海外学人

ISSN: 1556-861X

大数据专刊

2013年

CHINESE ASSOCIATION FOR SCIENCE AND TECHNOLOGY U.S.A.
中國旅美科技協會
Chinese Association for Science and Technology, USA (CAST-USA)
www.cast-usa.net

中国旅美科技协会名誉顾问

杨振宁　清华大学教授、诺贝尔奖获得者
周光召　中国科协名誉主席、中国科学院前院长
朱丽兰　中国科学技术部前部长
邓文中　美国工程院院士、美国林同琰国际公司董事长
路甬祥　中国科学院前院长、全国人大常委会副委员长
宋　健　中国工程院前院长、全国政协前副主席
陈香梅　美国国际合作委员会主席

中国旅美科技协会机构与主要负责人

总会会长	董事会主席	理事会主席	预备会长
蔡逸强	盛晓明	方　彤	于　浩

总会副会长

陈志雄　协助会长工作
郭　光　学术活动及出版，负责新闻新讯及杂志
庞　涛　中美交流－学术，NACC 协调人
蔡路凯　中美交流－企业，协调人
沈　琦　中美交流－企业
胡　近　协会及会员发展，协调人
马　超　协会及会员发展，中国北方
汪　洋　协会及会员发展，美国东部兼会员信息库
陆　强　《海外学人》主编
潘星华　学术活动及出版，协调人兼财务
彭　伟　学术活动及出版
王志梁　中美交流－学术
谢长春　中美交流－企业
涂子沛　中美交流－企业
贾鹏程　协会及会员发展，中国南方
蒋为民　协会及会员发展，美国西部
张　宣　法律顾问

海外学人
Overseas Scholars

OVERSEAS SCHOLARS

2013 年大数据专刊
ISSN: 1556-861X

本期责任编辑：盛晓明 涂子沛 陆 强
发 行 人：曾大军

《海外学人》编辑部

杂志主办

中国旅美科技协会
Chinese Association for Science and Technology, USA (CAST-USA)
www.cast-usa.net

CHINESE ASSOCIATION FOR SCIENCE AND TECHNOLOGY U.S.A.

中国旅美科技协会简介

中国旅美科技协会（旅美科协）是 1992 年夏在纽约成立的非政治性，非盈利性的民间团体。旅美科协的宗旨是促进中美之间文化、科技、教育、经贸等领域的合作与发展；弘扬中国传统文化，促进中美两国人民的相互了解；促进旅美学人、华人专业人士之间的团结、合作与交流。

旅美科协是一个跨地区（美国）、跨行业的综合性科技团体，会员主要由来自科技、文教、工程、法律、金融、人文等各个领域的中国旅美专业人士组成，现有会员八千人。许多会员是在世界 500 强跨国企业或美国知名大公司、高等院校或研究机构从事科技开发、研究工作，部分会员已经成为了中、高层管理人员。目前在全美有十余个分会及学会，会员分布在几十个州，并在中国国内几个大城市设立了联络处。总会设执行委员会负责日常工作，还设有董事会、理事会、专业学术协调委员会和顾问委员会。旅美科协成立以来的知名名誉顾问包括陈省身教授，宋健教授，杨振宁教授，朱光亚教授，陈香梅女士，田长霖教授，周光召教授，朱丽兰教授等学术及社会知名人士。

旅美科协总会及各分会举行定期学术研讨活动，为会员提供学术交流的平台。旅美科协总会定期出版《海外学人》杂志及实时通讯，杂志及实时通讯内容包括介绍协会的学术活动与中美科技界、工商界的最新动态及各种工作与投资机会等许多会员们切身关心的内容。

每年总会及各分会举办包括全国年会及分会年会、学术讲座等在内的几十次大中型学术研讨活动，活动中旅美科协邀请中美各界知名人士对会员所关心的学术及社会问题进行了探讨。旅美科协注重与其他专业协会的交流与合作，加强不同学科华人的交流，同时促进中美之间科技人才的沟通和科技的发展。旅美科协各分会也注重参加当地的华人社区活动，与所在地的其它侨团建立了良好的关系。

旅美科协总会现任会长蔡逸强，董事会主席盛晓明，理事会主席方彤，候任会长于浩。旅美科协历任会长为周华康、章球、徐震春、陆重庆、马启元、周孟初、谢家叶、肖水根、石宏、邹有所、林民跃、王飞跃、李百炼、左力、沈陆、陆强、曾大军、方彤、盛晓明、蔡逸强。

Table of Contents

《大数据专刊》前言

涂子沛

由于数据的积累和爆炸，人类进入了大数据时代。这个时代的到来，已逐渐成为了社会各界的共识，从 2012 年起，“大数据”的现象和意义开始引起非常广泛的讨论。随着大数据时代的到来，我们也面临着巨大的机遇和挑战。Big Data: Big Challenges，Big Opportunities， Technologies and Tools 一文介绍了与大数据有关的一些“大”的概念，以及近来涌现的数据分析的工具、技术和方法。

作为一种现象，大数据的产生，主要源于信息技术的进步，因为这些进步，人类收集数据、使用数据的能力得到前所未有的增强。人类进入信息时代以来，数据总量一直在增加，在近十年来，开始真正爆炸，其直接原因，是因为社交媒体的出现。在 FaceBook，推特、微博、微信等社交媒体出现之前，人类收集数据的主要手段，是通过运营性信息系统和感应器，但社交媒体出现之后，每个社交媒体的使用者都开始大规模的贡献数据，这部分数据，在互联网在沉淀，我们称之为非结构化数据，从 2004 年 FaceBook 成立至今，不到 10 年，产生的非结构化数据已经占了人类数据总量的 75%左右。所以，社交媒体的出现，实在是让大数据时代一锤定音的标志性事件。

本期大数据分刊也正是从社交媒体出发，围绕大数据时代的几个重要问题展开讨论，可谓精彩纷呈。

《中文社交媒体的大数据舆情挖掘》一文重点反映了大数据为统计科学提供的崭新手段：以社交媒体为主的舆情挖掘。较之于统计科学的传统手段“抽样调查”，舆情挖掘不仅成本更低，而且更快、实时性更好，更加广泛到收集了关于社情民意的观点。文章描述了以自然语言处理技术（NLP）为核心的舆情挖掘系统的架构和实现，并例举了大量鲜活的案例，读完之后，令人印象深刻。Social Media Emoticons for Brand Sentiments 一文则在上文上更进一步，具体阐述社交媒体的品牌管理上的应用，这个系列的第三篇文章《社交媒体中的粤语情报挖掘》则关注了一个中文世界舆情挖掘的独特领域：粤语，系统地梳理了方言领域数据挖掘的种种困难，并提出了解决的意见，值得国内学者借鉴。

大数据时代，数据增值的一个重要前提是数据必须进行有效整合，这种整合，未来应该是“自动”的，而整合的关键又在于在全社会流通在各种数据必须有统一的“主键”，胡善庆博士的文章将其称为“识别码”，他以中美两国为例，重点探讨了合理设定人和商业组织“识别码”的重要意义和方法；我的文章则从中国社会的实际出发，在宏观上探讨中国政府应该采取什么样的具体措施来应对这场由于信息技术革命带来的社会挑战。

关于大数据的讨论，国内国外正方兴未艾。之所以将其称为一场“革命”，是因为大数据将给人类社会的方方面面带来深刻的改变和挑战。希望我们这一集专刊，能给读者带来启发、激发新的研究和讨论。

Big Data: Big Challenges, Big Opportunities, Technologies and Tools

Yugang Hu

中文摘要

近来关于大数据的讨论变得越来越热门。随着大数据时代的到来，我们也面临着巨大的机遇和挑战。本文旨在介绍与大数据有关的一些“大”的概念，以及近来涌现的数据分析的工具、技术和方法。

Big Data is becoming an increasingly popular buzzword these days. Huge amount of data were generated every second including from Facebook, twitter, Google searches etc. In the meantime, we have heard a lot about NOSQL, predictive analytics, real-time analytics, data visualization etc. So what is Big Data?

Definition

Big Data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. In 2012, Gartner updated its definition as follows: “Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.”

The three V’s are commonly used to characterize different aspects of big data:

1. Volume - amount of data
2. Velocity - speed of data in and out
3. Variety - range of data types and sources: Lots of them are semi-structured or even unstructured data, which traditional SQL databases can’t handle efficiently.

Additionally, a new V “Veracity” is added by some organizations as an indication of data integrity and the ability for an organization to trust the data and be able to confidently use it to make crucial decisions.

How Big is Big Data?

In the last couple years, we have been seeing enormous data generated. Most of them are from Internet, for example, tweets, Facebook updates, web click history, online transactions, interactions over mobile devices etc. Walmart controls more than 1 million customer transactions every hour, which are saved into database with over 2.5 petabytes of information.

According to IBM, users create 2.5 quintillion bytes of data every day, which means that 90 percent of the data in the world today has been created in the last two years alone. It is not just more streams of data, but entirely new sources. For example, there are now countless digital sensors/cameras worldwide in industrial equipment, automobiles, electrical meters and shipping crates. They can measure and communicate location, movement, vibration, temperature, humidity, even chemical changes in the air. Big is relative. Today’s Big Data might not be big any more in the future given that computer memories and hard drives are getting cheaper every year.

Challenges

The data we are trying to handle is too big to fit in main memory or a machine, which makes lots of the existing machine learning algorithms and analytics tools no longer useful. One way to get around is downsampling, however it comes with pitfalls sometimes, for example, accuracy suffered, events misclassified, user experience impaired etc. Before we adopted HDFS and MapReduce, MPP (massively parallel processing) architecture was mainly used by Data Warehouse appliances to provide high query performance and platform scalability, typically an expensive option for business intelligence. MPP architectures consist of independent processors or servers executing in parallel. Most MPP architectures implement a "shared-nothing architecture" where each server operates self-sufficiently and controls its own memory and disk. Data Warehouse appliances distribute data onto dedicated disk storage units connected to each server in the appliance. This distribution allows Data Warehouse appliances to resolve arelational query by scanning data on each server in parallel. Thanks to Hadoop and open source technologies, the specially configured hardware is no longer the only option. Now we have lots of system to run SQL-like queries, for example: Hive, Impala, Stinger, Shark, Lingual etc.

The foundation of big data is the technologies that support distributed storage and data processing.

* Storage

Big Data put lots of challenges to the traditional data storage players. Because data are too big to fit in a single machine, data need to be distributed across multiple machines, sometimes even thousands of machines. That's where HDFS (Hadoop Distributed File System), Google BigTable, NOSQL databases were designed for. Interesting enough, Apache Solr joined the NOSQL camp recently from search engine arena since Solr 4.0.

* Data Processing

Data processing prompts the needs to process data from distributed storage within reasonable timeframe. Traditional software can NOT handle the data volume any more, for example, R, Weka, traditional Business Intelligence Software which were doing great to handle small amount of data. Recently there have been some efforts to make them work in a cloud environment, for example, RHadoop is introduced to allow R to run in hadoop clusters.

* Data Mining and Machine Learning

For traditional SQL users, we have Hive, Cascading Lingual as well as Shark, Impala, Stinger, all of which claimed to be at least 50 times faster than Hive. Algorithms need to be revised to support distributed computing.

Big Opportunities

The big data market is booming. IDC, a research firm, predicts that the market for Big Data technology and services will reach $16.9 billion by 2015, up from $3.2 billion in 2010.

Career Opportunities

To exploit the data flood, America will need many more data scientists. A report last year by the McKinsey Global Institute, the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million

more data-literate managers, whether re-trained or hired. Data Scientist is the next sexiest career, which requires the following skills:

- Statistics
- Data Munging including parsing, scraping and formatting data
- Visualization

Business Opportunities

We are using data products all the time. For example: AutoComplete in Google Search, Friend Recommendation in LinkedIn/Facebook. People start following trending hashtags on Twitter to find out patterns of influence and peaks in communication on a subject. Because some data products are so impressive, business owners know what they can expect from data mining to grow their business and remain competitive. Thus more business requirements are coming up. Retailers are using sales data, customers' geo-location information etc. to forecast inventories and adjust prices. They are looking for couponing/retargeting opportunities based on customers' shopping patterns as well.

Technical Opportunities

New technologies keep coming up, especially open source technologies, are fuelling the Big Data Trend. They are making it easier for data scientists to understand data better. Lots of companies start seeing the values from data mining, for example:

- News Clustering - Clustering similar news into new groups
- Product/Movie/ Recommendations - Almost all of the online retailers have recommender system
- Spam/Fraud Detection

Improved access to information is also fueling the Big Data trend. For example, government data (data.gov), social media data, Wikipedia data etc. Data is not only becoming more available but also more understandable to computers. At the forefront are the rapidly advancing techniques of artificial intelligence like machine learning, pattern recognition and natural language processing. More and more machine learning theories are put into practice, for example, Recommender System, Customer Segmentation, Trend Prediction, Personalization.

There are huge opportunities in the following areas in Big Data as service provider:

- Advanced analytics
- Value-add services
- Professional services

Big Data Technologies And Tools

NOSQL (Not Only SQL)

A NoSQL database provides a mechanism forstorage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. Motivations for this approach include simplicity of design, horizontal scaling and finer control over availability. Even where there is not a radical data type mismatch, a disadvantage of the relational database is the static nature of its schemas. In an agile, exploratory environment, the results of computations will evolve with the detection and extraction of more signals. Semi-structured NoSQL databases meet this need for flexibility: they provide enough structure to organize data, but do not require the exact schema of the data before storing it.

Below are the different types of NOSQL Databases:

- Key-Value

Key–value stores allow the application to store its data in a schema-less way. Below are the main players in this area:

- Riak
- Redis
- Memcached DB
- Berkeley DB
- Amazon Dynamo
- Hamster DB
- FoundationDB
- LevelDB

- Key-Document

The central concept of a document store is the notion of a "document". Below are the main players in this area:

- MongoDB
- CouchDB
- RavenDB
- Elasticsearch
- Terrastore
- OrientDB

- Column Family Database

A column family is aNoSQL object that contains columns of related data.

- Cassandra
- HBase
- Hypertable
- Amazon SimpleDB
- Cloudata
- Accumulo
- HPCC

- Graph Database

This kind of database is designed for data whose relations are well represented as a graph (elements interconnected with an undetermined number of relations between them).

- Neo4J
- Infinite Graph
- OrientDB
- Gremlin
- FlockDB
- Titan

- Real time Processing

- Twitter Storm
- Spark/Shark
- Impala
- Stinger/Tez
- Facebook Presto
- Drill
- Solr/Lucene

Machine Learning Packages

Below are some of machine learning packages written in Java:

- Mahout
- MLBase (Part of Spark)
- Weka
- Mallet

In general, machine-learning algorithms learn from data. The more data, the more the machines learn.

Summary

Data is in the driver's seat. Big Data is a shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions. The predictive power of Big Data is being explored in various industries: Retail, Financial Service, public health etc. "It's a revolution," says Gary King, director of Harvard's Institute

for Quantitative Social Science. “We’re really just getting under way. But the march of quantification, made possible by enormous new sources of data, will sweep through academia, business and government. There is no area that is going to be untouched.”

According to Wikibon, the total Big Data market reached $11.59 billion in 2012, ahead of its 2011 forecast. The Big Data market is projected to reach $18.1 billion in 2013, an annual growth of 61%. This puts it on pace to exceed $47 billion by 2017. That translates to a 31% compound annual growth rate over the five year period 2012-2017.

