSenze	丰田凯美瑞 💮 ▼ Week:201414 ▼			👤 🝷 chuats
	E	A Home /			The second second
	Home	SUB TOPICS			
:=	Timeline				
₽	Sub Topics	邹 美瑞悦宣版街 東38800 元	混合动力凯美瑞•尊瑞	凯美瑞新增红	骨灰级的凯美
躑	Entity Graph	引夫・而化子 旅先 38800元 新美瑞悦求版矩意38800元, 当天参加团购订车客户可求受每增加一台优惠500元, 以此累计, 上不封顶, 报名热线: 0391-3669988	科技改变生活。在提倡绿色、节能、环保的今天,混合动力凯美瑞 • 導瑞帶着先进、成熟的技术将怎样改变我们的生活呢?低油耗、 低排放,绿色未来不是幻想,就在眼前,来广州车展看看吧!	 色车型全新配 置年内上市 「常田: 飢業務新増紅色 辛型全新配置年内上市 「常本田官方称, 飢業務 「常本田官方称, 飢業務 	瑞粉丝,你们 左哪里?
	Trending Words				在咖里!
100	Sentiment				【骨灰级的凯美瑞粉 丝,你们在哪里?快来
<u>1</u>	Influence Users				看看这货】由丰田和 RK Motors联合打造。
8	Hot Zone		将会新增红色车濠车 另外配置方面也作出		5.9升TRD NASCAR V8 发动机,最大功率680
°°	System Management *		凯美瑞跨界极限拉力	修改。目前在告凯美瑞有6 种颜色,新款为红色版	马力,配合Tremec T- 56
			【丰田SEMA疯狂阵容凯美瑞跨界极限拉力】CamRally凯美瑞。	本,取代了香槟色,看…	
			由NASCAR车手PorkerKligerman、KyleBuschMotorsports以及 Mooresville、NCshopDetroitSpeed等联手打造,是一台纯粹的	纽约车展发布 丰田美版新款凯美瑞 消息 【 _{銀約车展发布}	
		丰田将凯美瑞•尊瑞作为领跑车	拉力获车,配备了宽体车身、外观升级签		
			2013年国产B级车前十车型销量分析 析 【2013年国产B级车前十车型销量分析】2013年1-8月, 我国B级车 销量前十车型依次是大众帕萨特、大众迈腾、丰田凯美瑞、现代素		
		秉承者"绿色=未来"的环保理念,广汽丰田将凯美瑞•尊瑞作为短跑车,其绿色环保的			
		品牌形象与马拉松和谐、健康的理念完美融合。		凯美瑞携全系车一起亮相美丽蓉城 新中国汽车工业已走过了漫漫60周年,凯美瑞国产以来	
			纳塔8代、雪佛兰迈锐宝、本田雅阁、别克新君威、日产新天籁	也经历了7年风雨。悄然逝去	的是年华,沉淀的却是…



