社交网络时代中文信息处理新机遇

黄河燕
北京理工大学计算机学院
北京市海量语言信息处理与云计算
应用工程研究中心

2012.4

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

纲要

中文处理

I 社交网络时代背景

II 中文信息处理的新挑战新机遇

III 以“人”为本的计算

IV 几点思考

社交网络发展现状

- 今年全球人口中的五分之一将会使用社交网络（13.4亿人），到2014年将达到四分之一。
- 世界著名的社交网站Facebook的全球用户总量超过了8亿，并很可能在8月前达到10亿用户。facebook预计最高可以募集到的资金预计为100亿美元左右，有可能成为有史以来最大宗的IPO案之一。有互联网产业最近报道：Facebook的上市标志着Web 2.0时代已经结束，社交网络将成为新的王者。

市场研究公司eMarketer于2012年3月发布的《世界社交网络使用：市场规模与增长预期报告》

社交网络发展现状

2011年上半年，我国微博用户数量从6311万迅速增长到1.95亿，半年新增微博用户1.32亿人，增长率高达208.9%（增长最快的应用），在网民中的使用率从13.8%提升到40.2%。

新浪微博

腾讯微博 t.qq.com

搜狐微博 t.sohu.com

网易微博 t.163.com

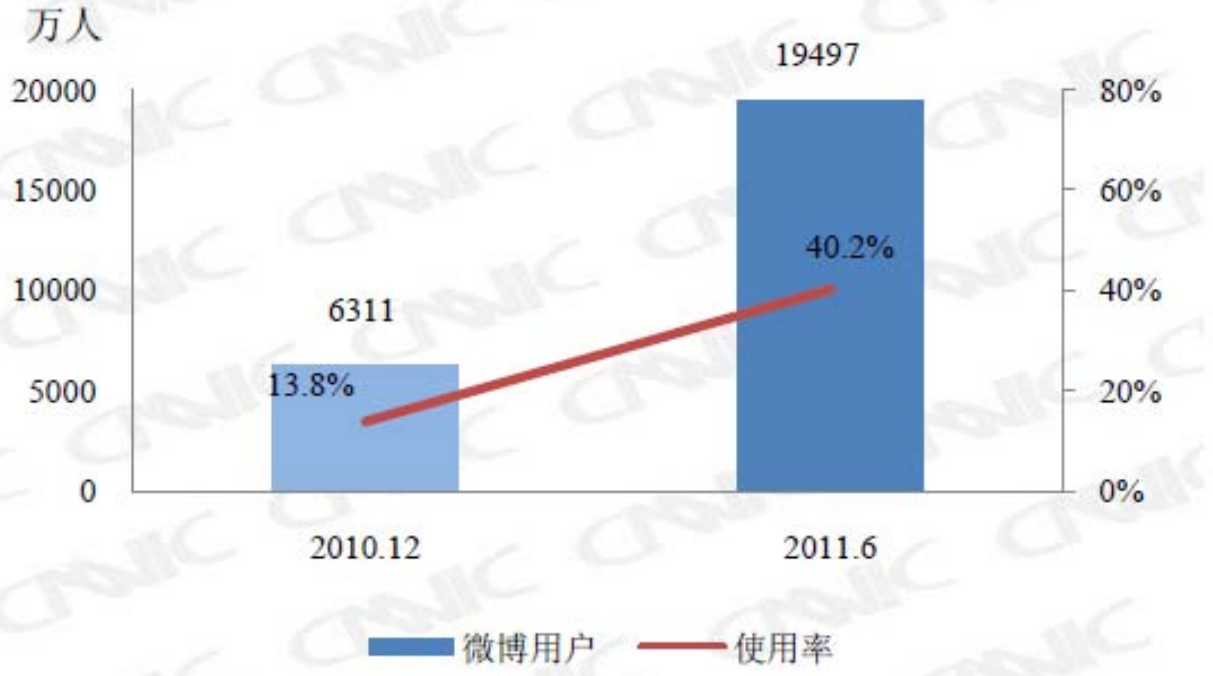


图 26　2010.12-2011.6 微博用户数及使用率

社交网络发展现状

研究背景——微博改变网民的上网习惯

即时通讯工具

电子邮件

微博

20%

搜索引擎

网址导航站

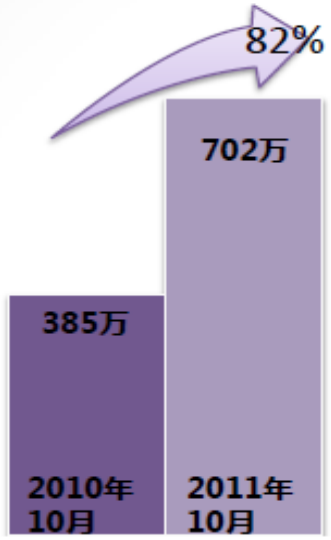
……