Reference

[1] Bill Franks, “Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics”, 2012

[2] http://bits.blogs.nytimes.com /2012/03 /07 /idc-sizes-up-the-big-data-market/?_r=0

[3] http : // www . nytimes . com / 2012 / 02 / 12/sunday-review/big -datas-impact-in-the-world.html

[4] http://searchenginewatch.com/ article /2201030/Big-Data -An-Introduction-for-Search-Marketers

[5] http://wikibon.org/wiki/v/Big_Data_Market_Size_and _Vendor_Revenues

[6] http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue _a nd_Market_Forecast_2012-2017

作者简介

Yugang Hu is currently Principle Scientist at Overstock. Before that, he was a Lead Software Engineer at Overstock. He received his Bachelor’s degree from Zhejiang University, and Master’s degree from Peking University, both in Computer Science.

中文社交媒体的大数据舆情挖掘
Mining Public Opinions from Chinese Social Media

Wei Li, Tian Tang, Lei Li
NLP Department
Netbase Solutions, Inc.
Mountain View, USA
wli,ttang,lli@netbase.com

Abstract—Social media in Chinese has seen an explosive growth in the last few years. It becomes a more and more influential outlet for the public opinions, on any topics, including the sentiments of consumers towards a brand or business. Due to the large quantity of social media big data, manually collecting public opinions is impractical. Automatic mining from Chinese social media enabled by Natural Language Processing is a way of doing just that, in scale. This paper presents the general architecture and implementation of a Chinese parser-supported social media mining system, primarily designed and optimized to mine consumer insights on brands to serve businesses, but can also be used to mine any other topics of interest. Benchmarks show that the sentiments extracted by this system reach high precision (over 80%) and fair recall (comparable to that of English) at this point. For popular topics, the experiments already show very impressive and insightful results, thanks to the information redundancy of big data.

Keywords-NLP; Chinese parsing; social media; sentiment extraction; public opinion; text mining

摘要—以新浪微博、腾讯微信为代表的中文社交媒体在近几年呈爆炸式增长。社交媒体的影响力越来越大，它们成为普罗百姓表达意见的一个重要平台，是公众舆论的窗口。由于社交媒体的大数据性质，手工收集汇编舆情对企业或机构越来越不现实。舆情挖掘自动化是大数据时代的必由之路。本文介绍我们正在试运行中的基于自然语言处理技术（NLP）的中文社会媒体舆情挖掘系统，描述其架构和实现。第三方的测试显示目前的 Alpha 系统精准度（precision）已达 80%左右，查全率（recall）与比较成熟的英语系统相若。但由于中文比英语情感密度大，因此查全率应该还有较大研发改进余地。即便如此，由于大数据中自然存在的信息冗余，对于热点话题，试验结果令人印象深刻。

关键词：NLP； 中文自动分析；社会媒体；舆情挖掘；情感抽取；褒贬分析。

I . 引言

以新浪微博、腾讯微信为代表的中文社交媒体在近几年呈爆炸式增长。社交媒体的影响力越来越大，它们成为普罗百姓表达意见的一个重要平台，是公众舆论的窗口。社交媒体的大数据性质使得手工挖掘已经完全不堪负荷。舆情挖掘自动化是大数据时代的必由之路。

首先明确处理对象。系统挖掘的是社交媒体的文本大数据。所谓大数据（big data），实际上是社交媒体火热以后的专指，与数据施事（帖

者）的背景相关联，而不是搜索引擎从开放互联网搜罗来的混杂集合。没有社交媒体及其用户社会网络作为背景，纯粹从量上看，“大”数据早已存在，它催生了搜索产业。但那不是现代意义的大数据。如今的（文本）大数据与社交媒体其实是一回事。可以比照的是非文本的大数据，譬如交易记录，用户点击记录等。数据挖掘（data mining）领域把用户背景信息和消费或行为习惯的数据结合起来，已经有很多成果和应用 (Linoff and Berry 2011；Han 1999)。社交媒体上的文本大数据的处理可以看作是数据挖掘的自然延伸，叫做文本挖掘（text mining）。

本文介绍我们开发的基于自然语言处理（NLP）技术的中文社交媒体舆情挖掘系统，描述其架构和实现。第三方的测试显示目前的 Alpha 系统精准度（precision）已达 80%，查全率（recall）与比较成熟的英语系统相若。但由于中文比英语情感密度大，因此查全率应该还有较大研发改进余地。即便如此，由于大数据中自然存在的信息冗余，对于热点话题，试验结果令人满意。

社交媒体的特点概括来说，就是：不断翻新的海量信息源；满是不规范的字词和表达法。这就要求研发的系统，首先必须具有海量处理能力（scalable），还要有分析系统的鲁棒性（robust）。

社交媒体舆情挖掘的一个直接应用是自动民调（Automatic Survey），指的是针对特定话题，从数据中自动抽取挖掘相关的民间舆论。自动民调是对传统的问卷调查（questionnaire）一个补充或替代。在移动互联网日益普及的今天，民间情绪和舆论通过社交媒体管道铺天盖地而来，为了检测、采集和吸收这些舆论，自动民调势在必行。民意调查（poll）可以为政府、企业以及民众的决策提供量化情报，应用范围极其广泛。总统大选是一个突出的例子，对于总统候选人本人及其竞选团队，民调的结果可以帮助他们调整策略。产品发布是企业的例子，譬如 Windows 8 发布以后，民调的反馈可以帮助微软及时发现问题。社交媒体中这些自发地对产品的评价对于企业增强客户服务、改善产品功能和研发新产品，具有很高的情报价值。抽取挖掘这类信息的客户情报（customer insights）系统是自动民调在企业的应用。对于消费者，自动民调的结果有助于他们在购买、等待还是转向别家的决策时，不致陷入盲目。

本文第二节综述相关工作。第三节讨论深度情感抽取的概念及其定义。第四节描述基于语法分析的汉语情感抽取方法。第五节讨论挖掘应用层面的情报整合和表达。第六节展示并解说舆情挖掘的实际案例，表明本系统并非实验室的研究原型，而是投入大数据实际使用的应用。第七节对相关实验结果进行讨论，最后是结论和展望。

II. 相关研究与讨论

舆情挖掘的基础是情感抽取（sentiment extraction），也叫情感分析（sentiment analysis）。情感抽取是近年大数据的概念出来之后自然语言处理（NLP）学界和业界的一个热点。从语言表达方式来看，自然语言可以大体分为客观语言（objective language）和主观语言（subjective language）两大类 (Bruder and Janyce 1990；Yu and Vasileios 2003)。前者陈述事实，后者表达情感（喜怒哀乐等）或做出褒贬评价。传统的信息抽取的对象是客观语言，目标是对事实（实体关系和事件）的抽取（Chinchor and Marsh 1998）。主观语言的研究和情感抽取起步较晚，但是特别受到大数据时代工业应用界的重视，因为它为企业渴望了解的客户情报（无论是对产品品牌的抱怨还是夸赞）提供了自动解决方案。

情感抽取根据抽取目标可以分为浅层抽取和深度抽取两种。多数情感抽取系统做的是浅层抽取工作，核心是情感的褒贬分类（sentiment classification）。相关的任务包括：（1）主观语言和客观语言的区分；（2）主观语言的情感褒贬分类：情感包括爱恨、喜欢不喜欢，褒贬包括优劣评价：二者表达角度不同，但语义倾向是一致的，可以归为一类（说一个品牌“很好”隐含着“喜欢”这个品牌）；（3）褒贬的语气强度。

开始的情感分类系统多以情感词典为主（Wiebe 2000；Turney 2002；Riloff 2003）。有情感分类结合词典和上下文的计算（Ding, Liu and Yu 2008；Taboada et al 2011； Wilson, Wiebe and Hoffmann 2009）。情感抽取的主流是利用机器学习基于所谓“一袋子词”（bag of word）模型的的情感分类（sentiment classification），通常的做法非常粗线条，就是把要处理的语言单位（通常是文章 document，或帖子 post）分类为正面（positive）和负面（negative），叫做 thumbs up and down classification (Pang， Lee and Vaithyanathan 2002；Turney 2002)。这种做法流行快捷，在某个特定领域（譬如影评、产品意见网站），分类质量可以很高 (Titov and McDonald 2008)。这是因为在一个狭窄的领域里面，评论用语相当固定有限，正面负面的评价用词及其分布密度不同，界限清晰，通过一袋子关键词的不同密度来识别褒贬自然不难。而且现在很多领域都不愁标注好的数据，因为越来越多的用户评价系统在网络上运转，如 Amazon、Yelp，积累了大量的已经分类好的数据，给情感分类的广泛应用提供了条件。

但是，上述分类遇到了下述一系列挑战。首先，领域移植性不好，影评数据训练出来的分类器（classifier）换到电子器件的客户评价分类上就不管用。要对多个领域训练出多个分类器，很耗时，效果也不能得到保证。如果面对的是整个社交媒体，就必须具有跨领域分类能力。

第二个挑战是，语言单位的缩小使得分类所需要的词汇证据减少，分类难为无米之炊，精度自然大受影响。多数系统针对文章（包括新闻）或者较长的帖子（如产品评介），机器学习比较擅长。句子或短语级的情感分类的尝试刚刚起步（Khan, Baharudin and Khan 2010； Yu and Hatzivassiloglou 2003； McDonald et al. 2007; Titov and McDonald 2008； Wilson, Wiebe and Hoffmann 2009）。多数这类研究带有很强的实验室性质，或者过分依赖领域，因此离开大规模实用还有距离。从文件到帖子到段落再到短句，语言单位每一步变小，舆情分类就日益艰难。在语句或者小于语句的单位上做分类工作，机器学习很难施展。语言单位缩小的挑战比领域挑战更加严重。可是社交媒体的趋势是短平快的微博极为流行，所占比例越来越大，因此，善于精雕此刻的基于规则的褒贬分类及抽取在保证微博情报质量上占有优势

第三是大数据切割（slicing and dicing）的挑战。本来我们利用机器来应对大数据时代的信息挑战，起因就是信息时代的数据量之大，无法依靠人工去捕捉、监测情报的变化。但是，观察具体应用和情报需求的现场就会发现，用户不会仅仅满足于一个静态的、概览似的情报结果，他们所需要的是这样一个工具，它可以随时对抽取情报进行各种各样的动态切割。最典型的切割应用是以时间为维度的《动态晴雨表》，可以反映一个研究对象的舆情走势（trends）。譬如把一年的总数据，根据每月、每周、每日、每小时，甚至每分钟予以切割，然后观察其分布走势。这对于监测和追踪新话题的舆情消长（譬如实时追踪总统辩论的舆情反应），对于新产品的发布，新广告的效用评估（譬如美式足球赛上的巨额品牌广告的客户效应）等，都有着至关重要的情报作

用。总之，大数据很可能在具体应用时要被切割成小数据，一个分类精度不高的系统就会捉襟见肘，被大数据信息冗余遮盖的缺陷凸显。

第四个挑战是找舆情对象的问题。在几乎所有的舆情分析应用中，舆情与舆情的对象必须联系起来，而这一基本要求常常成为舆情分类系统的软肋（Davidov，Tsur and Rappoport 2010）。多数分类系统利用品牌词的共现作为确定舆情对象的依据，这条经验（ heuristic）在遇到多于一个品牌共现时遇到困难。比如，“谷歌比雅虎强老鼻子啦”这样的语句，正面评价“强”到底是指雅虎还是谷歌，这看似简单的问题，对于基于一袋词的分类系统就是难题。这样的问题在基于句法结构分析的系统中较易解决。

上面列举的机器舆情分类系统的挑战，并不是要否定机器学习在情感抽取领域的价值，而是要阐明以下的观点：粗线条的机器分类只是舆情自动分析的开始，万里长征的第一步。一个真正有价值的舆情挖掘系统还需要更多更细致的深度舆情自动抽取的技术来支持。这一点第三节第四节将予详细分析。

对汉语句子做浅层情感抽取的工作近年开始引起研究者的关注 (Tsou et al 2005、朱嫣岚等 2006、李钝等 2008、王素格等 2009、党蕾等 2010、赵妍妍等 2010）。特别值得一提的是 Yang and Hou (2012) 基于规则的工作。他们从人工词典标注出发，利用规则系统，对汉语语句进行自动情感分类，步骤清晰，工作做得比较踏实。该系统设计的输出是一个表示趋向的数值。这个人为的设计带来了操作上的困难，包括人工标注词典的困难和人工标注语句的困难。这两个困难的叠加使得测量系统的准确度以及指导系统的进一步发展，变得不太可靠。系统的测试量很小，与大多数中文情感系统一样，它依然是一个离开实际语料和应用非常遥远的实验室系统。

这些研究的局限在于：（1）情感分析很粗疏，以分类为主；（2）分析主要依赖词典和非常有限的上下文（immediate local context），没有句法分析的支持；（3）离开实用距离远。本文的研究和应用在这三方面均有突破。

对大数据的舆情挖掘是建立在对具体语言单位的舆情抽取的基础之上。只有当语言海洋中千千万万的舆情表达被抽取存贮到某个数据库以后，我们才有条件针对具体的舆情问题（如某特定品牌的网络形象或某话题的舆情走势），搜索有代表性的舆情资料，并将搜索结果整合提炼，然后以用户可以调控的各种方式（譬如《品牌舆情图》或《话题晴雨表》）表达给情报使用者。文献很少有中文文本挖掘和舆情表达方面的研究工作，我们的工作也是刚刚开始，但已经显示出大数据及其挖掘的力量（详见第五节第六节）。

III. 深度情感抽取的概念

在浅层抽取之上，进一步走向细颗粒度的深度抽取是本系统的一个特色。本节专门讲解深度抽取的概念、定义及其设计思想。

3.1 情感分类的颗粒度

前面说过，情感分类的二分法，或者加上中性的三分法，对于情感抽取是太粗疏了。也有系统把正反情绪类别细分为喜怒哀乐等多种（Strapparava and Mihalcea 2008）。究竟分成几类合适应该由应用面的信息需求，兼顾可行性而定。重要的是在情感（sentiments）分析中要区分从施事（agent，譬如消费者）出发的主观性情绪（emotions）或评价和从对象（object，如品牌、产品）出发的客观性优劣得失（所谓【特点】）。它们之间的概念关系如下：

【情感】

【施事主观性】

【情绪：（爱/恨/喜欢/不喜欢/无所谓）】

【评价：（好/坏/还行/不咋样/中等）】

【对象客观性】

【特点：(优劣、得失、长短…)】

【对象：(品牌、产品、话题…)】

【方面：（材料、服务、价格…）】

【施事】

【性别：（男/女）】

【年龄组：（青少年、中年、老年…）】

【职业：（学生、工程师、退休者…）】

【民族：（亚裔、拉美裔、墨裔…）】

上述情感有关的层级信息（hierarchy）对于开发一个客户情报系统特别有意义。在抽取了上述信息的基础上做文本挖掘可以揭示对于企业很有价值的情报。值得一提的是，【方面】+【评价】常常可以组合成【特点】，譬如：外观+漂亮、价格+合理、服务+周到、材料+坚实。

褒贬评价在语言事实上是一个从褒到贬的连续体，用到很多强烈程度不同的语词，如：盖帽了、很棒、不错、较好、还行、一般、差强人意、不行、糟透了。对于客户情报的应用目标，我们选择首先区分正面、反面和中性，然后识别情绪或评价的强弱，相当于对情感切五刀的分类法：强褒（好/爱）、褒（还行/喜欢）、中性（一般/无所谓）、贬（不咋样/不喜欢）、强贬（坏/恨）。这是我们褒贬评价分析的预定标注。