Example 2: Automobile Sales in China Details of a Sub-Event

Li	veSenze	丰田凯美瑞 🧱 ▼ Week:201414 ▼	👤 🝷 chuats
	÷	A Home /	
	Home	TOPICS DETAIL	
:=	Timeline		
;;;	Sub Topics	(i) 丰田将凯美瑞·尊瑞作为领跑车 (142)	
268	Entity Graph		
	Trending Words	Topic Detail	🖾 Images
200	Sentiment	爱是至奢华的一件事:【将推出混动版本2013款留克萨斯ES细约车展亮相】据悉,ES300hd的动力总成与目前在售的混动版凯美珊相似,均为2.5升油	
<u>1</u>	Influence Users	电混合发动机。全车标配十个气囊,并可以选装车道偏离报警和盲点监控系统。而在音响方面甚至可以选装督克萨斯配备在高级车型上的15扬声器的 高级marklevin	
8	Hot Zone	◎▲ 彩木・【滋稚山福祉新水2013新雪古茜斯FS短約车里式相】据录 FS300bd的动力单点与目前在供的福祉新期单端相似。均为2.5升油由温全设动机	
°°	System Management 💌	-**** 全年标配十个气囊,并可以选续车道偏离报警和官点监控系统。而在音响方面甚至可以选续留克萨斯配备在高级车型上的15扬声器的高级marklevin	
		东方诗蕊:【将推出混动版本2013款留克萨斯ES纽约车展克相】揭怨,ES300hd的动力总成与目前在售的混动版凯美瑞相似,均为2.5升油电混合发动机,全车标配十个气室,并可以选装车道偏离报警和盲点监控系统,而在音响方面甚至可��◆◆法装留克萨斯配备在高级车型上的15扬声器的高级marklevin	TOYOTA
		韩字:【将推出混动版本2013款留克萨斯ES细约车展亮相】据悉,ES300hd的动力总成与目前在售的混动版凯美瑞相似,均为2.5升油电混合发动机。 全车标配十个气囊,并可以选续车道偏离报警和盲点监控系统。而在音响方面甚至可以选装留克萨斯配备在高级车型上的15扬声器的高级marklevin	
		happyjy! 【将推出滚动版本2013款督克萨斯ES經約率展亮相】据悉,ES300hd的动力总成与目前在售的混动版凯美瑞相似,均为2.5升油电混合发动 机,金车标配十个气囊,并可以选装车道偏离报警和盲点监控系统,而在音响方面甚至可以选装督克萨斯配备在高级车型上的15扬声器的高级marklevin	P refilitor
		夏霆:【将推出混动版本2013款督克萨斯ES纽约车展亮相】据悉,ES300hd的动力总成与目前在售的混动版凯美瑞相似,均为2.5升油电混合发动机。 全车标配十个气盛,并可以选装车道偏离报警和盲点监控系统。而在音响方面甚至可以选装督克萨斯配备在高级车型上的15扬声器的高级marklevin	
		●婷:【将推出混动版本2013款留克萨斯ES經約车展亮相】据悉,ES300hd的动力总成与目前在售的混动版凯美确相似,均为2.5升油电混合发动机, 全主版配十个气空,并可以法装主借值离报整知首占监控系统,而在音响方面甚至可以法装留立萨斯配条在高级主型上的15扬声器的高级marklevin	



Example 2: Automobile Sales in China From Data to Event to Reports

5 CAR BRANDS WEEKLY TRENDS IN CHINA

Data Gathered from Weibo & Tencent Weibo [Mar 24 - Mar 30] 5款汽车微博每周趋势分析【3月24日~3月30日】 数据来源:新浪微博 腾讯微博



OUTLINE

- The Live Social Observatory
- The Technologies behind the System
- Next Generation Systems and Challenges
- Other Related Research
- Summary



Key Challenges

 As decided 5 years ago, the key challenges are Live, Big Data and 3M

What are they?



What is Big Data?



Big Data vs. Live

- Big Data is not about big, but about online and live
 - It is easy to handle (static) big data
 - But difficult to handle big live data streams
- Big Data Analytics
 - Data online and Live
 - Gather and Filter data (not Store, Sample and Search)
 - Feedback (close-loop) and React



Key Challenges in Live Social Media Analysis

ISSUES	KEY TECHNOLOGIES		
Live Data	How to gather "representative" live data from multiple sources w.r.t. an entity		
Multimodal	To analyze multimedia data (text, images, videos, locations and users)		
Multilingual	Deep NLP for multiple languages (English, Chinese,)		
Events	To detect and track all events happening around any entity		
Prediction	Predict future trends and detect hot posts / events		
Prescription	Generate comparative reports, user demography, and prescribe actions		
Automated	For scalability, the entire system should be fully automated		