网民上网第一站登录微博的比例达到将近20%，直逼即时通信工具和电子邮件。

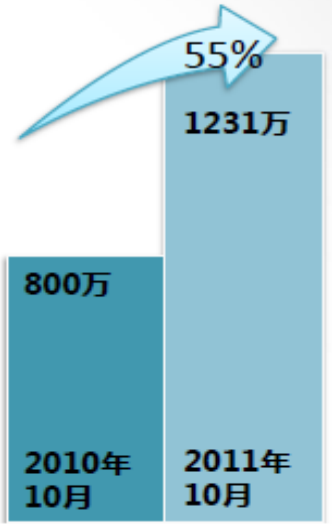
——《2011年社会心态蓝皮书》

社交网络发展现状

微博对门户网站流量贡献度在逐步增强

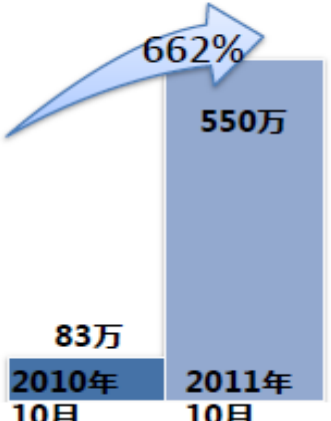


导入用户数量



导入页面浏览量

信息爆炸，微博外链条数的增长速度远远大于微博外链用户数



微博外链被点击条数

社交网络发展现状

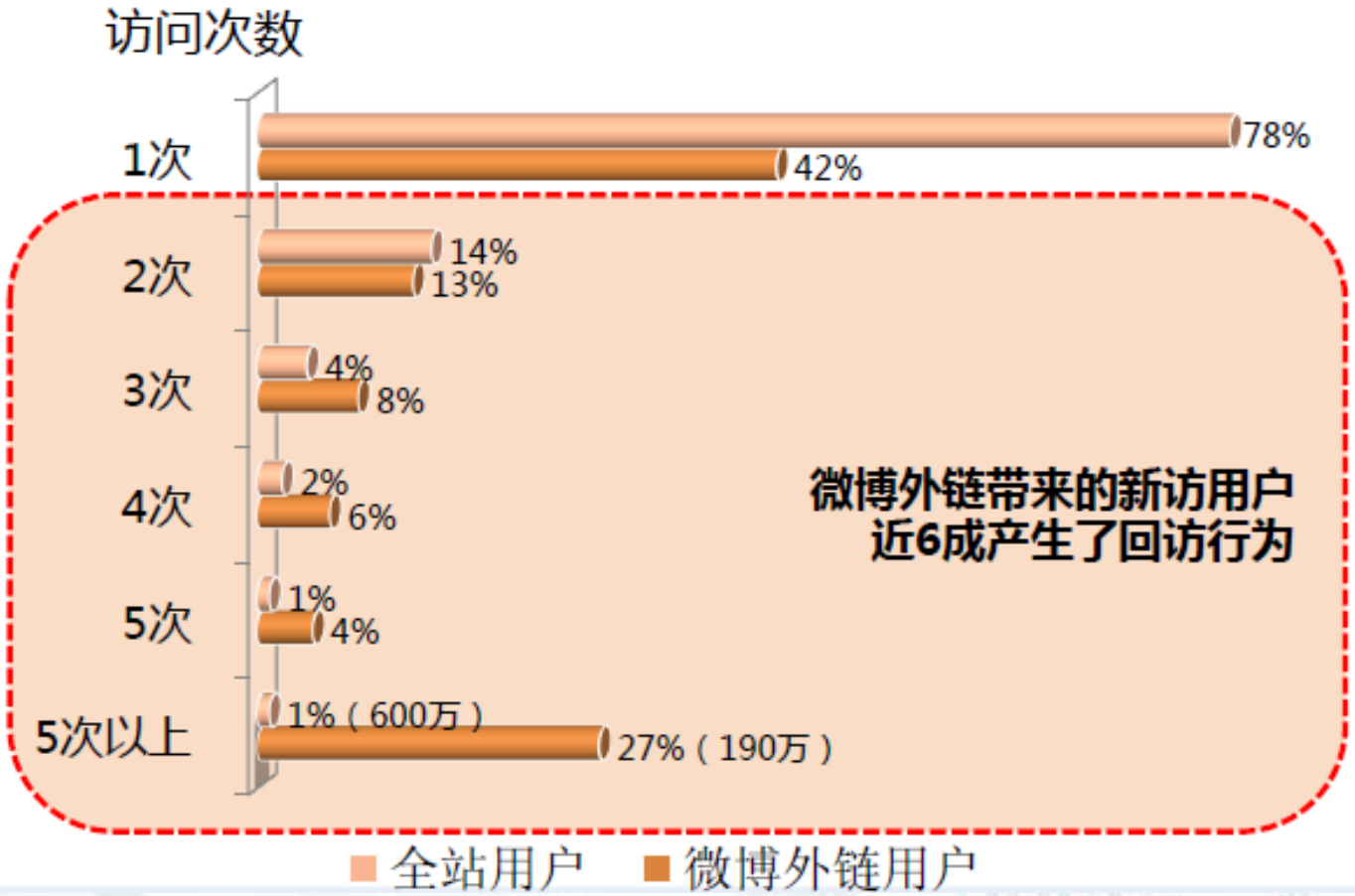
微博导入用户为网站的优质用户

微博外链用户VS全站用户访问粘性指标对比

指标	微博外链用户	全站用户
平均访问停留时长（秒）	402	374
访问频率（次/人）	4.3	3.1
访问深度（页/次）	7.1	6.2
用户活跃度（页/人）	30	19