为什么如此设计？情绪分类过细较难实现，也没这个必要。从个体而言，语言表达中“喜爱”比“喜欢”强烈，“迷恋”又比“喜爱”强，“痴狂”似乎比“迷恋”强，“非常痴狂”比“痴狂”更强。这样下去，别说三刀、五刀，就是 10 刀也不能区别语言表达的种种细微差别及其语义组合的可能性，因为人的情绪是连续体，其表达也相应千变万化。但是，设计的时候还是要用离散来对连续模型化，而且要限制离散度，以便宏观把握：语义上太细琐了标准就难掌握，难以操作、实现和检测。最重要的是，劳而无功。作为连续的模型化，切五刀已经很不错了，这也正好与多数客户调查使用五颗星的制度相吻合（学生总成绩的评判也多采用五分制，偶尔采用两分制 pass/not pass，单个测验为方便计算采用百分制，但随后还是要整合到五分制去）。分类过细的思路也不是大数据的思路。采样大可以弥补个体颗粒度的粗疏，这在机器学习领域被一再证明（也是很多统计学家不屑于语言学家雕虫小技的缘由之一），也在我们的客户情报系统的最终结果中一再得到印证。

可是并非每一类情感相关的情报都可以用分类的思路来解决，譬如下面要讨论的褒贬的理由，不同客户不同产品就千差万别，无法分类到预先定义的有限类别去，这给机器分类提出挑战，却给规则系统的精雕细刻创造了发挥力量的地方。机器学习领域的对此的应对是自底而上的归并（clustering），属于非监督学习（unsupervised learning）的路子，研究性强，应用上不易把握。

3.2 褒贬的理由

情绪自然要抓，更重要的是要抓情绪背后的理由。只知道数大拇指还是中指的数量，知其然不知其所以然是单纯的情绪分类系统的缺陷。情绪虽然对舆情的总体概览有意义，但远远不是直接帮助决策的情报（not actionable insights）。知道很多人喜欢或者不喜欢一个品牌，企业还是不知道怎么办，最多是在广告宣传投资量的决策上有些参考价值，但对于改进品牌产品，适应用户需求，褒贬情绪的二分法（或五分法）太过抽

象，不能提供有价值的指导。这就要求舆情分析在褒贬分类以外，进一步发掘这些情绪的背后动因，回答**为什么**网民喜欢或不喜欢一个品牌。譬如挖掘发现，原来喜欢麦当劳的主要原因是它发放优惠券，而不喜欢它的原因主要是嫌它热量太大，不利减肥。这样的舆情才是企业在了解自己品牌基本形象以后，最渴望得到的高价值情报，可以据此调整产品方向（如增加绿色品种和花样，水果、色拉等），改变市场化的策略（如强调其绿色的部分）。

值得注意的是，人的情绪和评价属于主观语言的范畴，而情绪评价背后的理由则需要用客观语言描述以增强说服能力。这也符合客户做评语或一个人论证自己立场（make an argument）时的套路，所谓以理服人：先表达观点（好恶如“我太喜欢海底捞了”，或评价“海底捞太棒了”），然后给出支持观点的种种客观理由，譬如“口味鲜美”、“价格合理”、“进口食材”等等。可见，包含原因和论据的情感抽取已经超出了传统情感抽取局限于主观语言的框框。因此，以前那种先做主观语言客观语言分类，再行主观语言内部的褒贬分类的做法格局太窄。

褒贬情绪背后的动因往往是品牌的具体【特点】（优点缺点，pros/cons）。这些长短优劣的表述千差万别，不同领域的产品很不同，反映了品牌的各个【方面】（譬如，智能手机的方方面面包括尺寸大小、屏幕精度、电池时间等），抽取挖掘难度比简单的褒贬分类大很多。从问答（question answering）系统的角度看，why 的问题和 how 的问题是公认的最难自动回答的问题（Srihari and Li 2006）。舆情分析着重于挖掘一个对象的优劣长短从根本上为企业回答客户舆情背后的 why 问题提供了解决方案。

3.3 定义情感抽取框架

情感抽取与传统的信息抽取类似，抽取什么情报都是事先定义好的，一般用一种语义框架（semantic frame，亦称模板，template）作为表达，它是用户信息需求的反映，作为抽取系统的目标。根据本节舆情深度分析的讨论，情感抽取的语义框架的完整定义如下（其中【理由】和【特点】都可以是情绪评价的依据，前者常常是显性的原因，而后者往往是隐性的原因）：

【分类：正面情感/负面情感】
【对象：（品牌/话题）】
【情绪：（喜欢/不喜欢/高兴/悲伤…）】
【行为：（采纳/舍弃/购买/杯葛 …）】
【强度：（强/弱）】
【评价：（好/坏…）】
【特点: (优劣、得失、长短…)】
【方面：（对象的不同侧面或部分…）】
【施事：（帖者 …）】
【理由：（因为…）】

IV. 深度情感抽取的方法

4.1 总体设计思想

情感抽取引擎的总体设计是以自浅而深、层层推进的句法分析来支持社交媒体的深度舆情抽取。整个引擎形成一个自底而上、由线性字符串逐步解构成平面树结构的多层管式系统（multi-level pipeline system）。多层管式系统的设计有如下好处：分而治之，把复杂的语言分析问题简化为一个个子任务，有利于模块化开发。

引擎由两大构建组成。底层是一个中文语法分析器（Chinese parser），顶层是情感抽取器（sentiment extractor）。语法分析器旨在把中文线性字符串（文句）从语言学角度层层处理，条分缕析，生成相应的结构句法树。这样的结构树为下一步利用子树匹配（sub-tree matching）的方

法抽取褒贬信息打下了语义逻辑基础。自动分析通过编制一连串有限状态文法（finite state grammars，Roche and Schabes 1997）的管式规则系统实现，类似的形式系统的架构和 NLP 开发平台见（Silberztein 1999；Hobbs 1993；Srihari et al. 2006b）。

句法分析支持舆情抽取的总体设计的根本好处在于：（i）句法分析原则上独立于领域，它是纯粹语言学模块，同一个句法分析器因此可以支持不同领域、不同用户、不同产品的各种信息需求；（ii）由句法分析支持的信息抽取，无论为了舆情还是关系事件，都可以做到高效、灵活和深入。高效灵活表现在信息抽取系统的开发时间大为缩短，因为繁难的语言现象已经在句法分析阶段涵盖并结构化了。此外，因为有了句法结构，深度抽取成为可能，譬如比较结构（comparative structure）在一袋子词抽取中是无解的，但在句法结构帮助下就容易实现。

信息抽取（包括褒贬抽取）当然也可以绕开繁难的句法分析，建立在浅层的语言处理（譬如切词）之上。事实上，多数情感抽取系统就是这么做的，然而没有句法支持抽取领域移植性差，也难以做到深度抽取。分析器和抽取器的简单道理就是：分析越深，抽取就越简单；分析越浅，则抽取就不得不繁复。总之，工作量是基本恒定的。很多系统信息抽取做得很辛苦，那是因为分析深入不下去。

4.2 不规范的社交媒体及其处理对策

什么叫社交媒体的语言，它与规范语言（譬如新华社新闻）有什么不同？下面是从微博摘取的客户评价的社交语言真实样本，可见系统面对的处理对象与规范语言的词汇风格有多大的不同：

地球人已经阻挡不了中国人了,连日剧里都用联想电脑, 那是多么奢侈的事情啊,都不用买东芝了.

联想电脑真是垃圾,但更垃圾的是要 5 年才能更换,这 是电子设备啊,怎么跟铁锹榔头一样折旧??????

联想电脑现在越来越烂了, 还变得和小日本一样狡猾, 一过保质期就坏!

东方航空的飞行员太给力了,这一路各种转弯啊,害我替旁边的阿拉伯大哥担心一路,怕他胃里的羊肉和馕溢出来啊.

@东方航空 95530 也忒难打了,天亮打到天黑,累计 n 小时,愣是"对不起, 坐席正忙".

我在萧山机场被告知飞机要晚点半小时,不喜欢东方航空

坑爹的东方航空啊, 上次是取消这次又给我晚点,再上次又改签... 乃跟我有仇吗?

为啥这东方航空公司网站经常性地打不开, 老是搞得黑客热爱他们的航班似的...

严重表扬@东方航空. 又一次被早餐惊喜到,竟然是荷叶饼夹肉沫炒鸡蛋+黑米粥

图 4.1：社交媒体数据样本

图 4.2：社交媒体用语极不规范

鲁棒性（robustness）是管式规则系统预防错误放大（error propagation）的关键，亦是应对不

规范社交媒体的必需。为了取得语言处理的鲁棒性，一个行之有效的方法是在系统研发中实现四个形容词的所指：词典主义（lexicalist）；自底而上（bottom-up）；调适性（adaptive）；和数据制导（data-driven）。这四条是相互关联的，但各自重点和视角不同。系统设计和开发上贯彻这四项基本原则， 是取得鲁棒性的良好保证。有了鲁棒性，系统对于不同领域的语言，甚至对极不规范的社会媒体中的语言现象，都可以应对。限于篇幅，下面简单谈谈词典主义和调适性开发这两点。

实用 NLP 系统要使用海量词典早已是业界的共识了。词典工作做得越足，系统质量就越高。电脑擅长的是记忆，多少词典都吃不饱它。本系统的词典资源如下：

(1) segmentation-only lexicon: 184k
(2) feature lexicon: 59k （including HowNet）
(3) location names: 3k
(4) person names: 64k
(5) product/brand names:21K
(6) company names: 1.7k
(7) other names: 2k
(8) idioms: 52k
(9) Cantonese-only lexicon: 4.9k
(10) Social media jargon: 1k

Total:

393k participated in segmentation, and
208k with lexical features to support parsing and sentiments

带有语法语义参数的语言学词条已达 20 多万（图 4.2），包括《知网（HowNet）》六万词条的前 200 多语义类别（董强、郝长伶、董振东 2003），也包含五万多条汉语成语和熟语。作为参照，【现代汉语词典】（第六版）收录六万九千词条；【辞海】（1999 版）共收录十万左右词条。除了专业词典外，现代汉语通用词汇基本一网打尽了。

坏水|n|minusEntity,personRelated
坏死|v|minusAction,disease
坏心眼儿|n|minusEntity,personRelated
坏血病|n|disease
坏账|n|minusEntity,financeD
坏种|n|minusEntity,human
欢畅|a|joyful
欢唱|n|joyful,sing
欢度|v|joyful,pass
欢歌笑语|a|joyful,sound
欢呼|v|joyful,cry
欢聚|v|joyful,assemble

图 4.3：语言学机器词典样本

词条总量达到 39 万三千。系统还有一部社交媒体词典（图 4.3），粤语词典，以及经过统计得来的用于切词的 18 万多高频词语。

强：　弓虽
非常:　灰常
谢谢:　3Q 3X THX
没有:　米有 莫有 莫油 木油 木有
我操:　卧槽 卧草 wk 我擦
傻逼:　煞笔 傻B 傻× 傻叉 狗逼 狗B 傻屄 狗屄 二逼 2逼 2b 2货
他妈的: 他喵 特么 尼玛 尼码 尼码比 草泥马 你妈逼 娘西逼
乐极生悲:乐极则忧 乐极则悲 乐极悲来 乐极悲生 乐极生哀
乱乱哄哄:乱乱纷纷 乱乱腾腾 乱乱轰轰
争奇斗艳:争妍斗丽 争妍斗艳 争奇斗妍 争芳斗妍 争芳斗艳

图 4.4：错别字及网络用语规范化词典样本

调适性作为管式系统的开发原则是克服错误放大的反制，保障鲁棒性的必需。理想世界里，系统模块之间的接口是单纯明确的，铁路警察，各管一段，步步推进，天衣无缝。但是实际的系统，特别是自然语言系统，情况很不一样，正误夹杂，后面的模块必须设计到有足够的容错能力，针对可能的偏差做调适才不至于一错再错，小错铸成大错。大多数的中间错误类型都是可预测的，至少是可观察的，因此后面的模块可以矫枉过正，负负得正。很多人夸大了管式系统的错误放大问题，他们忽略了调适性系统的容错能力。

4.3 中文句法分析

文句是隐含着结构的线性字符串，自动句法分析的任务是把它层层解构，最终转换成平面的结构图。中文分析器对社交媒体的文句进行自底而上的层层模式匹配（pattern matching），由浅层分析（包括切词、词类消歧、词组抱团等短语结构的处理）到深层分析（包括主谓宾、定状补等的句法关系分析），最终生成依从关系的句法结构树（syntactic dependency trees）。图 4.5 是中文分析从浅层到深层的架构图，限于篇幅，分析内部的各模块功能及其实现留待另文专述。

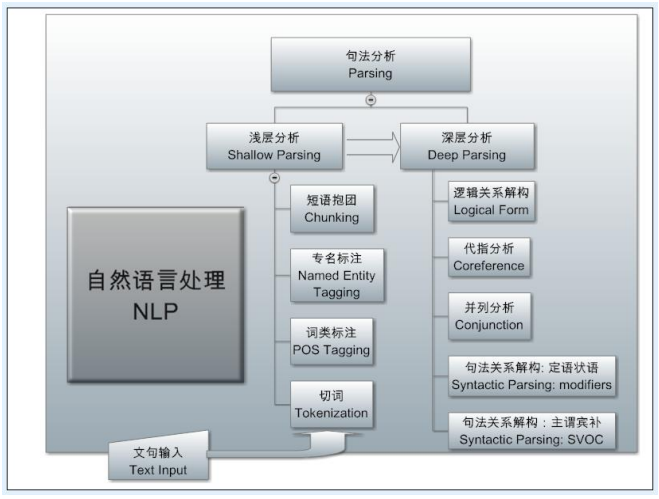


图 4.5：中文分析器架构图

我们使用的句法结构图以依从关系为主体，但包含了浅层分析结果的短语结构（phrase structure，如下图中的名词短语结构“ios5 的升级”，“我的想象”和“提升的功能”），是一种混合式依从关系树形图（hybrid dependency tree，Li 1989）。依从关系直接表达句法单位之间以及短语结构之内的语义关系，同时又兼得短语结构对于树节点的抽象性，因此方便了下一步抽取句型的实现。图 4.1 样本第一句经过自动句法分析输出的句法树见图 4.6。

为什么要结构化？盖因语言表达（乔姆斯基所谓表层结构）是无限的，但句法结构是有限的。只有结构化了，抽取规则才有表达力和概括

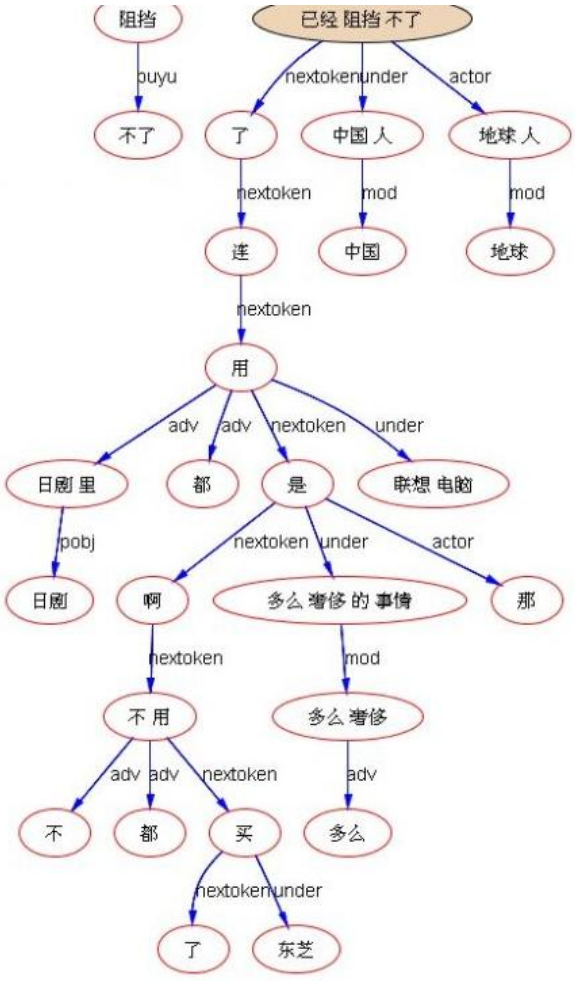


图 4.6：句法树样本

性。考虑语言单位的组合可能性，一条抽取规则所代表的句型常常等价于几百甚至几千条语言表层模式（surface patterns）的规则。依从关系树形结构的形式为下一步高效情感抽取打下了坚实的基础。