Live Event Detection & Tracking from live social media streams

- Aims: monitor pulses of organization/ event/ location/ people:
 - What is going on; what is hot (Event detection)
 - Who is involved; who says what (User community detection)
 - Where are the users (Geo location detection)
 - What do users share in images/videos (Image Content Analysis)
 - Complain, admiration (Sentiment analysis)
- Main challenges
 - Huge amount of tweets filtering problem
 - Dynamic & noisy vocabulary conversational nature f social media
 - Dynamic user community
 - Media content analysis
 - Scalable distributed architecture



1. Gathering of Representative Data -1

- Key issue: How to deal with Live/Dynamic Data?
 - Some studies show that there will be 6 times more data if crawled in real-time
- 1st Source: Known accounts- Most organizations have substantial social media presence in building customer relations
 - Reliable but from Organization's point of views

Australia · http://www.vodafone.com.au



Vodafone Australia @Vodafone_AU Follow us for all the latest network news, product and service announcements, and special promotions. For help and support, please tweet @vodafoneau help.



 2nd Source: Dynamic Keywords- Generate dynamic keywords to address dynamic vocabulary issue

1. Gathering of Representative Data -2

- 3rd Source: Active Users- Identify dynamic user community
 - Needs to remove bot users



- Overall: Reliable Data Collection Strategy
 - Employ multi-faceted approach to gather representative data
 - Simulate real user login behaviors
 - Our strategy ensures high relevance, coverage and diversity



• Results: High recall, but too much noise

2. Mutilingual Linguistic Analysis (in English and Chinese)

- Key Phrase Extraction
- Named Entity (NE) Extraction
 - Identify name of people, organization, location and products etc.
 - May need finer grained NE's in vertical domains

The New York Philharmonic Orchestra will make a Person historic trip to North Korea in February, it has announced. Dominique de Villepin a été nommé Premier ministre Location ce mardi en fin de matinée par Jacques Chirac. Organization The orchestra's president and executive director, Zarin Date Mehta said it would play in the capital Pyongyang Time on February 26. In August, the reclusive communist country's Ministry of Culture sent an invitation to Title the orchestra at Lincoln Center in Manhattan. 朝鲜外务省发言人11月1日在平壤宣布,朝鲜将重返六方会谈,但前提条件是朝鲜与 美国在六方会谈框架内讨论解除美国对朝鲜金融制裁问题。针对朝鲜方面的动向, 各方均表示欢迎。美联社11月1日报道说: "长期以来一直拒绝与平壤进行直接对话

的美国总统布什认为,各方达成一致、同意恢复六方会谈应归功于中国的斡旋。



One Approach to Phrase Extraction

- Combine Short Text and Long Text
 - Short Text: Tweet, Weibo, Review from e-commerce website
 - Long Text: Wechat, News
- Procedure
 - Syntactic & semantic analysis
 - Filter potential bag of words (using LTP tools)
 - Keyword analysis
 - Generate word weights from different data source [Chen et al. ESWA2014]
 - Calculate Phrase
 - Keep top-score bag of words as phrase



Entity Extraction

- Prior Entity Database
 - Wiki, Baike (combine entity alias via link redirection)
 - Sogou Input lexicon database
 - Geographic entity database from Tsinghua University (Guoliang Li's group)
- Features
 - Word level: unigram, bigram
 - Part-of-Speech level: postag-unigram, postag-bigram
- Algorithms
 - Online learning algorithm
 - Semi-supervised learning algorithm



Multilingual Linguistic Analysis

- Entity Relations
 - Infer from co-occurrences of entities in messages
- Sentiment Analysis
 - Determine sentiment of posts at event, sub-events and entity level
- Classifier/ Filterer for noise
 - Employ text, social relations and temporal relations
 - To filter irrelevant contents
 - To filter noise





Filter Irrelevant Content

- Consider 4 Factors:
- Content: For each tweet, perform analysis to:
 - Informal language normalization
 - Irrelevant text tokens filtering
 - Identify evolving text features
- Location information
- User Social Relationships
- User Tweeting Tendencies Over Time
- Train a classifier based on training set using, say, kNN or SVM.
- For a new tweet <u>T_{new}</u>, use the classifier to assign the class(es) of <u>T_{new}</u>.