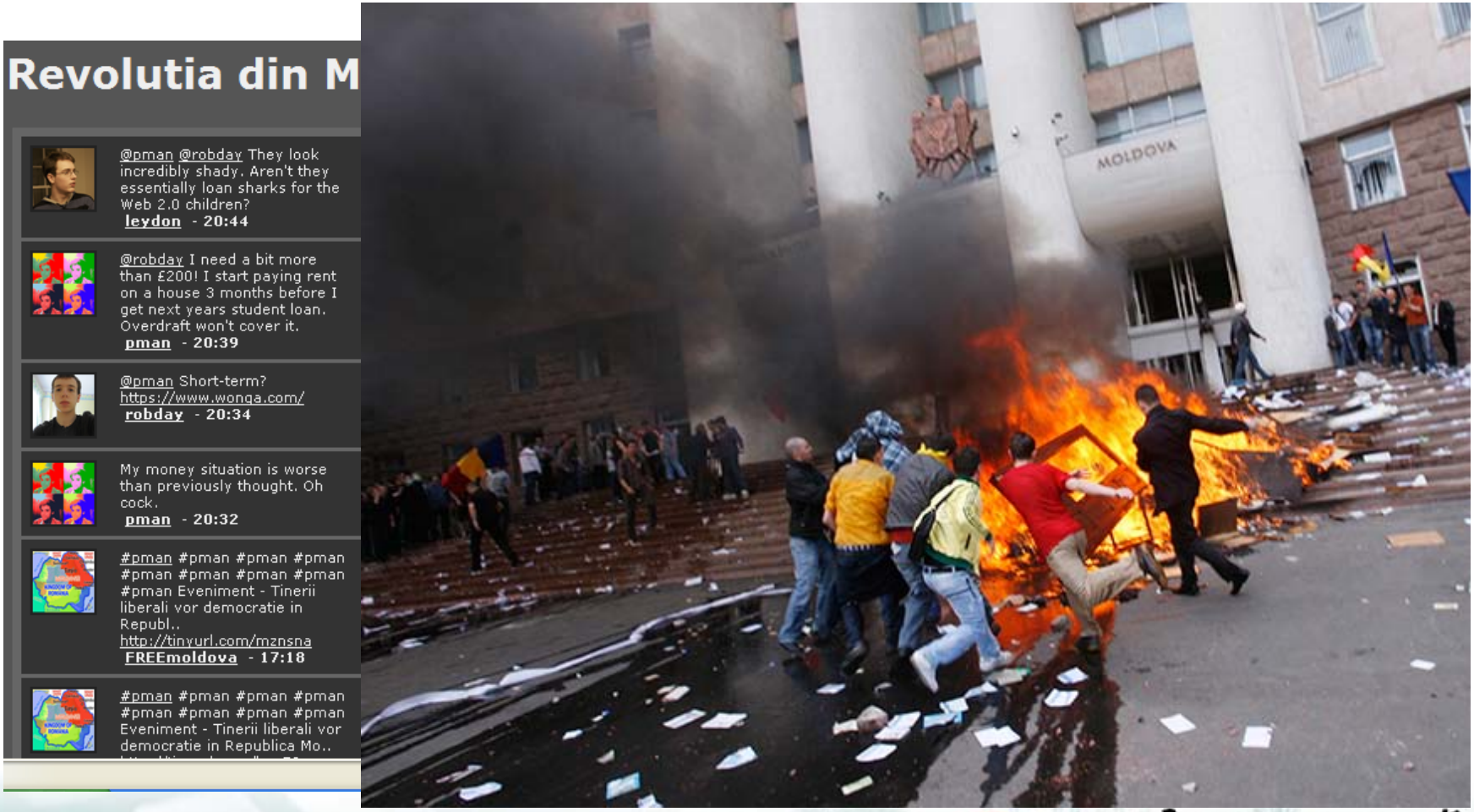
社交网络发展现状

附图：10月微博来源用户产生的回访用户数分布



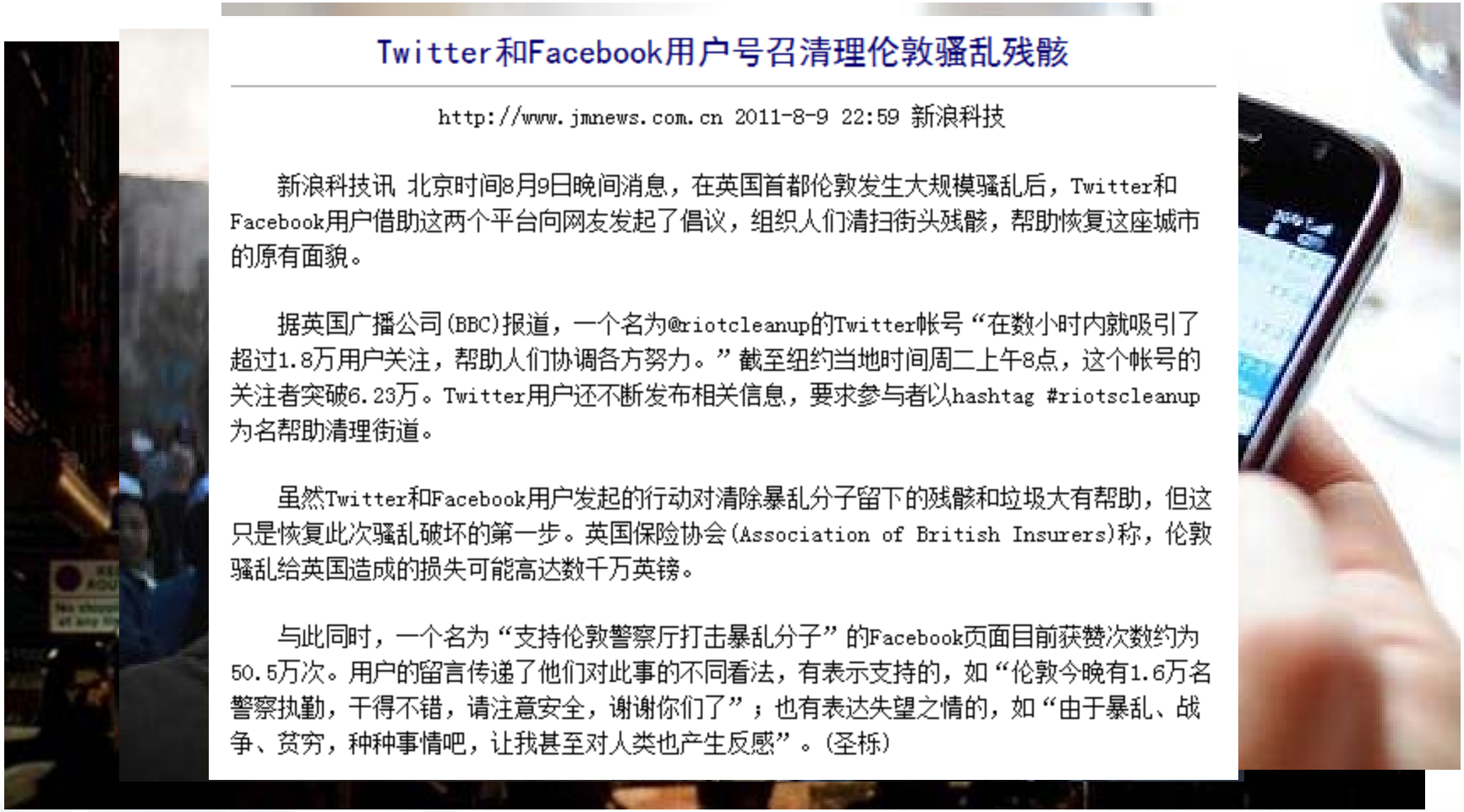
社交网络发展现状

- 2009年 摩尔多瓦 Twitter革命，阿拉伯之春



社交网络发展现状

2011 伦敦骚乱



Twitter和Facebook用户号召清理伦敦骚乱残骸

http://www.jmnews.com.cn 2011-8-9 22:59 新浪科技

新浪科技讯 北京时间8月9日晚间消息，在英国首都伦敦发生大规模骚乱后，Twitter和Facebook用户借助这两个平台向网友发起了倡议，组织人们清扫街头残骸，帮助恢复这座城市的原有面貌。

据英国广播公司(BBC)报道，一个名为@riotcleanup的Twitter帐号“在数小时内就吸引了超过1.8万用户关注，帮助人们协调各方努力。”截至纽约当地时间周二上午8点，这个帐号的关注者突破6.23万。Twitter用户还不断发布相关信息，要求参与者以hashtag #riotscleanup为名帮助清理街道。

虽然Twitter和Facebook用户发起的行动对清除暴乱分子留下的残骸和垃圾大有帮助，但这只是恢复此次骚乱破坏的第一步。英国保险协会(Association of British Insurers)称，伦敦骚乱给英国造成的损失可能高达数千万英镑。

与此同时，一个名为“支持伦敦警察厅打击暴乱分子”的Facebook页面目前获赞次数约为50.5万次。用户的留言传递了他们对此事的不同看法，有表示支持的，如“伦敦今晚有1.6万名警察执勤，干得不错，请注意安全，谢谢你们了”；也有表达失望之情的，如“由于暴乱、战争、贫穷，种种事情吧，让我甚至对人类也产生反感”。(圣栎)