4.4 情感抽取

图 4.7 展示句法分析基础上的信息抽取架构图。舆情抽取从架构上与传统的信息抽取（窄义）同属信息抽取（广义），二者所用的资源和方法并无二致。

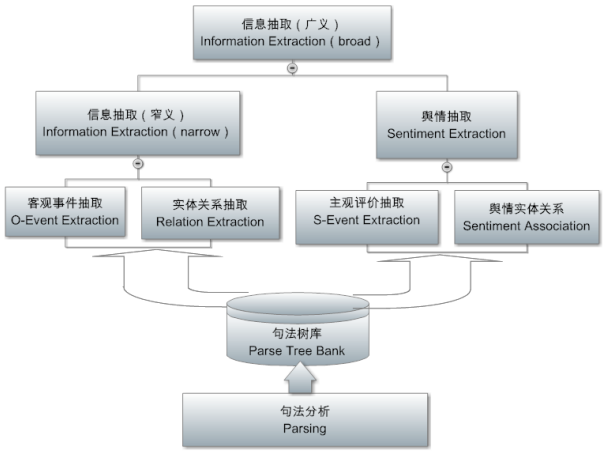


图 4.7：信息抽取架构图

舆情抽取通过对于句法结构树的子树匹配来实现。图 4.8 显示抽取子树的模式规则：其中主语（actor）是可有可无的节点，以问号表示；父节点“用/使用/运用/采用”与其子节点宾语（undergoer）是必要条件。

【分类：正面情感 subtree rule1】
【父节点：WORD =“用/使用/运用/采用”】→【行为】
【子节点 actor: FEATURE = human】?→【施事】
【子节点 undergoer: FEATURE = brand/product】→【对象】

图 4.8：子树匹配规则样本

一旦与图 4.6 句法分析树匹配成功，情感抽取框架的相应角色【行为】、【施事】和【对象】得以赋值，抽取工作完成（见图 4.9）。

【分类：正面情感】
【对象：*联想电脑*】
【行为：*用*】

图 4.9：情感抽取结果

图 4.10 与图 4.11 展示分析与抽取的全过程。

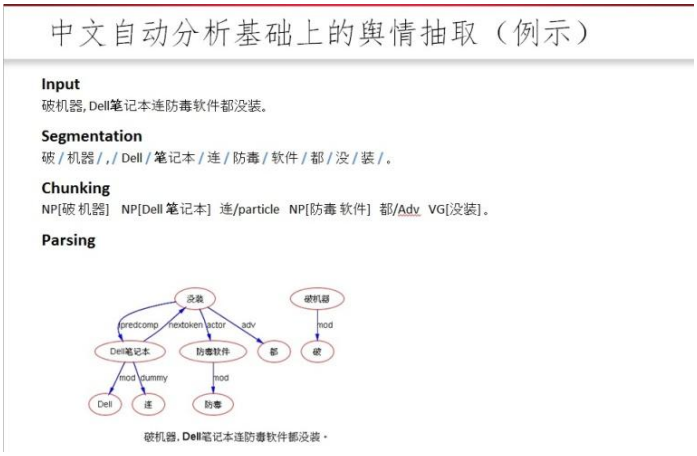


图 4.10：句法分析步骤例示（walk-through 1）

图 4.11：舆情抽取步骤例示（walk-through 2）

V. 舆情挖掘和表达

整个舆情挖掘系统由前后两个子系统组成。第四节描述的核心引擎是后台子系统（back-end indexing engine），用于对社交媒体大数据做自动分析和抽取。分析和抽取结果用开源的 Apache Lucene 文本搜索引擎存储（lucene.apache.org）。生成后台索引的过程基于 Map-Reduce 框架，利用计算云 (computing cloud) 中 200 台服务器进行分布式索引。对于过往一年的社会媒体大数据存档（约 300 亿文档跨越 40 多种语言），后台索引系统可以在 7 天左右完成全部索引。

5.1 前台应用的后备式模型

前台子系统（front-end app）是基于 SaaS（Software as a Service）的一种类似搜索的应用。用户通过浏览器登录应用服务器，输入一个感兴趣的话题，应用服务器对后台索引进行分布式搜索，搜索的结果在应用服务器经过整合，以用户可以预设（configurable）的方式呈现给用户。这一过程立等可取，响应时间不过三四秒。前台系统负责搜索、挖掘、整合和表达，设计成一个三层的混合后备式模型（hybrid back-off model），以求最大程度地满足不同用户对于情报精度（precision）与召回率（recall）之间的平衡的不同需求。第一层是句法分析支持的高精度情感抽取，第二层是统计支持的中等精度的情感分类，第三层是基于关键词的相关情报搜索（图 5.1）。

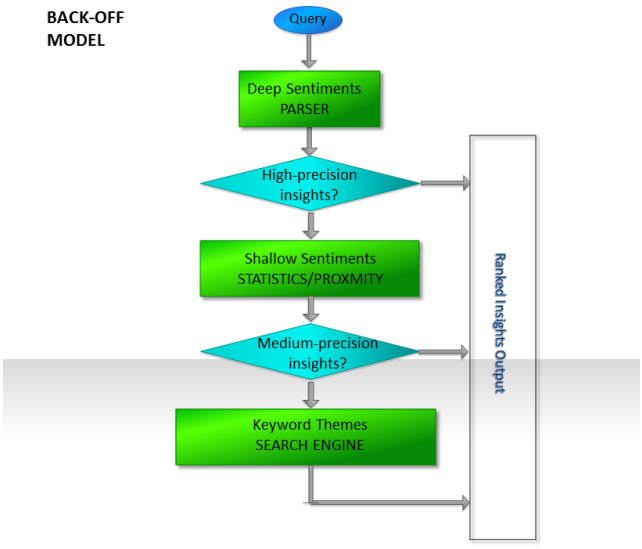


图 5.1：前台应用系统的混合后备式模型

第一层是高精度情报，情报来自数据库中深度舆情抽取的结果。其所以叫做高精度情报，乃是因为所搜索的品牌或话题与舆情抽取的模式完全匹配。数据质量因此得到保障，情报精度(80%以上)在中文舆情实用系统中得到验证。

第二层中等精度的情报是第一层的后备。它的基础是机器学习加上一些 heuristics（譬如 emoticons）支持的粗线条情绪分类，利用所搜索的品牌或话题与分类的情绪在帖子中的共现关系作为关联依据。这一层与多数情感抽取系统大同小异。

第三层低精度情报用的是一个内部建立的搜索引擎，利用关键词与搜索的品牌/话题之间的共现关系返回一切相关的主题词（theme words）和其可能隐含的情报，理论上的召回率为 100%。

如前所述，统计方法和机器学习擅长文件分类，从宏观上粗线条把握语言现象，而计算文法则擅长细致深入的语言学分析，从细节上捕捉语言现象。如果把语言看成森林，语句看成林中形态各异的树木，总体而言，机器学习是见林不见木，计算文法则见木不见林。从效果上看，机器学习常常以广度胜出（high recall），而计算文法则长于分析的精度（high precision）。由于自然语言任务比较复杂，一个实用系统（real-life system）常常需要在在粗线条和细线条以及精度与广度之间取得某种平衡，因此结合两种方法的 NLP 混合式系统（hybrid system）往往更加实惠好用。一个简单有效的结合方式是把系统建立成

如上描述的后备式模型，先让计算文法做高精度的处理，再行机器学习出来的统计模型（statistical model），以便粗线条覆盖遗留问题。关键词索引则是保证召回率的最后一道保障。

5.2 舆情表达和图示的设计

对于应用，搜索和挖掘舆情是问题的一面，表达（presentation）和图示（illustration）是问题的另一面。在信息爆炸时代，每个搜索（query）都可能有几十万甚至几百万个舆情数据点，如果把结果铺天盖地一股脑输出给用户，用户很容易被淹没。另一方面，做品牌搜索研究的分析员不仅仅需要大数据的精准情报，他们更需要把代表性情报制成各种容易理解的图标，报告给管理层在决策时做参考。因此，表达和图示的设计也是舆情应用系统不可或缺的重要环节。

先说一下《多品牌舆情对比图》的设计。《多品牌舆情对比图》是大数据自动民调的非常有价值的长项，可以把一个行业的一系列品牌的客户形象和竞争态势一目了然地展示给企业用户，而手工客户调查由于成本考量几乎不可能做多品牌的调查对比。有两类不同层次的情报，第一类是在挖掘（mining）层，表达的是社会媒体对品牌的总体评价（褒的数量与贬的数量的比例）；第二类是在抽取（extraction）层，表达的是褒贬语言是如何表达的，言辞激烈与否。这两类信息虽然可以整合到一起去，比如把第二类信息加权或者打折计入数量底数然后再做比例，但整合以后不利于情报用户对这两类情报分别处理和理解。因此，我们把褒贬指数与情绪强度作为两维标示在《多品牌舆情对比图》上，还有第三维热议度，系统根据挖掘层检索出来的数据由相应品牌的圆圈大小来展示（图 5.2）。这样设计的舆情对比图一目了然，适合于反映所调查领域的多品牌的社媒形象，受到企业用户的欢迎。

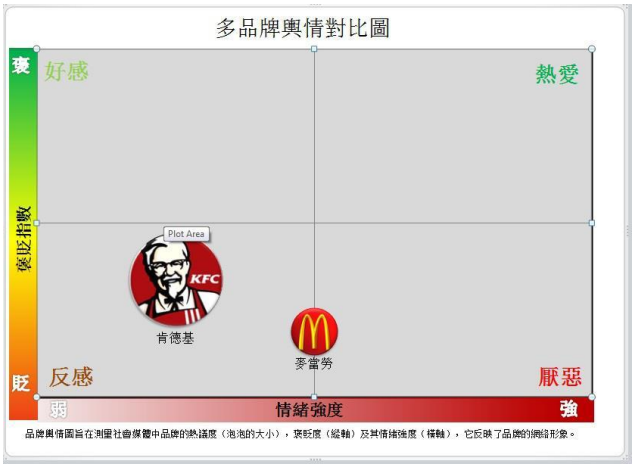


图 5.2：多品牌舆情图的设计

一般来说，单个品牌的数据有可能不好解读。但是如果是做同类品牌的比较，相对数据及其质量（包括褒贬挖掘的精准度与覆盖面）不会有问题，因为机器不懂歧视。至于褒贬基准（中间的十字线），工具里面提供了按比例微调（configure）的可能性，为的是可以突出品牌之

图 5.3：馬英九施政一年来《輿情晴雨表》

间的区别（如果太拥挤或者跌出界限的话）。

另外一个重要的舆情图示工具是时间轴上的褒贬趋势（trends），叫《舆情晴雨表》（图 5.3，关于此舆情表的点评见 6.5 节）。在这里褒贬指数依然存在，但是情绪强度被时间的维度取代了，这样出来的曲线可以看到一个对象的形象

图 5.4：KFC 品牌的主题词云

图 5.5：针对馬英九的情绪云图

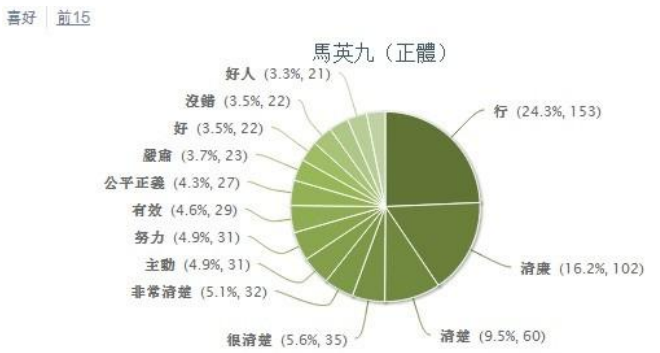


图 5.6：支持馬英九的主要理由

消长，对于监测和预示品牌的走势很有意义，在产品发布和广告效用的分析中常用。（其缺点是把褒贬舆论分解到时间长河中，数据量有时候不足，结果显得不太牢靠。而《舆情对比图》往往以一年的数据量作为基础，常常有几十万、几百万个数据点（data points），反映出来的形象真实可靠，即便有杂音，也被大数据自然过滤掉了。）

其他的舆情表达图示手段还有近几年流行起来的词云（Word Cloud）和传统的图表（如 Pie Chart，Table 等）。词云对于揭示与话题相关的主题和情绪表达非常有效直观，对人的视觉有冲击力，往往让人印象深刻，例示如下。

图 5.4 是肯德鸡（KFC）的主题词云，除了快餐业最大的竞争对手“麦当劳”外，主题中比较突出的还有“鸡腿王”和“禽流感”，前者似乎是肯德鸡在消费者心中的定位，后者是肯德鸡

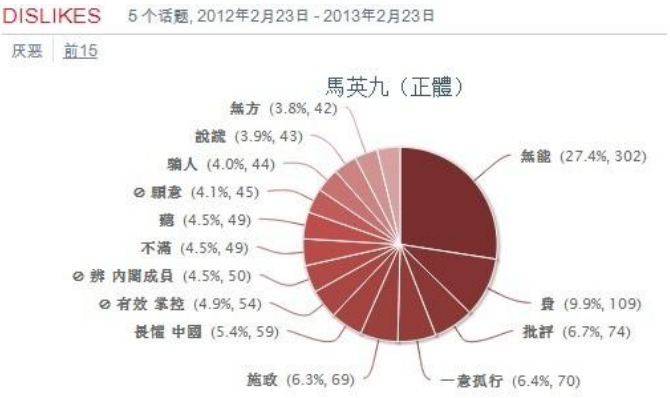


图 5.7：反对馬英九的主要理由

的克星。记得有报道说，上次禽流感流行时肯德鸡门庭冷落，营业损失惨重。

图 5.5 是针对马英九的情绪词云，虽然有不少铁杆“支持”的声音，但总体批评声浪也不小，红红的一片（图示中红贬绿褒）。

对于所抓取的某话题的成千上万的优点缺点（pros/cons）或者不同的侧面和特点（aspects/features），我们常常展示前 n 项结果，利用 pie chart 来直观显示，以此来应对优劣长短不好归纳的困难，反映舆情中最突出的情绪背后的理由，以利于宏观把握舆情的重点。譬如，網民眼中馬英九之毁誉究竟如何？图 5.6 和图 5.7 列举了前 15 项褒贬馬英九的缘由。