Noise Removal

- Identify Spam for Short Text
 - Filter out Tweet/Weibo which contain apparent commercial intent
 - Build a manually labeled dataset (10000+ labels)
- Features
 - Content-based: lexical pattern, part-of-speech pattern
 - Network-based: user follow information, retweet information
- Algorithms
 - KL divergence retrieval model [Qazvinian et al. EMNLP2011]
 - Random walk model [Li et al. IJCAI2015]



3. Event Detection (SIGIR' 2012)

- Live Event Detection and Tracking
 - Extract named-entities in text messages, and visual concepts/ logos in associated images
 - Filter noise using a learned classifier
 - Cluster remaining (multimedia) message into events using an incremental clustering algorithm
 - Further divide the detected events into emerging and evolving events
 - Identify influence/active users and update evolving keywords
- Key research Issues:
 - How to do it Live??
 - How to extract relationships between events
 - and between events and users

C SUB TOPICS				
Bjp,Seat,Give,Ticket,Sir st please also give ficket to BJP candidate from sitamarki (Bihar) the birth place of maa janaki (don't give this seat 2 ally	Bjp,Follow,Tweet, Twitter,Rss Here is not even a single symbol of B/B/B(but) in your hylitectover , being from RSS we should not follow "Vyakti pooja"		Constant Con	Take, Sir, dus, Frm if for great or the soliders piz te action t naxais
	Bjp,Varanasi,Contest,Poll,S eat BJP's Narendra Modi to contest from	Kejriwal,Arvind, Gujarat,Question ,Development	Media,Aap,C ongress,Say, Kejriwal	Win,Election, Wish,Sir,Guja rat
Sir,India,Pm,Wish,Hope	MM Joshi's Varanasi seat in Uttar	to the 16 Questions asked by Arvind		you sir and wish you all the best
sir,l hope u r the next PM of India sir.	Bjp,Seat,Varanasi,Contest, Candidate	Kejriwal to Narendra Modi	No Congress - No AAP - YES	for upcoming lok sabha elections. You
	BJP's prime ministerial candidate Narendra Modi to contest from	India,Pm,Candidate, n	Party,Electio	Speech,Know, Support,Bjp,T hackeray
Vote,Congress,Bjp,Party,Time	Rally,Address,Shri,Delhi,O			
MY vote for mr narendra modi bjp, aap has become paap, and congress is sell made secular, rest partys another face of CONGRESS	disha Today WATCH LIVE: Shri Narendra	Rahul, Gandhi, Gujarat, Turf, Educati on		Narendra Modi, but his Sena may hurt BJP

4. Predictive Analytics -1 (SIGIR' 2012, 2013)

- Identify hot emerging events:
 - Aim: to detect such events before they become viral



- Has high rate of: User #, Message #, Re-tweet #, cumulated msg influence weight
- Has high level of overlap between: org key users and topic key users, org keywords and topic influence keywords
- 6 features for classifier with high accuracy:

$$f_{1} = \frac{|U^{t}|}{\sum_{x=0}^{t} \frac{1}{t-x+1} |U^{x}|} \quad f_{2} = \frac{|Tw^{t}|}{\sum_{x=0}^{t} \frac{1}{t-x+1} |Tw^{x}|} \quad f_{3} = \frac{|Rtw^{t}|}{\sum_{x=0}^{t} \frac{1}{t-x+1} |Rtw^{x}|}$$
$$f_{4} = \frac{\#(ku_{tp} \cap ku)}{\#ku_{tp}} \quad f_{5} = \frac{\#(kw_{tp} \cap kw)}{\#kw_{tp}} \quad f_{6} = \frac{|A^{t}|}{\sum_{x=0}^{t} \frac{1}{t-x+1} |A^{x}|}$$