社交网络是重要的国家战略资源

- 美国国防部长罗伯特·盖茨2009年6月表示：Twitter等在伊朗德黑兰抗议活动中起到重要作用的社交网络是“美国的重要战略资产”。
- 美国当局在近三年中拿出超过2000万美元的“竞争性赠款”，在2011年追加2500万美元赠款，“以支持正在利用尖端手段对抗互联网压制行为的新涌现的技术人员和活动人士群体”。

纲要

中文处理

I 社交网络时代背景

II 中文信息处理的新挑战新机遇

III 以“人”为本的计算

IV 几点思考

社交网络时代的中文信息处理意义

- 虚拟的社交网络和真实社会的交融互动对社会的直接影响越来越大，直接影响国家安全与社会稳定，事关各国的国家战略安全。
  - 借助社交网络发布和接收信息的简便性，人人都有了网络话语权，各类涉及到国计民生的话题和观点可以随时发布，信息一旦发布就能通过“核裂变”的方式传播扩散，期间经过意见领袖的放大作用，促使具有相同观念和诉求的虚拟社区快速形成，并在线下快速组织并发动群众参与到社会活动中，有可能构成社会动员力。
- 社交网络已成为人们生活的一部分，深深地影响了人们生活的方方面面，对整个国家和社会的信息获得、传播、思维和生活产生不可低估的影响。
  - 2011年2月，美国企业网络Merchant Circle对美国8,456家小型企业的随机抽样调查分析显示，70%的当地企业使用Facebook用于市场营销；有近40%的受访者表示，他们使用Twitter用于市场营销。
- 社交网络是商业的未来，也是未来的商业

社交网络时代的中文信息处理需求

- 社交网络时代对中文信息处理提出了更广阔更深入的技术需求：
  - 社交网络消息的自然语言处理与内容挖掘；
  - 网络个体的个性特征建模；
  - 社交网络的分类与聚类；
  - 行为模式与群体行为挖掘;
  - 大数据量的存储、分析与挖掘。

中文信息处理的新挑战

- 信息内容短小（微博140字）但规模巨大（每天6亿条）；
- 内容不规范，语言口语化严重，且有上下文背景；
- 信息快捷，稍纵即逝；要求在线处理；
- 信息多维复杂，涵盖主体基本资料、内容信息、用户社会行为等综合信息；
- 需要涵盖自然语言处理、社会网络分析、传播网络建模等诸多过程。

传统的中文信息处理渐入困境！

NLP：自然语言处理？身心语言程序学

每天用点心理学-湖南NLP学院：你要相信”当下你的选择，一定是你能做出的最好选择“我们做的任何事情，都是为了满足自己的一些需要。在那些特定的环境里，也许你事后会后悔自己当时的选择，但其实当给多你一次机会重头来过，你还是会做同样选择，因为那是你在当时的最好。想让自己学会好的选择吗？NLP可以告诉你

5分钟前　来自 皮皮时光机　转发 | 收藏 | 评论

中微子u：//@李方涛2011: 之前有NLP的ACM Fellow吗

@刘知远THU：今年ACM Fellow揭晓。http://t.cn/SqgiC0 其中Dan Roth (UIUC)和Amit Singhal (Google)是与NLP和IR相关的，关注。

12月9日01:04　来自 新浪微博　转发(5) | 评论(2)

20分钟前　来自 微博搜索　转发 | 收藏 | 评论

信诚人寿-冯艳：当我以为最年轻的NLP执行师在我们班时(18岁),花美女说她们班上有个16岁的。。。。嗯!这么早接触NLP真好!