VI. 社交媒体挖掘实例

本节给出围绕热点话题在应用层测试系统的实例，说明本文报告的自然语言技术已经投入实用。具体说，在多语言客户情报挖掘的产品开发

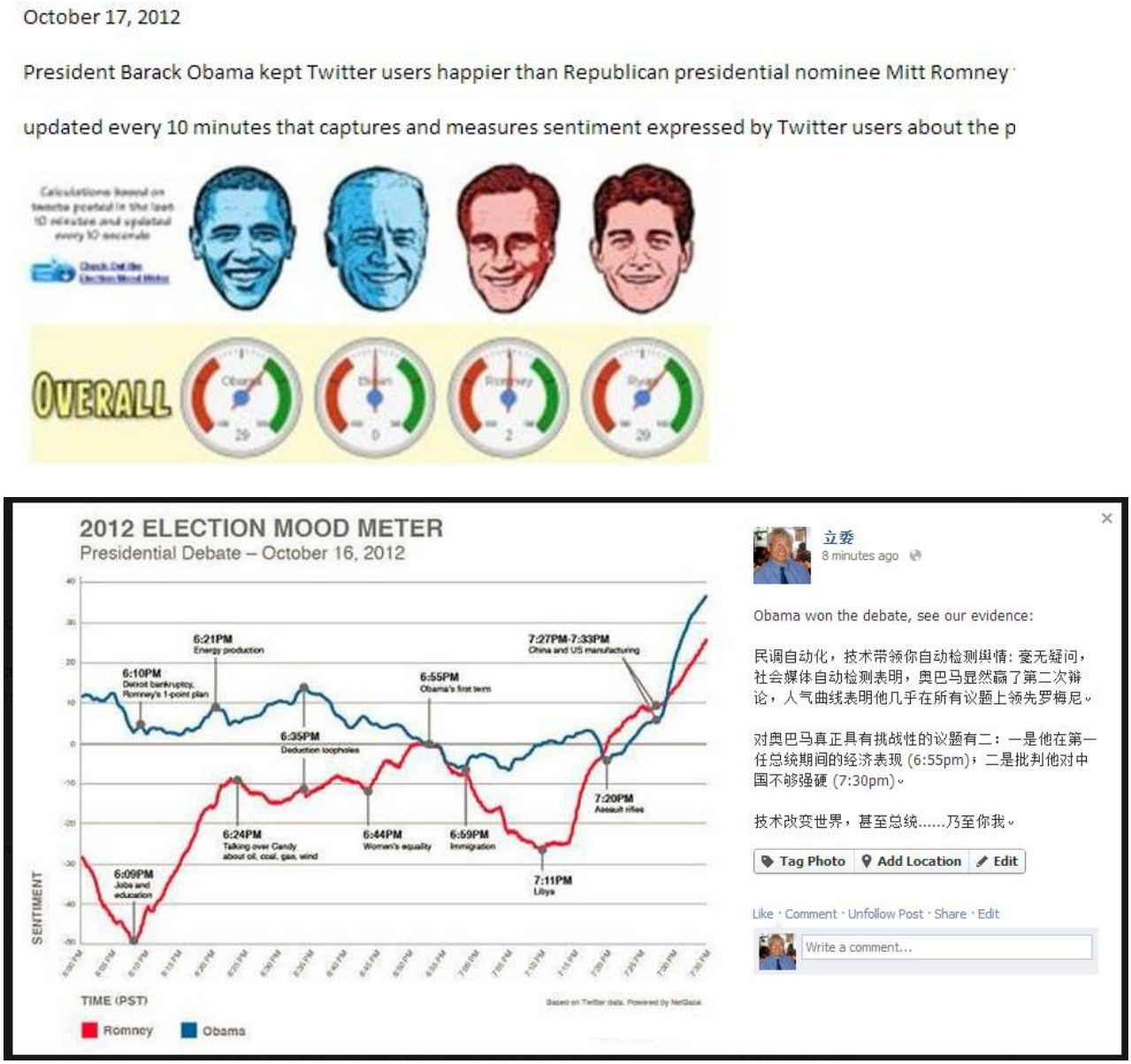


图 6.1：大选年美国总统辩论社交媒体实时舆情监测

中，以英文最为成熟，已经有很多 500 强客户，有不同数量客户使用的其他语言有西班牙语、德语、葡萄牙语、意大利语、法语和日语。中文是最近一加入的语言，但技术开发的成熟程度仅次于英语。由于中文社交媒体的主流微博和微信尚未正式纳入，因此目前只能是供内部测试的 alpha 系统。相信不久的将来可以解决好数据源的问题，届时正式发布产品。（说明：本文所有舆情挖掘和句法分析的测试结果都是本系统的实际输出，完全没有后编辑，错漏之处在所难免。）

6.1 实时舆情监测

先给一个英文社交媒体实时监控的真实案例，图 6.1 是民调自动化在去年大选年美国总统现场辩论时的现场使用，美国主要电视台曾予报道。精细到分钟的两条人气曲线（【舆情晴雨表】）反映了大数据实时舆情挖掘，比以前任何手工民调及时靠谱。以 twitter 为主的社交媒体实时监控的曲线图相当准确地反映了随着辩论话题而改变的舆情走势。

对奥巴马真正具有挑战性的议题有二：一是他在第一任总统期间的经济表现（6:55pm 始）；二是批判他对中国不够强硬 (7:30pm 始)。人气曲线在第一点经济表现的话题辩论时显示出奥巴马人气的下降和罗梅尼的急升，表明美国选民对经济迟迟不能大幅度复苏非常不满。但辩论到第二个与中国外交的话题时，奥巴马一度失利，随后两人的人气差不多同时上升，说明奥巴马的回应不错。这个话题对罗梅尼天然有利，因为美国社会一直弥漫着中国威胁论，害怕中国的崛起抢走美国的工作和机会，这种情绪要求总统对中国采取更强硬的态度。

6.2 中文社交媒体来源

我们转向中文社交媒体的热点话题挖掘（李维 2013）。首先说明一下挖掘的数据来源。

图 6.2 详细列举了数据源的分布。大体说来，本节展示的中文自动民调，除了另做说明以外，来自中文世界社交媒体过往一年的档案，简体文档三亿五千万。大约有一亿论坛帖子来自百度（贴吧等），两千多万来自搜狐，两千五百万来自天涯论坛。数据总量的 70%为论坛的帖子，15% 来自博客。遗憾的是最有影响力的微博微信数据暂时阙如，但由于所选都是热点话题，这些其他社交媒体的大数据也已足够反映民间舆论的概貌。

6.3 热点话题一：国际快餐品牌舆情比较

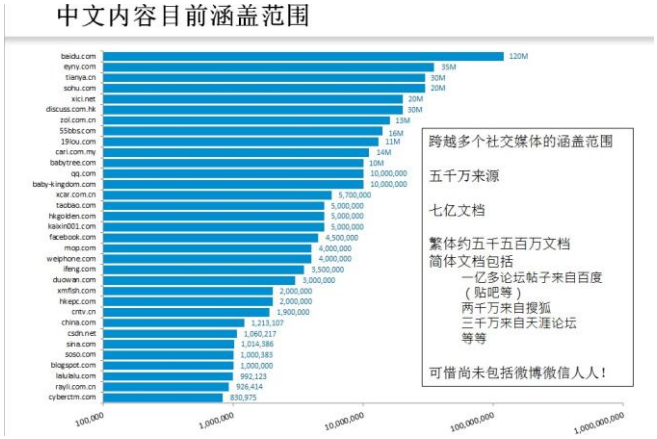


图 6.2：中文社交媒体来源分布

【品牌形象比较图】特别适合随时调查对比一个行业的竞争品牌的社会形象。一个行业大众热议的流行品牌之间的口碑人气和竞争态势一目了然。所选的四家快餐国际知名品牌是麦当劳、肯德基、必胜客和吉野家。

图 6.3 显示，肯德基最为人们热议（buzz），超过麦当劳。这一点与美国有鲜明对比，肯德基在美国本土根本无法与快餐的航空母舰麦当劳相提并论。这四家快餐店只有必胜客的形象还算正面，处于褒贬议论的中线上。其他三

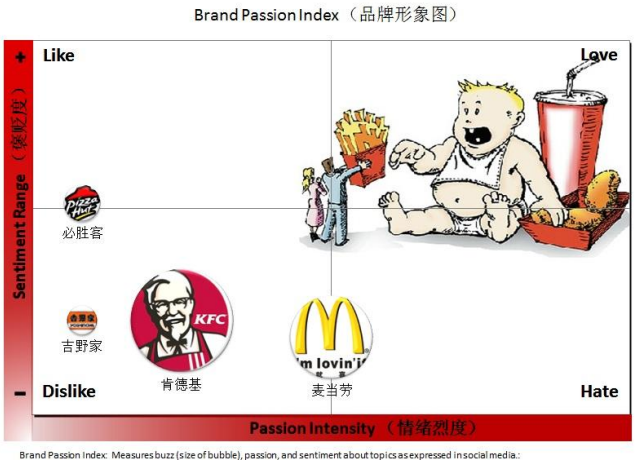


图 6.3：国际快餐品牌舆情比较图

家均在中线之下，表明客户的抱怨多于喜爱。在舆论强度的轴上，麦当劳刚好在中线上，表明讨厌它咒骂它的人都不少，其他两家（肯德基和吉野家）尽管总体形象也是负面的，但大家抱怨的强度不烈。必胜客呢，虽然总体形象不错，却与吉野家一样处于情绪强度的最左边，说明喜欢它和抱怨它也都不激烈。

曾几何时，以国际餐饮大王麦当劳为代表的西方快餐店纷纷进军中国市场。当年国门乍开，国人对西洋东洋的东西甚觉新鲜，清洁卫生规范快捷的外来快餐店在东土大受欢迎，一时门庭若市。但中国毕竟是舌尖上的中国，中国人对吃最挑剔，最讲究。在食这一块儿，要想长期扎下去赚钱，与本土的各种经济便餐以及农家小菜竞争，其实并不容易。外来快餐，首先是价格上没有优势，其次是口味太单调。从上图也可以看出，老百姓对这些外来餐饮不满多於喜爱，外来快餐店似乎风光不再。

KFC/肯德鸡 Summary 2013/2/4 - 2013/6/3 资料来源：主体为新浪微博和腾讯微博 2013 四月至五月					
品牌	提及	净情绪	强度	正面	负面
KFC	209,016	23%	68	19,348	11,990

图 6.4：肯德鸡微博舆情概貌

正好最近收集了新浪微博和腾讯微博 2013 年四月到五月的一个月数据样本，计一百八十六万微博，我们专门对 KFC（肯德鸡）在这批微博数据上做了详 尽民调，结果汇报如图 6.4。

值得注意的是，在微博数据里（图 6.5），肯德基的褒贬指数（净情绪）明显高过前面的其他社交媒体（论坛），达到正 23%，说明喜爱肯德鸡的人很不少。图 6.5 的微博数据样本令人印象深刻，满是网民在微博上的短小帖子，好几句

图 6.5：肯德鸡褒贬微博数据样本

图 6.6：肯德鸡情绪词云

都是比较句式，自动中文分析都对了。特别是那句“麦当劳还是不如肯德基好吃”很典型，里面有两个比较的品牌，这种令一袋子词的情感抽取头痛的现象，我们的系统因为有结构分析，表现不俗。

图 6.4 至图 6.11 展示了这次调查的方方面面，显示大数据被自动处理的情报丰富性。这些情报均立等可取。

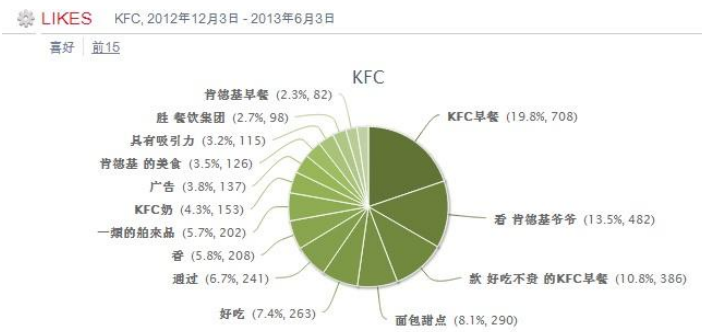


图 6.7：肯德鸡的主要优点

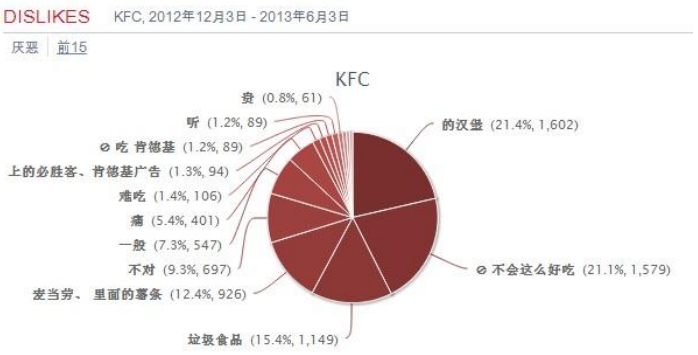


图 6.8：肯德鸡的主要缺点

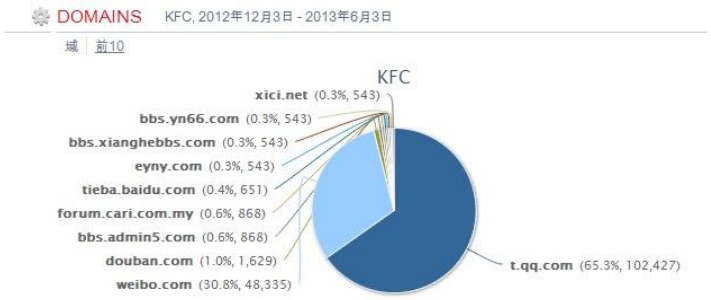


图 6.9：肯德鸡民调的数据构成
（65%腾讯微博，31%新浪微博，豆瓣贴吧等不足 4%）

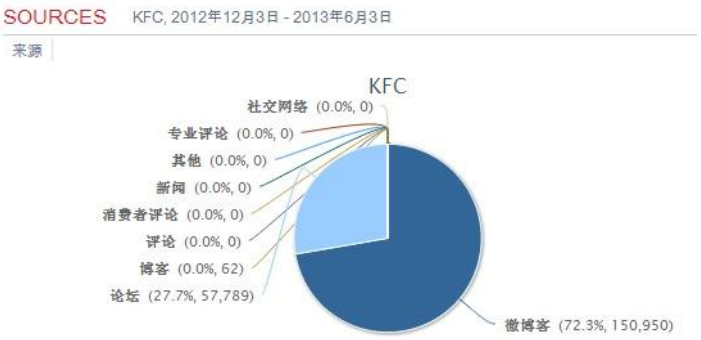


图 6.10：肯德鸡民调的数据种类分布
（微博类 72%，论坛类 28%）

图 6.11：议论肯德鸡最多的网民 IDs

6.4 热点话题二：中国第一夫人彭丽媛

图 6.13 是社交媒体热议第一夫人的实例（产品中叫 soundbytes）样本。自动挖掘的好处之一是用户可以对挖掘到的任何情报随时查验（drill down）社交媒体，看网民究竟是怎样议论的，也可以帮助系统改进质量。

以前都是媒体热议西方第一夫人如何光彩照人，如今终于可以一睹一议新中国自产的第一夫人的风采了，网民的热情一浪高过一浪。自宋家姐妹以来，还没有一位中国女性受到如此多的聚焦和赞美（见图 6.14 的网民情绪词云）。

图 6.12 是 2013 年前三个月的第一夫人的人气走势图，展示了大众对第一夫人的热度（净情绪）曲线，多数时候居高不下，更在二月 20 号左右达到 100%的顶点。如此的高评价，在对各种品牌和人物所做的系列自动媒体调查中，是绝无仅有的。