4. Predictive Analytics -2 (SIGIR' 2012, 2013)

- Framework for Viral Message and Event Detection:
 - Extract live social signals for early cues of trending events
 - Track multiple indicators: involvement of influential user, number and increase in msgs and re-tweets..
 - Identify also users likely to participate in diffusion process
 - Classifier for viral events/messages
 - Identify hot emerging events/ messages before they become viral
- Influential/ Active User Detection:
 - Characterization of influential and active users
 - Identification of mid-range influence users
 - Leading to prescriptive actions



5. Prescriptive Analytics From Data to Review/ Report

- Converting Data into Actions:
 - Translating past, present, and future data into actionable items for better user engagements
- Reviews and Reports:
 - Smart linguistic program that aggregates social media sentiment and overall opinions about an entity
 - Generate comparative reports based on key indicators
 - Identify strengths and weaknesses for each entity
 - Towards prescribing (current and future) actions to improve image and user engagements



5. Prescriptive Analytics - 2 From Data to Review/ Report

- From Data -> Analysis -> Predictive -> Prescriptive Analytics
 - Towards analysis of "why" and "how"...
 - Work on characterization of problem, and solution to alleviate the problems



6. Information Organization: from unstructured UGCs to Structured Knowledge

- Mining structure of data from multiple UGC sources
 - UGC sources includes: Wikipedia, Blog, cQA, Forum and Twitters, or equivalent




Information Organization The Framework



Key Research Focus: Topic Relations Identification

- Sub-topic relation: $R(t_a, t_b)$ indicates that t_b is a sub-topic of t_a
- We use multiple evidences to estimate probability $p(R(t_a, t_b))$ of both directed and undirected relations

$$p(R(t_A, t_B)) = \left(\sum_{e_k \in E_{direct}} w_k \cdot e_k(t_A, t_B)\right) \cdot \prod_{e_s \in E_{undirect}} e_s(t_A, t_B)$$



Information Organization: Topic Hierarchical Generation



Algorithm 1 Hierarchy Generation Algorithm Input: T: the candidate topic term set; **Output:** R_{ret} : the sub-topic relation set of the resultant hierarchv: T_{ret} : the topic set of the resultant hierarchy; 1: Initialize $T_0 = \{\mathcal{C}\}, R_0 = \phi$ 2: for $i = 1 \text{ TO} \infty \text{ do}$ $t \leftarrow \text{selectTermFrom}(T - T_{i-1})$ 3: if t is NIL then 4: $R_{ret} \leftarrow R_{i-1}$ 5: $T_{ret} \leftarrow T_{i-1}$ 6: 7: break8: end if $T_i \leftarrow t \cup T_{i-1}$ 9: $R_i \leftarrow R_{i-1}$ 10:for all t_k IN T_{i-1} do 11: 12:if edgeBetween (t_k, t) exists then $R_i \leftarrow R_i \cup \text{edgeBetween}(t_k, t)$ 13:end if 14:end for 15: $edgeWeighting(R_i)$ 16: $R_i = \text{hierarhcyPruning}(R_i)$ 17:18: end for

Resultant Hierarchy

Evolution of a Hot Topic over time



Application in Multimedia QA

- OneSearch: Multimedia/multilingual question-answering
 - Long term research: turning social QA into dynamic knowledge structures
 - Focus on returning timely multimedia answers



Overview of Information Organization Unstructured to Structured Knowledge

• Towards knowledge structures/ graphs for different domains:

- Information and relation extraction
- Multi-source knowledge integration
- Domain: Entertainment & Wellness (Critical Illnesses)





7. Scalable Infrastructures



Back-End Tech Highlights

- Fully automated
- Cloud-based
 - 80 VMs
- Big data
 - >100 TB, 2.6 billion records
- Live + Batch Processing
 - Storm + Hadoop + Hbase
- Distributed
 - Crawlers, Database, File Storage, Processing
- Cross indexing
 - Elasticsearch index for text
 - Own hash-based image index
 - Graph index