43分钟前　来自 UC浏览器　转发 | 收藏 | 评论

工大學 OF TECHNOLOGY

纲要

中文处理

I 社交网络时代背景

II 中文信息处理的新挑战新机遇

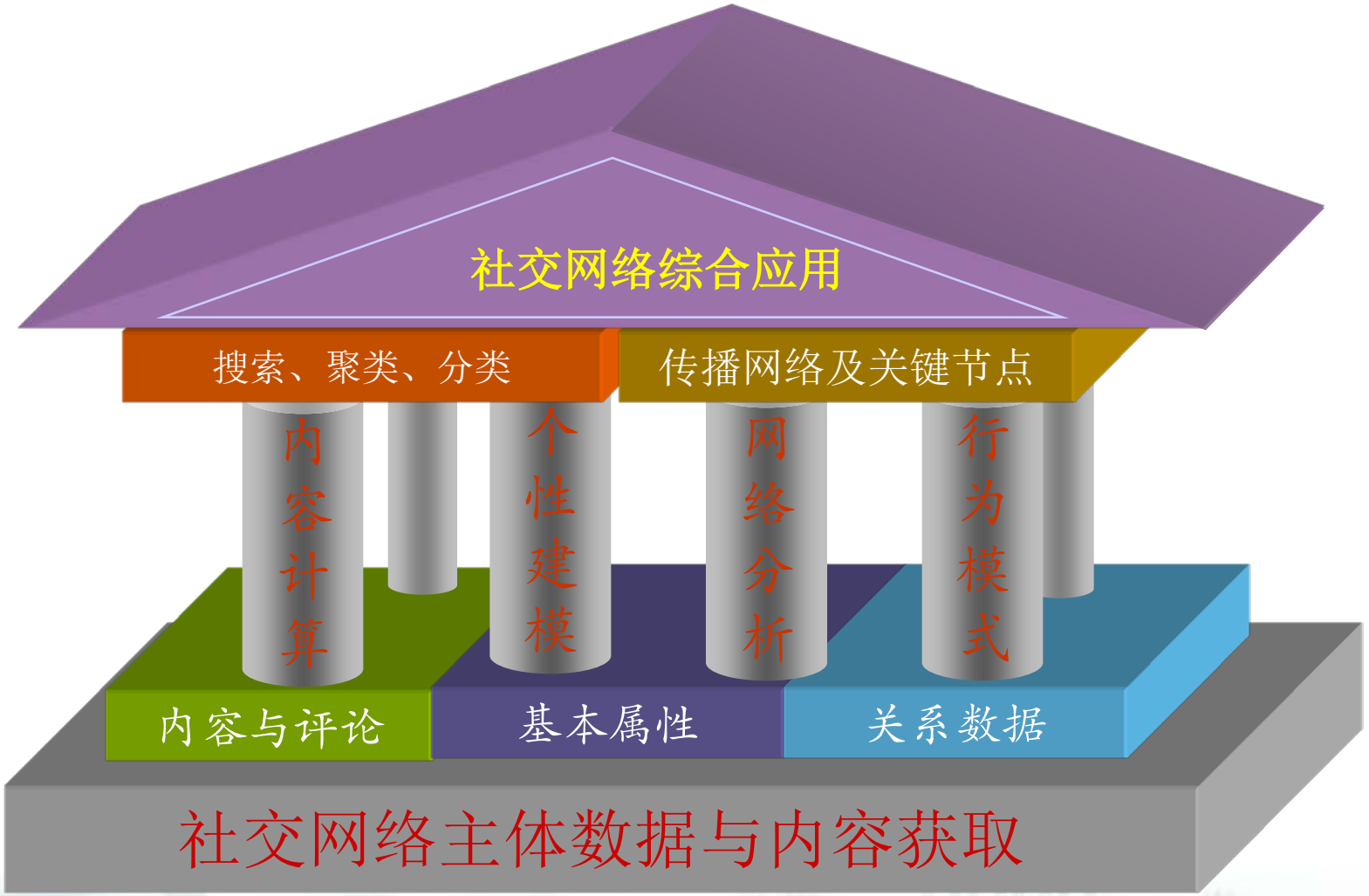
III 以“人”为本的计算

IV 几点思考

社交网络中文信息处理微革命：以人为本

社交网络信息综合计算与应用

社交网络中文信息处理微革命：以人为本



面向微博的社交网络研究

- 微博获取
  - 微博主体及内容的定向与元搜索采集
- 微博主体个性建模
  - 微博博主个性化建模及应用
  - 密切关系网挖掘
  - 微博情绪感知
- 微博行为模式分析
  - 微博博主行为模式挖掘-微观上
  - 微博主题传播网络及关键节点分析
- 微博综合应用
  - 基于微博的股票多空判别
  - 微博宏观统计分析与挖掘-宏观上的
  - 微博的多模式自动分类
  - 微博舆情监测与疏导
  - 微博广告推荐
  - 微博综合态势推演与群体情绪指数