图 6.15 是一年以来社会媒体对第一夫人具体

但是从来没有像彭丽媛那么受关注。享誉盛名的中国艺术家歌唱家彭丽媛。 最开

bbs.mzsky.cc

美丽大方气质! 支持彭丽媛!

club.china.com

彭丽媛做中国第一夫人, 真是为国争光, 我敢说, 当今世界没有一个"第一夫人"能超过

club.china.com

第一夫人坚定且深情, 真有范儿, 有气质! 好喜欢彭丽媛, 美丽, 高贵, 大气。

图 6.13：彭丽媛社媒舆情样本

图 6.12：中国第一夫人人气曲线

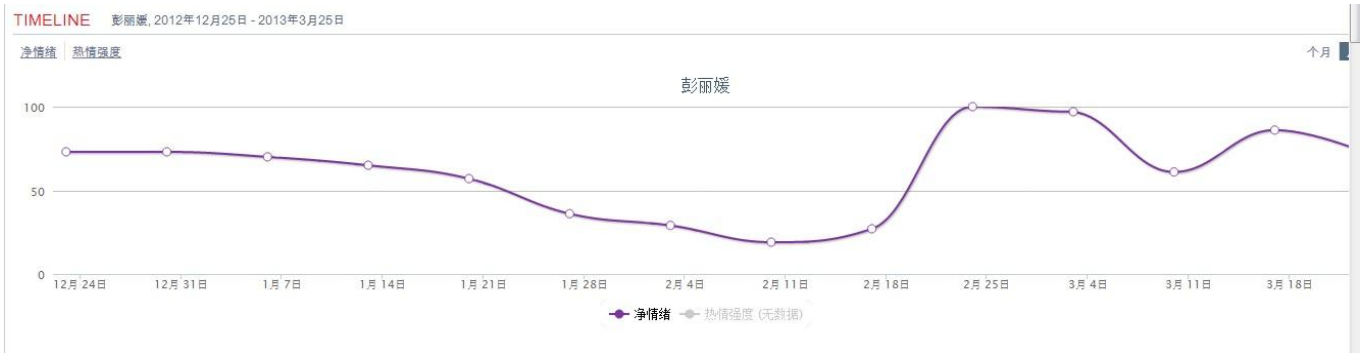


图 6.15：第一夫人的优点缺点词云

评价的词云，几乎一面倒的赞誉。唯一一条比较显眼的批评是彭丽媛不知何时何地曾经“穿肥大的军裤”，似乎影响了人们心目中的优雅形象。

除词云外，舆情也可以用传统的 pie-chart 图示。图 6.16 回答了网民为什么喜欢第一夫人的前 15 条理由，以形象好为主，也提及她的知名度和为人处世，包括姣好（41%）、聚焦和 知名的歌唱家（21%）、漂亮和亮丽（20%）、微笑（4%）、不俗（3%）和低调（3%）。

图 6.16：喜欢第一夫人的主要理由

6.5 热点话题三：台湾政坛舆情图和晴雨表

测试中文舆情挖掘的繁体系统，借此了解一下台湾政坛舆情。舆情图 6.17 旨在计量社会公仆在社会媒体中的被关註度、褒贬度和爱憎情绪强度，反映其公众形象。

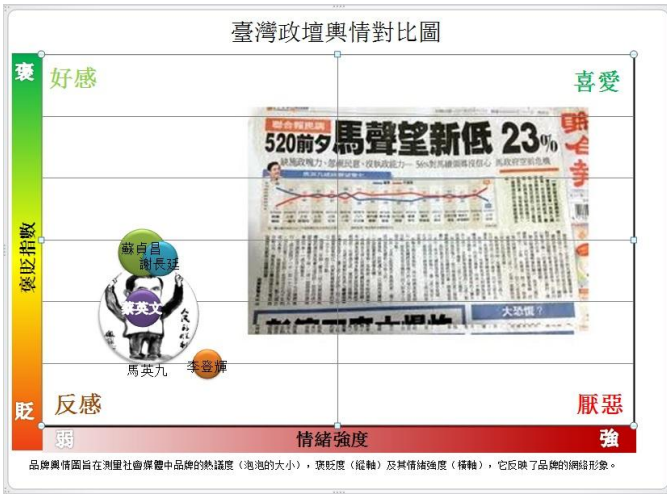


图 6.17：台湾政客舆情对比图

一眼看去，台湾的蓝绿政客几乎全部挤在舆情图的左下角反感（弱贬）的角落，说明他们在民众中的形象都不太好。不仅如此，大家对他们的情感也不强烈，大概是失望已久，又没有其他备选项，已经疲怠了。仔细比较，可以看出，苏贞昌名声最佳，毫无疑问是这次自动民调中矮子丛中的将军。谢长廷紧随其后，然后才到蔡英文和马英九。蔡（指数 19）比马（指数 18）略高，但由于是当选总统，马的议论最多（泡泡最大）。老总统李登辉的声望则日落西山，更在马蔡之下。至于阿扁前总统，名声太差，淨情绪指

台湾政坛舆情自动民调: 资料来源：一年的社交媒体文档（正體）12/2/22-13/2/23			
話 題	提 及	淨情緒	強 度
馬英九	303,137	18%	18
陳水扁	76,184	-12%	17
蘇貞昌	68,717	28%	18
蔡英文	41,707	19%	17
謝長廷	36,924	27%	20
李登輝	26,234	10%	28
連 戰	18,427	36%	34
肖萬長	7,835	33%	29
宋楚瑜	5,611	19%	39

图 6.18：台湾政客舆情一览

标-12，根本浮不上舆情图的臺面。连战、萧万长、宋楚瑜也未能浮现舆情臺面，原因不是被民众唾弃，而是被民众遗忘。他们根本就没有多少议论，泡泡太小。

其实，论褒贬指数，连战的淨情绪 36 才是冠军，萧万长也有 33，二者均高出苏贞昌的 28 一头。详细数据比较见图 6.18。

回看【马英九施政一年舆情晴雨表】（图 5.2），其看点及分析如下：（1） 一年来马英九的总体形象偏低，淨情绪指标在零度以下居多。他一直试图改善形象，但总也不大成功。究其原因，凡臺上的政客，除非社会经济出现奇迹般改善，作为常规，总是招致的批评远多于赞扬。民众总是愤怒的，而在野党不会放过任何一个机会推波助澜。（2）不过一年来也有 10 多次短暫的亮点，声望处于零度以上（褒大于贬），虽然都好景不长。从图上看，去年七月初到九月初之间是正面声望持续最长的区间（只在八月短暫跌入零度以下），不知道有什么亮丽的政治表现还是由于团队公关得力，有兴趣的读者可以查證一下。马总统宣誓就职的五月中，淨情绪指标尚在零下 30 度左右徘徊，怎么到了七月就迅速回暖至零度以上，持续约两个月，直到九月 2 日的 +35 的峰值。这个区间似乎确是马总统二度当选以来得到民众认可的最佳时期。此后就一蹶不振，只在十月、十一月与今年元月短暫回升。一年来的最低点在叁月四日的零下 44，十二月 16 日一度跌入零下 42。

总而言之，马英九自从去年初当选以来，不是很顺，民众的失望抱怨情绪弥漫网坛。图 6.19 从总体上来对比一下马施政一年来的得失荣辱。从数据点上看，还是红色负面评价居多。

正面评价中最大的亮点是『清廉』，负面评价除了『无能』外，最大的批评就是『一意孤

图 6.19：对马英九的褒贬词云

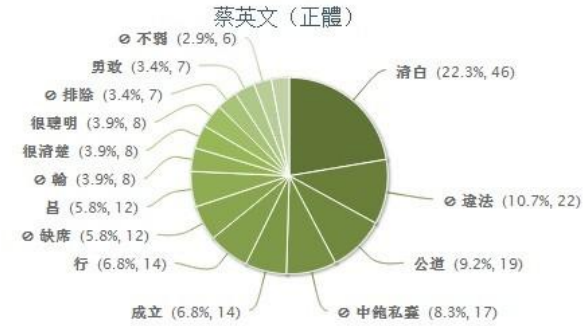


图 6.20：蔡英文的优点

图 6.21：谢长廷的优点

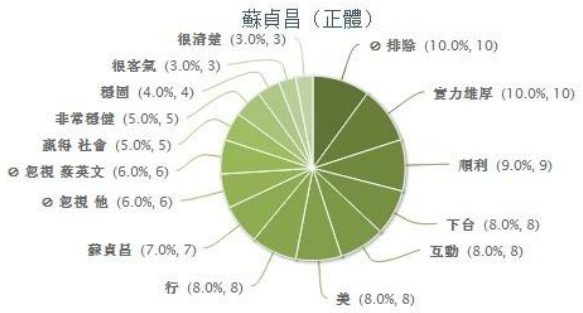


图 6.22：苏贞昌的优点

行』和『畏惧中国』（嫌他在两岸关系中对大陆不够强硬）。总之是，臺湾民众比较烦。马总统要想赢得民心，光靠已有的清廉勤政形象远远不够。他要青史留名，改变无能总统的批评浪潮，扭转其颓势，今后几年一定要交出一份更大的成绩单。

而在野的台湾绿营大佬除了民众舆论已经贬出局的在押阿扁外，主要包括蔡谢苏（蔡英文、谢长廷、苏贞昌），褒贬指数都比马总统高。具体来看，蔡英文为人形象不差，普遍认为她为人“清白”、“公道”，“（没有）违法 ”、“（没有）中饱私囊”。谢长廷的长处主要就是一个“行”字（80%）。苏贞昌则“实力雄厚”而且“ 顺利”。如图 6.20-6.22。

6.6 热点话题四：“剩女”问题

我们的系统本来是为了挖掘品牌的客户情报而设计的，但是我们也不妨尝试其他话题的挖掘。社交媒体大数据里面关于品牌的舆情只是各种话题之一，其他的热点话题除政客和名人外，也包括引起热议的任何事件或概念。前面说过，与以机器学习为技术基础的舆情挖掘不同，以句法分析为基础的舆情挖掘对于不同 的话题移植性能较强。这是因为机器学习的方法对于所训练的数据有较强的依赖性和调适性，一旦转移数据、领域或话题，系统往往无所适从，难以保证挖掘的质量。句法分析为基础的规则系统则不同，作为技术基础的句法分析是独立于领域（domain-independent）的纯粹语言学模块，只要使用的是中文，无论哪个领域的话题，分析抽取原则上依然可以进行，对质量的影响不大。这一点已经为我们测试过的五花八门的话题挖掘所印证，这些话题包括：爱情、婚姻、切糕、春晚 ……， 还有“剩女”。

BBC 曾有新闻报道中国的所谓剩女问题，于是想到做一个自动社会调查，看看这个话题可以挖掘到什么。剩女这词对女同胞有侮辱之嫌。然而，剩女成为热词已经多年，它反映的社会问题是我们躲不开的。人类社会走到今天，男婚女嫁的自然关系变得越来越复杂。虽然婚姻的性爱基础不变，但从物质和精神上看，都是相互越来越挑剔。可是男女性别带来的心理差异以及男女客观社会地位的差异，使得现代社会（不仅限于中国）比古代更加注重男才（财）女貌，为此才（财）男宁愿下娶。可是女性的心理及其社会压力总是不能下嫁，结果造成了高学历大龄女青年与低学历男光棍的双重挤压，成为巨大的社会问题。这件事的尴尬在于，剩女和剩男处于社会的两极，虽然都躁动不安，向往婚姻，却天差地别，无法成双。我们还是看看社交媒体中的议论吧。

如图 6.23 所示，最突出的一点是，很多人认为剩女问题是一个『伪问题』。这应该不是否认社会问题的存在，而是强调这不是剩不剩的问题，而是具有更深更复杂的社会成因。高学历、高收入者果然与剩女紧密相连：这不但印证了我们每日所见的现实状况，而且指明了问题的主要社会原因：女方追求高学历的代价往往是耽误了自己的终身大事。在古代就没有这个问题，女子无才便是德，一般女子及其家庭对女子的教育都不很重视，社会对女子的期望也多在相夫教子、三从四德上，并不很在乎女子的才学。那样的环境下，一个长相平平的女子一般也会早早嫁给一个人家，根本等不到大龄的那一天。如今的时

图 6.23：剩女的得失词云

代，男女平等了，女性的自我意识和社会抱负也相应增长，于是越是聪明的女子，越不甘落后。可是，现代社会男性的择偶标准却依然停留在以前的观念上，依然是重貌重德远胜重才。其结果可想而知。

有意思的是，社会媒体对于“剩女”优缺点的调查其实是针对两个不同的概念。优点方面大多说的是“剩女”本人，因为这些所谓剩女其实资质才学都很不错，除了年龄偏大相貌平平外，剩女自身集中了相当多的优点，如图 6.24 所示。

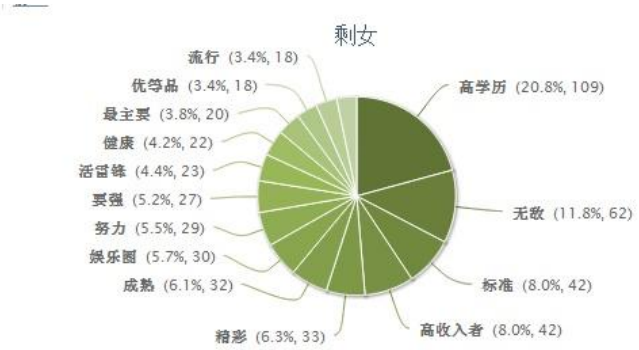


图 6.24：剩女（群体）的优点

可是，当调查剩女的缺点（图 6.25）时，社会媒体的指向却有所改变，议论的对象从剩女这个群体，大体转移到了社会。无论定性为“伪问题”也好，还是议论造成剩女的社会因素，主要着眼点都不仅仅是这个群体本身的问题。“没有时间”是忙于学业和工作的剩女的主要问题之一，但实际上也是社会使然：现代社会竞争日趋激烈，时间在学历和奋斗中很容易飞逝。就算忙里偷闲约会几次，恋爱几段，稍有差池，转眼 30 仍滞于婚姻殿堂以外自然不在少数。

这次调查所反映的社会情绪，与我们平时的印象基本吻合。不过，看到这些舆情图示，还是有一种警示。问题是突出的，可是解决问题的方案却不甚了了。似乎也没有什么灵丹妙药医治这个现代社会的顽疾。

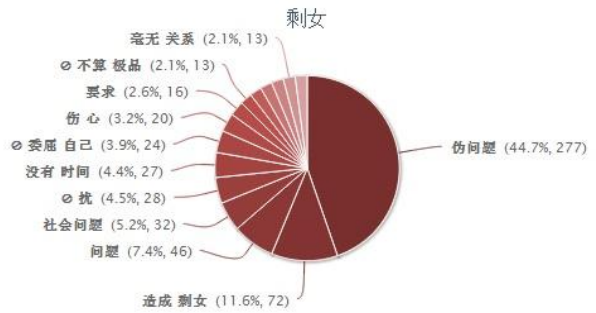


图 6.25：剩女的问题

6.7 热点话题五：中国手机市场

关于中国手机市场，自动挖掘的结果颇令人意外。本来以为社会舆论一定是爱憎分明，各家手机品牌的社交媒体形象相差应该很大，比如爱疯（iPhone）应该是贵族，高高在上；小米下里巴人应该居于下风。结果呢，居然大家都挤在一起，社会形象大同小异，如图 6.26。