8. Large Scale MM Content Analysis (ICMR 2014, ACM MM 2014)

- Information is becoming increasingly multimedia
 - >30% of microblogs have images/videos
 - Over 40% of such microblogs do not have relevant text descriptions
 - Purely text based approach is inadequate
- Key problems: How to identify images relevant to entities/topics without relevant texts
 - O Huge amount of social-tagged mm content available
 - O Earge semantic & intension gaps in mm contents
 - Classifiers to detection certain classes of visual concepts, logos are effective







Large-Scale Social Video Analysis Deep Learning Architecture for Fusing Multiple Cues



Observation 1: Actions, objects and scenes compose the semantics of videos → We devise a multi-channel deep NN.

Observation 2: Direct modeling of these semantics are not effective → We extract more higherlevel semantics that are more

easier to be fused.

Large-Scale Social Video Analysis Deep Learning Architecture for Fusing Multiple Cues

- Experimenting with Multi-modal deep learning architecture
 - combining object, scene, motion and text features
- Automatically discover mid-level attributes
- Used in benchmark classification tasks
 - Out-performed existing state-of-the-arts features/systems on image annotation and video event detection
- Towards continuous learning system



Example of Attributes Discovered (Airplane)



Example of Attributes Discovered (People)



Content Representation via Embedding Learning Features from Images, Tags & Users

Deep Semantic Relation Embedding Architecture



mAP% of the proposed features (last two) as compared to state-of-the art on benchmarks

NUSWIDE 38.6 32.7 41.3 42.9 CCV 62.5 58.4 67.2 67.8	Dataset/Method	DeCAF	ICMAE	Ours-fc7	Ours-4kSc
CCV 62.5 58.4 67.2 67.8	NUSWIDE	38.6	32.7	41.3	42.9
	CCV	62.5	58.4	67.2	67.8

Content Representation via Embedding Examples: Image-Image Embedding





Content Representation via Embedding Examples: Word-Image Embedding





Content Representation via Embedding Examples: Image-Word Embedding



beach, girl, ocean, sand, water, playing, white, bikini, chair



soccer, field, game, players, zijn, groen, running, roelof, januari, ball



bike, girl, street, yellow, car, white, woman, boy, man, dog



wedding, wearing, dress, white, girl, red, black, pink, dressed, flower



Content Representation via Embedding Examples: Word-Word Embedding

Dog	foster, sleeping, service, cute, cat, puppy, stray, freshfields, spaniel, pug	Girl	dress, cute, hat, pink, boy, portrait, baby, shirt, birthday, blonde
Iorning + Drink	drink, champagne, sky, juice, drinks, coffee, foggy, cold, soda, sun	Evening + Drink	evening, drink, wine, beer, bartender, alcohol, beverage, martini, liquor, bottle

Table 1: Spearman correlation of 971 word-pair similarities computed by different methods and MEN human judgements. Our method learned from SBU is very close to Word2Vec learned from Google News.

Method	Word2Vec	S-CNN	Ours
MEN	0.65	0.36	0.64



Large-Scale Social Image Annotation Mining Curated and Weibo Images

- Extend to social images from different types of social networks
 - Classifying curated images from Pinterests
 - Identify brand images from Weibo



http://www.nextcenter.org/Brand-Social-Net/

Deployment of Visual Technologies -1 Variety of Visual Technologies



Deployment of Visual Technologies -2 Some vertical domain visual search engines

Company	Appl ⁿ	Technologies	
Snapfashion	Fashion find Apps	Similarity SearchApp development	
Superfish	Visual search API for image	 Similarity Search 	
Cortica	Contextual Ads Serving for image	 Matching-based recognition 	
Face++	Face, in finance & surveillance	Similarity searchObject recognition	
Dextro	Visual recognition for	 Visual analytics for images/videos 	
ViSenzeVisual Search & recognition API for image and video		 Similarity search Domain ontology Object recognition 	