纲要

中文处理

I 社交网络时代背景

II 中文信息处理的新挑战新机遇

III 以“人”为本的计算

IV 几点思考

几点思考

中文信息处理新的定义和理解：

- 特点：
  - 信息规模海量化
    网络环境：因特网、物联网，获取途径多样
  - 信息产生与传递的群体性
  - 信息使用方式的自然交互性
  - 信息处理的实时性
  - 信息应用的更趋智能化

几点思考

未来中文信息处理的着眼点：

- 海量化：云模式实现
- 应不仅局限于对文字内容的分析处理，与物联网的融合
  背景+情景+个体感知
- 以人为本的社会计算：深度语义分析计算
- 基于多模态信息的脑认知的语言理解

几点思考

可能的重要原始创新

- 信息表示的基础理论：多模态感知信息表示、语义知识的有效表示
- 信息感知与获取：基于物联网和现有Web资源，自动收集TB级的信息资源库，基于云计算架构和信息质量评估，研究新型不限元的语言模型建模方法，提供海量知识库的支撑；
- 信息认知与处理：如社交网络的综合计算模型；从超大规模的UGC(用户产生内容)数据集合中，研究提炼知识与情报的基础理论与工具等；
- 信息的交互与应用：多模态自然的人机交互接口、各种特性化、个性化的信息服务。

几点思考

产业发展方向：

- 中文信息处理技术应用于社交网络，采用云计算架构，打造与Google/baidu等量的信息服务，如社交网络的个体搜索、个性化服务与推荐、云语音、云翻译等。
- 中文信息处理核心技术中间件化，采用Oracle等结构化数据分析厂商的模式，并形成快速应用原型，可方便中文信息处理技术的产业化推广与应用；
- 结合行业特点定制中文信息处理技术，可降低语义处理的难度和复杂度，如新闻、娱乐、教育、邮政、气象等行业。

产学研战略合作模式：

- 不同性质组织机构的合理定位与分工协作
  优势互补、合作多赢
- 资源平台适度的开放共享，减少重复投入
- 合作模式灵活多样：
  合作研发、技术兼职、技术授权、技术入股、直接经营

几点思考

提升中文信息处理在国家重大科研计划中地位：

- 围绕国家重要战略需求，根据国际研究发展现状，研讨制订领域发展战略；
- 依据发展战略和国家重大科研计划的不同特点提出相关的指南建议；
- 加大对领域内有显示度的成果和成功产业化应用示范的宣传，团结、联合，扩大学会及本行业在社会及相关部门的影响；
- 不同专家学者，不同场合的集体呼应；
- 根据科研基础及研究力量进行适度的战略布局与协作，大力提倡同行的互相沟通、支持与配合。

Big Data is a Big Deal!

Big Data is a Big Deal

Subscribe

Posted by **Tom Kalil** on March 29, 2012 at 09:23 AM EDT

E-Mail | Tweet | Share | +

[Editor's Note: Watch the live webcast today at 2pm ET of the Big Data Research and Development event at http://live.science360.gov/bigdata/]

Today, the Obama Administration is announcing the "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning.

To launch the initiative, six Federal departments and agencies will announce more than $200 million in new commitments that, together, promise to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data. Learn more about ongoing Federal government

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

结语

- 奥巴马政府先期投资2亿美金，启动“大数据研究与开发促进计划”，用于提高从大规模数据集中提取知识与商业智慧，用来加速科学与工程的探索历程，加强国家安全，优化相关的教学。
  - 社交网络是迄今最大最实用的实时在线大数据集合；
  - 中文信息处理是知识挖掘的基础工具；
  - 社交网络时代的中文信息处理尽管面临巨大挑战，同时更意味着前所未有的机会，是可与互联网诞生相媲美的新时代新机遇！

Thank you

Contact
Email: hhy63@bit.edu.cn

谢谢!