图 6.26：手机品牌中国市场舆情一览图

苹果作为龙头老大，在这次自动调查中只是 buzz 比其他几家要大很多，因为有更多的人在社交媒体中谈论它，但是其褒贬的比例（净情绪）和烈度（intensity）与其他几家基本处于对等地位，并无迹象表明它得到用户特别的青睐或抱怨。具体说来，市场对这几家流行的手机都不大满意（净情绪在中线之下），但褒贬的情绪并不

激烈（几家手机全部集中在情绪烈度的左边）。除了爱疯被议论最多外，其他几家的 buzz 都差不多（泡泡的大小差不多）。有意思的是，HTC 的客户评价度居然略高于苹果爱疯，三星离爱疯只有半步之遥，大有后来居上之势。小米和诺基亚虽然低于前三位，但也相差不远。小米作为下里巴人，也只比各位略微低了小半头。总而言之，社会形象彼此彼此。照目前看来，手机市场不但离一统天下还很遥远，甚至三足鼎立也不可及，仍然处于军阀混战。每个手机品牌的数据点均在 40 万到一百六十万之间。这样的样本比起常常只有几千份数据点的手工客户调查高出两到三个量级，可置信度不可同日而语，这是大数据的力量。

最近收集了新浪微博和腾讯微博 2013 年四月到五月的一个月数据样本，计一百八十六万微博，我们专门对 HTC 在这批微博数据上做了更详尽的自动民调，结果如图 6.27。

HTC Summary 2013/2/4 - 2013/6/3 资料来源：主体为新浪微博和腾讯微博 2013 四月至五月					
品牌	提及	净情绪	强度	正面	负面
HTC	526,890	82%	61	213,332	21,371

图 6.27：HTC 微博舆情一览

图 6.28：HTC 品牌微博舆情样本

从样本（图 6.28）上看，舆情抽取质量不错。譬如正规文体中不见，但微博中流行的情绪表达法“HTC 好屌！”，也抓住了。

图 6.29：HTC 情绪词云

图 6.30：HTC 长短优劣词云

VII. 测试结果和讨论

系统舆情挖掘的质量检测（QA）是这样进行的。由于本系统主要是为挖掘品牌的客户情报而研发，我们通常选取各主要行业有代表性的 10-20 个品牌名作为系统的输入，然后测试舆情情报结果的平均质量。我们使用 CrowdFlower 的问卷服务来做检测。每个结果提交四个匿名判

员，必须至少达到 75%（四人中至少三人）的一致意见才计入计算。最近的测试选取下列 15 个品牌作为情报搜索对象：

iPhone|中国电信|丰田|伊利|南航|可口可乐|宝马|家乐福|必胜客|携程|淘宝|苏宁|茅台|蒙牛|麦当劳

下表（图 7.1）是最近两次检测的结果和对比：

Sentiment	Release 1	Release 2
Positive	81.4%	73.2%
Negative	89.8%	89.1%
Overall	85.7%	79.6%

图 7.1：Precision Change Over Last Two Releases

在精度与覆盖面指标的动态平衡中，我们以保持精度 80%作为底线，因为这是多数客户的要求。在 80% 左右的精度上，力求最大的覆盖面。图 7.2 可以看出，前一个 release 比较保守，精度高达 85.7%，于是我们做了一些松绑的尝试，增加了一些条件宽松的默认规则，结果精度降低到 79.6%，但是覆盖面大幅度提升，翻了约三番（298.7% - 417.1%），正面指标达到 14.5%，负面指标达到 5.8%，说明这次调整合理合算。

这里的覆盖面指标不是严格意义的召回率（recall），而是其间接反映。由于在随机选取的真实连续文本里面，情感句子相对稀疏，灰色现象也不少，想要标注一个足够大的，有代表性同时多人意见一致的测试库来精确计算绝对召回率是很困难的，经济上不太可行，维护起来也是负担。我们转而使用一种相对的覆盖面指标（relative recall）来检测系统覆盖面方面的进展。这个指标的公式如下：

1. Brand retrieval:
   品牌情感抽取的句子数 / 品牌出现的总句子数
2. Precision: 系统的平均精度
3. Sentiment concentration:
   情感句子数 / 所有的句子数

*Relative Recall = (Brand retrieval * Precision)/ Sentiment concentration*

图 7.2 里的相对覆盖面指标看上去很低，但与英语指标（6%-14%）相若，说明中文覆盖面并不差，相同数据量中文英文所抽取的情报量大体相当。英语是我们开发最成熟的语言，投入应用时间最长，开发的规则最多，边边角角磨了很久了，精度和覆盖面均让用户满意。当然，我们也注意到语言间的差别，单从情感词汇的统计（基于现代汉语基本词汇五万左右，英语基本词汇四万四千），现代汉语褒义词是英语褒义词数量的三倍，汉语贬义词也是英语的两倍以上，这从侧面反映了汉语的情感语句的密度可能远远高于英语。因此，汉语情感抽取的覆盖面在今后的进一步开发中应该还有不少改进余地。

	Release 1	Release 2	Change
Positive	2.8%	14.5%	417.1%
Negative	1.5%	5.8%	298.7%

图 7.2：Relative Recall Change

很长一段时间，学界测量一个系统，使用的是两个指标：精度（precision, 即抓到的有多大比例是抓对了的）和召回率（recall，即所有该抓到的有多大比例真地抓到了）。由于自然语言的歧义性和复杂性以及社交媒体的随意性和不规范，要想编制一套两项指标综合水平（所谓 F-score）都很高的系统非常不容易。但是，研发实践发现，自然语言系统能否实用，很多时候并不是决定于上述两个指标。还有一个更重要的指标决定

着一个系统在现实世界的成败，这个指标就是系统对于大数据的处理能力，可以不可以真正地scale-up 到大数据上。

由于电脑业的飞速发展，云计算技术的成熟，大数据处理在现实中的瓶颈往往是经济上的羁绊，而不是技术意义上的难关，其结果是革命性的。在处理海量数据的问题解决以后，精度和广度变得相对不重要了。换句话说，即便不是最优秀的系统，只有平平的精度（譬如 70%，抓100 个，只有 70 个抓对了），平平的召回率（譬如 30%，三个只能抓到一个），只要可以用于大数据，一样可以做出优秀的实用系统来。其根本原因在于两个因素：一是大数据时代的信息冗余度；二是人类信息消化的有限度。覆盖率的不足可以用增加所处理的数据量来弥补，这一点相对容易理解。既然有价值的信息，有统计意义的信息，不可能是“孤本”，它一定是被许多人以许多不同的说法重复着，那么覆盖率不高的系统总会抓住它也就没有疑问了。从信息消费者的角度，一个信息被抓住一千次，与被抓住一百次，是没有本质区别的，信息还是那个信息，只要准确就成。问题是一个精度不理想的系统怎么可以取信于用户呢？如果是 70%的系统，100 条抓到的信息就有 30 条是错的，这岂不是鱼龙混杂，让人无法辨别，这样的系统还有什么价值？沿着这个思路，别说 70%，就是高达 90%的系统也还是错误随处可见，不堪应用。这样的视点忽略了实际的挖掘系统中的信息筛选（sampling）、整合（fusion/merging）与排序（ranking）的环节，因此夸大了系统的个案错误对最终结果的负面影响。实际上，典型的情景是，面对海量信息源，信息搜索者的几乎任何请求，都会有数不清的潜在答案。即便有一个完美无误的理想系统能够把所有结果，不分巨细都提供给信息消费者，他也无福消受（所谓 information overload）。因此，一个实用系统必须要做筛选整合排序，把统计上最有意义的结果呈现出来。这个筛选整合排序的过程是我们系统挖掘层的一部分，它使得最终结果的质量远远高于系统的个案质量，极大改善了用户体验。总之，多了就不一样了，大数据改变了技术应用的条件和生态。

VIII. 结语

舆情是什么？人民的声音。人民是由个体人组成的。但白马非马，人非人民。人民的呼声通过冗余才能听得见，否则就不是人民的声音，只是可有可无、可以忽略、听不见也不必听见的个体意见。对于大数据，缺失部分数据不是大问题，只要这种缺失对于要挖掘的话题或品牌没有针对性。缺失数据的原因很多，譬如，服务器或数据库故障，由于成本考量只取一定比例的样本，还有垃圾过滤系统的误删，当然也有系统本身覆盖率的不理想，等。总之缺失是常态，而求全则是不现实也是不必要的。大数据追求的是有影响力的信息和舆情动态，而这些原则上都不会因为数据的部分缺失而改变，因为动态和影响力的根基就在信息的高冗余度，而不是大海捞针。重要的是，冗余本身也是情报的题中应有之义。这与同一个情愿诉求为什么要征集成千上万的签名道理一样，至于最终是 10 万签名还是 9 万五千人签名了，完全不影响诉求的内容及其整体效应。可以说，大数据加速了技术的应用。

在大数据和移动互联网时代，随着社交媒体的深入民间，民间情绪和舆论的表达越来越多。因此，语言技术支撑的舆情挖掘势必成为支持决策的基本工具，应用前景广阔。政府需要它来了解民意，调整政策；企业需要它来听取客户情报，维护品牌的声誉；普罗百姓也可以用它来调查任何品牌和机构的总体评价和走势，避免盲目购买或投资。作为中文社交媒体舆情应用的初步尝试，本文全面展示了这样一个系统的概念设计、技术基础、架构和应用示例。

致谢

感谢 Ray、Cheng-Ying 为中文系统研发提供 NLP 平台支持，感谢质量检测组 Xin 为情感抽取提供的质量测试报告。还要感谢 Zixin、Min Martin、Sophia 对于中文系统数据、词典以及规则开发做出的贡献。词典资源亦用到“知网”（HowNet）情感分析用词语集，句法分析也用到知网的部分语义分类（董强等 2003）。

参考文献

[1] G.A.Bruder and W.M. Janyce 1990. A psychological test of an algorithm for recognizing subjectivity in narrative text. In Proc. 12th Annual Conference of the Cognitive Science Society, 1990: 947-952.

[2] N. Chinchor and E. Marsh 1998. MUC-7 information extraction task definition (version 5.1). Proceedings of MUC-7.

[3] D. Davidov, O. Tsur and A. Rappoport 2010, “Enhanced Sentiment Learning Using Twitter Hashtags and Smileys,” Proceedings of COLING 2010: Poster Volume, pages 241–249, Beijing, August 2010..

[4] X. Ding, B. Liu, and P.S. Yu 2008, “A holistic lexicon-based approach to opinion mining,” Proceedings of the international conference on Web search and web data mining (WSDM ’08). NY, USA. 2008:231-240.

[5] J. Han 1999. Data Mining. In: J. U. a. P. Dasgupta (editor), Encyclopedia of Distributed Computing. Kluwer Academic.

[6] J. R. Hobbs 1993. FASTUS: A system for extracting information from text. Proceedings of the DARPA workshop on Human Language Technology, pp. 133–137. Princeton, NJ.

[7] W. Li 1989. A Dependency Syntax of Contemporary Chinese, BSO/DLT Research Report, the Netherlands.

[8] G. S. Linoff and M. J. Berry 2011. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Wiley Publishing, Inc. 2011

[9] R. McDonald, K. Hannan, T. Neylon, M. Wells and J. Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In ACL 2007.

[10] B. Pang, L. Lee and S. Vaithyanathan 2002 Thumbs up? sentiment classification using machine learning techniques. In EMNLP 2002, 79–86.

[11] E. Roche and Y. Schabes 1997. Finite-State Language Processing. MIT Press.

[12] M. Silberztein 1999. INTEX: a Finite State Transducer toolbox. Theoretical Computer Science 231(1). Elsevier Science.

[13] R. Srihari and W. Li 2006. Question Answering Supported by Multiple Levels of Information Extraction. In T. Strzalkowski and S. Harabagiu (eds.), Advances in Open- Domain Question Answering. Springer, 2006, ISBN:1-4020-4744-4.

[14] R. Srihari, W. Li, T. Cornell and C. Niu 2006b. InfoXtract: A Customizable Intermediate Level Information Extraction Engine. Journal of Natural Language Engineering, 12(4), 1-37

[15] C. Strapparava and R. Mihalcea. 2008. Learning to identify emotions in text. In SAC Symposium on Computational Approaches to Analysing Weblogs. AAAI Press

[16] M. Taboada, J. Brooke, et al 2011. “Lexicon-Based Methods for Sentiment Analysis,” Computational Linguistics, 2011,37(2): 267-307.

[17] I. Titov and R. McDonald 2008. Modeling online reviews with multi-grain topic models. In WWW, pages 111–120, New York, NY, USA. ACM.

[18] B. Tsou, et al 2005. “Polarity classification of celebrity coverage in the Chinese Press,” Proceeding of the 2005 International Conference on Intelligence Analysis. 2005:137-142.

[19] P. D. Turney 2002, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. USA.2002:417-424.

[20] J. Wiebe 2000. Learning subjective adjectives from corpora. In AAAI.

[21] T. Wilson, J. Wiebe and P. Hoffmann 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics, 35(3):399–433.

[22] J. Yang and M. Hou 2012, “基于规则的句子语义倾向计算,” Proceedings of the 13th Chinese Lexical Semantics Workshop (CLSW2012), Wuhan, China. 2012.

[23] H. Yu and V. Hatzivassiloglou 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. EMNLP

[24] 李维 2013，“走进‘大数据’- 洗衣机寻购记，”《硅谷简讯》，2013 年第七期第八期

[25] 董强，郝长伶，董振东，2003， 基于《知网》的中文语块抽取器，语言计算与基于内容的文本处理（全国第七届计算语言学联合学术会议论文集），孙茂松、陈群秀主编，清华大学出版社

[26] 朱嫣岚，闵锦，周雅倩，黄萱菁，吴立德 2006，“基于 HowNet 的词汇语义倾向计算，”中文信息学报，2006，20(1)：14-20.

[27] 李钝，曹付元，曹元大，万月亮 2008，“基于短语模式的文本情感分类研究，”计算机科学， 2008，35(4)：132-134.

[28] 王素格，杨安娜，李德玉 2009， “基于汉语情感词表的句子情感倾向分类研究，”计算机工程与应用，2009，45(24)：153-155.

[29] 党蕾，张蕾 2010， “一种基于知网的中文句子情感倾向判别方法，”计算机应用研究，2010，27(4)：1370-1372.

[30] 赵妍妍，秦兵，车万翔，刘挺 2010， “基于句法路径的情感评价单元识别，”软件学报，2010，21(8)：1834-1848.

Social Media Emoticons for Brand Sentiments

Martin Min, Tanya Lee, Margaret Zhang, Lei Li
Netbase Solutions, Inc.
Mountain View, USA
cmin,tlee,ttang@netbase.com

Abstract — Automated sentiment extraction from social media is enabling technology to support mining online consumer insights. From the input side, it is observed that social media as a sub-language often uses emoticons mixed with text to show emotions. Most emoticons, e.g. :=), are not natural language words, but textual symbols using characters to present a smiley face. Intuitively, such symbols are innately associated with emotions, whether happy, annoyed or don't care, hence important clues for helping sentiment analysis. Previous research has involved the limited use of emoticons as noisy labels in sentiment learning but detailed study on how noisy or useful they are has not been done. This paper presents a comprehensive data analysis of the role of emoticons in sentiment classification, focusing on its impact in the context of a brand. The study shows that emoticons alone are not sufficient for determining sentiments. Nevertheless, emoticons serve as an additional factor for extracting sentiment in cases where other linguistic clues are insufficient. A further study also discovers a use of emoticons as counter evidence to effectively block glaring errors in existing sentiment analysis. Finally, a recall study finds that emoticons' overall contribution to recall is limited; nevertheless, it is low-hanging fruit and valuable.