ViSenze: A Startup Company under NExT (www.visenze.com)



Size of Company:

- 30+ Employees
- 22 Technical (7 PhDs)
- Support over 30 million paid images
- Moving to USA and Europe

Key Customers:

(over 50% of key sites in SE Asia and South Asia) Flipkart, Rakuten, Zalora (Rocket), Lazala (Rocket), Reebonz, Patsnap Download Flipkart App from Phone store

VISUAL POWERED E-COMMERCE SITES









Snap Search Buy!

Shazam for Fashion Commerce !



Making Visual Content Targetable and Relevant to Advertisers





Contextual Advertising in Videos

Detect and recognize brands/logos in videos to associate with relevant Ads or Recommendations **In-Video Recognition**





10

Use cases for Visual Search & Recognition

- Offline to Online
- Online to Offline
- Large-scale database retrieval
- Visual recommendation
- Supports multi-categories:
 - Fashion & related
 - Packaging
 - Home décor, furniture
 - Diagrams/drawings
 - And MANY more



Deployment of Visual Technologies -3

- Recent products are targeting at mostly 5 vertical domains:
 - Fashion
 - Home furnishing
 - Products
 - Surveillance
 - Finance



Snap Search Buy!

Shazam for Fashion Commerce !

- Highlights of most current deployments:
 - Most are based on image matching & annotation technologies
 - Visual recognition has just started
 - Some impressive recognition systems are crowd-sourced-based
 - Hardly anyone offer real-time in-video products

OUTLINE

- The Live Social Observatory
- The Technologies behind the System
- Next Generation Systems and Challenges
- Other Related Research
- Summary



A Recap: Key Characteristics of Current Generation of Social Media Systems

- Live, Big-Data and 3M (Multi-source, Multimedia & Multilingual)
- Support users' need for communication, sharing and interaction
- Support co-viewing and co-creation of contents (by users and systems)
- Develop social analytics aim to understand contents, events and users - targeting at recommendation



1. Image/Video handling

Top 3 recent social media platforms are all image/video centric





 Live video and online devices common place







Visual-based Platform: SnapChat



 Snapchat: founded in 2011; Sharing 10-sec video moments that erases itself after certain period; 70% of users <24 yrs old (~100M active users); user base will be bigger than Twitter soon

16

- Temporal Real Time Messages: people decide how long others view their photos/videos
- Share live stories with friends
- "Discover" news from top sources, eg CNN.
- See also Vine: 6-sec loop-videos



Periscope



- Similar to SnapChat in supporting live broadcast from mobile phones
- Users can broadcast and watch any other live broadcasts ۲
- Why it is hot??
 - Support mobile-based broadcast
 - Track events live
 - Interactive
 - Integrate with Twitter Social Network



Broadcast Video LIVE from Anywhere

2. Live location analytics

- Sensor devices are everywhere capable of multi-form of sensing
- Multi-source information: location traces, POI and audio
- Towards better location estimation and mobility analysis





3. Live aspect is central now

- With Live, comes the ability for continuous sharing and feedback..
- Users want instant feedback from friends and systems
- Live videos are commonplace
 - Shoot, share, and feedback, all live
 - Many examples: home automation, SnapChat, Periscope
 - Problems in understanding and reconstructing the interesting places visited and/or events happened







4. Quality and structure of data

- Deteriorating quality of data, with about 70% of UGC's belongs to noise/ spam/ rumors category
- Representativeness of data is important, especially for large dynamic topics
- Key approach to making data usable is to structure them at both data and knowledge level



5. Co-Creation & Co-Invention

- With live instant feedback, comes the possibility to co-create and co-invent
- Not just the contents, but systems and design


Co-Inventing the Future -1

GoPro points towards future of Consumer Electronic Products

- Products online from Day 1
- Create social communities
- Feedback to product design
- Bootstrap new products and applications
- Another example of intelligent appliances: Home Air Filters
 - Filters online from Day 1
 - Understand users' living habits
 - Auto order when filters break down
 - Extendable to any home appliances
 - Foundation for intelligent home



Co-Inventing the Future -2

Product Design:

- Building a dedicated user communities
- Leverage opinion leaders in communities to feedback on deign of systems/products (hardware and software)
- Offer new design through communities to influence more users
- Government policies
- Early Stage Examples



What Has Changed in Last 5 Years?

6. Online & Offline Integration

- Online data represents certain segment of population
- Offline data often represent the majority, if available
 (However, other than ecommerce, offline data is often limited)
- Integration of both provides full picture of the world
- Examples:
 - E-commerce social media buzz; need offline data on actual sales and online sites that cater to buyers
 - Election of Singapore social media and actual voting pattern is totally different



Big Data Revisited

Big Data Analytics

- Data online and Live
- Gather and Filter data (not Store, Sample and Search)
- Feedback (close-loop) and React
- Big Data is not about big, but about online and live
 Through user and systems feedback, to realize better user experience

Towards C2B

- Alternatives to B2B and B2C, where business takes central roles
- Customers now take central roles in influencing business



Key Technologies

- Predictive Analytics
 - Predict users preferences and dislikes
- Prescriptive Analytics
 - Analyze user communities and identify local communities
 - Analyze user feedbacks, identify needs/ concerns
 - Formulate actions



OUTLINE

- The Live Social Observatory
- The Technologies behind the System
- Next Generation Systems and Challenges
- Other Related Research
- Summary



Multimodal Location Estimation Combining Text, Visual, Location & Audio Analysis

- Given a series geo-tagged tweets, how to map them to actual POI (Point of Interests)
 - Each geo-tagged locations may be mapped to 100s of venues (as defined in, say, 4Square)
 - Sampled audio signals to estimate venue categories (for elder-care applications)



- Results using Tri-modal Deep learning architecture (Timber, Rhythm, Pitch) > 67%
- Classifier to estimate exact venue



Location Analytics

- Mine relations between check-in venues
- Identify landmarks of local venues
- Identify popular trials (of individuals and their friends)
- Analyze user demographic and interests communities
- Generate multi-faceted user relation graph



Travel paths and flows of users



Interest community of food lovers in Singapore



Overview of Computational Wellness

ClealthSenze gathers & organizes big-data from multiple sources



Towards D2D (Doctor-to-Doctor) Network A Form of Social Life Network



- Able to reach hundreds of millions of patients at various Locations

Wellness Advisory System for End Users

ClealthSenze



Advisory System: based on 6 critical illnesses



End Users

- For end users
- Offer advises on appropriate activities and diets based on current conditions

OUTLINE

- The Live Social Observatory
- The Technologies behind the System
- Next Generation Systems and Challenges
- Other Related Research
- Summary



Summary

- We have built a live social observatory system:
 - Work on large-scale real-world problems of impact
 - Analyze large scale live UGCs and SDCs, and possibly MGCs
 - Monitor events and happenings in city to help users lead better life
- Research under NExT Research Center:
 - Over 40 researchers at all time
 - 2 Startup companies
- ViSenze Pte Ltd
 - 30+ staff
 - Mobile visual search
- 6Estates Pte Ltd
 - 10+ staff
 - From unstructured data to actionable knowledge



Summary

Many Challenges:

- Deeper analysis at entity and relation levels
- Multimodal analysis at various levels
- Real-time analysis and inference
- Prescriptive analytics
- Links to social understanding of events
- Privacy and Transparency

0

Towards Innovative Applications

- Social wellness
- Smart city, E-government
- Education..



THANKS

Visit our Web Observatory: http:///WWW.NEXTCENTER.ORG/

We have openings:

- Interns, Research Staff, and Visiting Research Staff
- For NExT Research Center and Companies