Keywords **-** emoticon; sentiment analysis; sentiment classification; customer insight; social media

摘要—从社交媒体自动提取网民情绪是客户情报系统的技术基础。除了文字，网民也经常使用网络表情符号（emoticons）来表达情绪，譬如笑脸 :=)，因此，表情符号自然应该成为情绪分析的线索。文献中有试图使用表情符号为文本自动标注情感（这种做法其实有问题），或者使用它作为机器学习的一个因素。但表情符号究竟可靠与否，它对于情感分析的质量有多大影响，迄今缺乏深入研究。本文报告了我们的详细数据分析，研究表情符号对于情感分类在精确度和召回率两方面的潜在作用，着重于社交媒体以品牌为中心的情感文句。调查表明，单单靠表情符号来做情感分类精确度远远不够，但不一致的部分多在灰色区域。进一步分析发现，表情符号的整体贡献是有限的，指望它大幅度提高情感分析的覆盖率并不现实。然而，合理利用情感符号对情感分析是有益的。此外，表情符号可以弥补其他语言线索的不足。这项研究还发现了一个利用表情符号作为反面证据的方法，有限制地使用这个方法可以有效防止系统情绪分析中正反混淆的惹眼错误。这类错误往往导致用户对系统质量失去信心，因此这个方法的发现在实践中有其特别的意义。

关键词：表情符号；社会媒体；情感分析；情感分类；客户情报

I. Introduction

The rapid growth of social data from such online social networks as Twitter and Facebook has aroused enormous interest in the mining and discovery of customer insights. For instance, Twitter has over 500 million registered users as of 2012, generating over 340 million tweets daily, which is equivalent to 3,935 tweets per second. Such huge amounts of unstructured text data involve enormous amount of customer voice that is invaluable to leading brands and businesses. They need to monitor, react, engage, and publish at the speed of social in real time in order to compete. The main motivation behind the adoption of social media intelligence tools like ours lies in the fact that people are increasingly sharing their opinions on products and services on social networks. Recent estimates indicate that on average one in every three blog posts and one in five tweets involve comments on products, services or brands (Hogenboom1 et al. 2013). Random users freely talk about whether they love or hate a brand, and lots of times they compare it with other brands in the same category. Apparently, such information would be really important for businesses to keep track of consumers' attitudes toward their brands and the management can make faster decisions based on the social intelligence when it is extracted from the huge social data pool and analyzed properly.

Sentiment analysis is an essential part of a commercial social media intelligence system and provides the core functionality for sentiment extraction from unstructured data. The majority of sentiment analysis systems are machine learning based, taking a traditional text classification approach to train models like Naïve Bayes, Maximum Entropy or Support Vector Machine. The text unit for classification is usually a document or paragraph consisting of multiple sentences. Typical examples of such data are movie or product reviews. When trained on domain data, those classifiers generally achieve over 80% precision for coarse grained sentiment classification as positive, negative or neutral (Pang and Lee 2008). While machine learning based sentiment classification works well in domain data at document or paragraph level, it faces challenges in handling short messages (e.g. tweets posted by mobile users).

Social media as a sub-language is sometimes full of emoticons mixed with text to show emotions of the poster. Most emoticons, e.g. :=), are not natural language words, but textual symbols using characters. Among various definitions, Wikipedia defines emoticon as "a meta-communicative pictorial representation of a facial expression … draw a receiver's attention to the tenor or temper of a sender's nominal verbal communication, changing and improving its interpretation". It mainly uses a combination of punctuation marks to mimic a smiley face to express a person's feeling or mood. Intuitively, people would assume that such symbols are innately associated with emotions, whether *love, hate* or *don't care*, hence important clues for helping sentiment classification. With the rapid growth of social media uses globally, it is generally recognized that emoticons have been playing an increasingly important role in online communications, especially for the younger generation. In light of this, a natural language system built for sentiment analysis is expected to take advantage of this special phenomenon, which is typically not in the scope of the grammar or vocabulary research of a language.

Unlike the main stream sentiment analysis based on statistical models using machine learning, we have built a rule-based high precision sentiment analysis system based on a parser for mining consumer insights from social media. This system is designed to address two

major challenges of machine learning systems: (i) sentence-level sentiment classification for short messages; (ii) extracting reasons behind sentiments to answer why questions. The research on emoticons reported in this paper is motivated by the need to enhance (i) in the context of mining customer sentiments towards a brand. But the analysis and experiments we have done also serve the purpose of enlightening the researchers in both machine learning world and the rule world with better understanding of the role of emoticons in sentiment analysis. In fact, it reveals a potential pitfall facing some earlier researchers (e.g. Read 2005) who assume emoticons are handy and reliable sentiment indicators and therefore use them as a gold standard to collect training corpus for sentiment classification.

The major contribution of this study lies in the fairly comprehensive study of emoticon's role in a sentiment analysis system. We aim to accomplish three specific goals:

- Provide a description of our sentence level sentiment extraction system built upon a robust natural language deep parser, focusing on how parser-supported sentiment extraction overcomes the obstacles from most machine learning based sentiment classification systems
- Provide a statistical analysis of how emoticons are used in social media from various perspectives
- Provide an evaluation of emoticon’s roles and contributions in our sentence level brand-oriented sentiment analysis.

Part of our initial study was presented at a symposium (Min, Lee and Hsu 2013) and further investigation with new data has been continuing since then.

The remainder of the paper is organized as follows. In Section II, we first review two types of related work: the work on sentiment analysis, with the focus on the issues arising from traditional statistical classification and the work on emoticons used as a resource for sentiment analysis. Section III describes our sentence level deep sentiment extraction system based on natural language parsing. Section IV presents the study of the emoticon's uses and statistical distributions in social media to give the reader a general view of this phenomenon. Section V focuses on the precision study of emoticon, especially in the context of brand-oriented sentiment. Section VI is the recall study. We summarize our findings and discuss future work in Section VII.

II. Related Work

2.1. Sentiment analysis

Sentiment analysis has been a focus of research in the last decade, especially when associated with big data of social media. Natural language mainly involves two types of expressions, one called subjective language, and the other objective language (Bruder & Janyce 1990; Yu and Vasileios 2003). The former is used for stating facts or evidence while the latter is to express sentiments or judgments. Traditional information extraction systems target facts such as relationships and events (Chinchor and Marsh 1998). More recent research on extraction of sentiments from subjective language has drawn special interests from industry as the underlying technology opens the door to the untapped unstructured text world in social media big data, for gaining business insights and listening to the customer's voice and sentiments towards a brand. Whether the sentiments represent customers' happiness with a product or

complaints, they are invaluable customer insights which the businesses so far mainly collect from surveys manually.

In the research literature, the main stream of sentiment analysis has been sentiment classification based on machine learning algorithms (Pang and Lee 2008). Such keyword-based learning supports course-grained sentiment classification, basically tagging the incoming document or post as positive, negative or neutral (Pang, Lee and Vaithyanathan 2002; Turney 2002; Taboada et al 2011). This approach for coarse-grained sentiment analysis has the benefits of quick implementation and high performance in a narrow domain such as movie or product reviews when labeled corpus is available for training (Titov and McDonald 2008; Pang and Lee 2005). Within a domain, the vocabulary for subjective language is fairly limited and unambiguous, the sentiments are more black and white, hence, the sentiment classification based on keyword density using even a simple learning algorithm can achieve high precision (80% or above), as long as the input is not a short message. In websites like Amazon and Yelp, there is no lack of online reviews data which are already labeled by the users using either thumbs up and down icons or 5-star ratings. The increasing availability of such labeled data creates an ideal condition for using simple machine learning approach to sentiment classification.

However, all coarse grained statistical sentiment classification faces a number of challenges, listed below.

The first is domain portability challenge. Traditional text classification approaches are domain dependent (Read 2005; Turney 2002). A sentiment classifier trained on movie reviews performs poorly on electronics reviews. In a commercial sentiment analysis system, social data are harvested from almost everywhere on the Web, including Twitter, Facebook and various review and blog sites and different domains, such as hotels, airlines, retailers, banking, automobile, foods, TV shows, and every other possible sector. With so many domains and verticals where a market researcher wants to apply sentiment analysis, training and maintaining domain-based classifiers seem to be a daunting task, even if the data from all these domains are available.

The second challenge comes from the length of a message. As mentioned before, the learning-supported classification is good at classifying a document or a long review post, but the quality will compromise significantly once the incoming message is short (e.g. a tweet). This is understandable as short messages do not provide sufficient number of data points (lexical evidence for sentiments) for an effective traditional classification system to work on. There is some research on applying classification to short messages in an attempt to address this challenge (Khan, Baharudin and Khan 2010; Yu and Hatzivassiloglou 2003; McDonald et al. 2007; Titov and McDonald 2008; Wilson, Wiebe and Hoffmann 2009), but most such study is either highly research oriented or too domain dependent, and hence cannot be applied to a real world sentiment analysis system. As the mobile platform for social media is getting popular, the social media world will soon be dominated by short messages. In fact, in our launched customer insight product, short tweets are already dominating, taking more than half of the entire social space while all other millions of social media sources put together cannot match.

The third challenge comes from the unsatisfactory classification precision in the practical scenarios when big data becomes small. For example, almost all market researchers need

the support for slicing and dicing the data for in-depth sentiment analysis from different angles, based on demographics, geo-locations, time, etc. In particular, representing sentiment insights on the dimension of time shows the trends of a brand and its history, which is particularly useful for monitoring a brand's ups and downs in real time. However, when sliced and diced based on the users' needs, big data becomes small quickly, and a precision-challenged classifier is bound to have trouble with the users.

The fourth is the association challenge. Sentiments are not meaningful unless they are associated with an object, such as a brand or product, a business, a person or any other hot topic. By nature, all classification based sentiment analysis has the trouble in associating with the target object (Davidov， Tsur and Rappoport 2010). Most such systems rely on co-occurrence and proximity heuristics for topic-sentiment association. Since there is no structure or any relational understanding of the message in these systems, they are powerless with simple comparative expressions like "Google is a lot better than Yahoo" or "I prefer iPhone over Blackberry", and thus the topic often loses connection with its sentiment content.

Finally, insights from sentiment classification only provide an overview of the sentiments, they are not actionable insights. Deep insights for sentiment analysis need to uncover the reasons behind the sentiments to answer business questions such as why customers like or dislike a product. For instance, "I like iPad for its apps", where the reason for the liking of iPad is "its apps". We want to extract not only iPad as Object but also "its apps" as the reason of the emotion. Decoding the underlying reasons is a critical need for actionable insights because business can take actions with such insights as a guide.

The implementation and deployment of our sentiment analysis engine in the customer insight system is designed to address the above challenges, taking the approach of deep sentiment extraction based on natural language parsers. Our system is multi-lingual (currently, English, Chinese, Japanese, Spanish, German, Portuguese and Italian), but for this paper, we limit our study to English.

2.2. Emoticons study

The popularity of emoticons (or smileys) comes hand in hand with the growth of social network. They have been extremely popular in social media among the younger generation and the seasoned netizens. Despite their use everywhere in the online text, linguistically, emoticons are not a "legitimate" part of natural language vocabulary or morphology, hence belonging to so-called Unnatural Language Processing (UNLP, Ptaszynski et al. 2011). Some emoticons are fairly universal as symbols of emotion, some are language dependent. Nevertheless, a serious Natural Language Processing (NLP) system, especially the sentiment analysis system, should not simply ignore them as they contain sentiment information by the nature of their semantics. Survey shows that they are the second most important vehicles for expressing emotions in online communication (Ptaszynski et al. 2011).

In the context of NLP, the use of emoticons and hash-tagged emotion words (e.g. #anger) has attracted machine learning researchers in sentiment classification. Emoticons seem to be a handy and reliable indicators of emotions and hence are used either to help automatically generate a training corpus for sentiment

classification (e.g. Read 2005) or to act as seeds or one type of evidence features to enhance sentiment classification (Davidov, Tsur and Rappoport 2010; Liu, Li and Guo 2012; Read 2005; Yang, Lin and Chen 2007; Hogenboom1 et al. 2013).

Asian netters seem to be even more creative and engaged in making and using a variety of emoticons. In addition to the universal emoticons started from the west, more and ever-growing Eastern types of emoticons have been created. Accordingly, there is considerable study of such phenomena in the context of Asian text (Ptaszynski et al. 2010 & 2011; Zhao et al. 2012).

Not much has been done in evaluating the contributions of emoticons in sizable real life social media corpora, in the context of brand-centered sentiment analysis. That is one major motivation and value for this study.

Ptaszynski et al. (2011) proposes that emoticon research consists of four lines of tasks: (1) detection; (2) extraction; (3) parsing; (4) semantic analysis; (5) generation; (6) evaluation. Our work involves (1), (2) and (6). The work involved in (3), (4) and (5) assumes the productive nature of emoticons, similar to the open morphology study in natural languages. For the following reason, at least for English, this is not a real issue.

There are thousands of varieties of emoticons due to the semantic compositionality of its components, in ways that are very close to flexible word formation in natural language morphology: a smiley face is typically made with various types of eyes, nose and mouth etc. (Strapparava and Mihalcea 2008). However, we observe that the frequency distribution is very different, and many theoretically possible combinations do not really add to the system due to the infrequency of their appearance in data. At least for English, the top n (n<1500) emoticons are easily listable in lexicon and can fulfill the identification of emoticons with very high precision and recall. The research on (3), (4) and (5) would be more meaningful in Asian context for future work.

III. Extraction of deep sentiments from social media

3.1 System Overview

The entire system involves the conceptual design of two-subsystems and four levels of processing. The two subsystems are often referred to as front end search app and backend indexing engine. The four levels are:

1. linguistic level
2. extraction level
3. mining level
4. app level

These four levels of two (sub-)systems basically represent a bottom-up support relationship: 1 ==> 2 ==> 3 ==> 4. Clearly, the core engine of NLP (namely, a parser) sits in the first layer as an enabling technology, and in the app level in the fourth layer lies our customer insight application.

The backend system is composed of three components: data acquisition involving Extraction, Transformation and Loading (ETL), NoSQL databases, and distributed indexing. Pooled from different sources in different formats in real time or non-real-time, the data is uploaded by the ETL subsystem into the distributed Cassandra NoSQL database. The uploading process involves language detection, spam detection, web page extraction and de-duplication. The data integration goes through distributed

processing after the data is loaded in the NoSQL database, using the open source text indexing engine Apache Lucene (lucene.apache.org) based on the MapReduce Framework. The NoSQL databases and the distributed indexing engines are configured in the Computing cloud in order to ensure the elasticity of the entire backend system. A large social media data archive accumulated since one year ago (about 30 billion documents across 40 languages) can be indexed within seven days by using about 30 Cassandra database servers and 150 indexing servers. The frontend system is a SaaS-based application very similar to a search engine. Users interact with the apps through the browser and the application server. The application servers and the query node cluster communicate via multicast based on JGroups. The users' queries are broadcasted to the query node cluster for distributed search. The search results go through a process of integration, ranking and necessary transformation in the application server before they are presented to users using our configurable dashboard in the app. One year archive of social media big data generates about 25 terabytes of index files stored evenly in about 170 query servers using SSDs (solid state drives). In most cases, the user can see their search results within 2-3 seconds for all the required data reports in the app platform.

At a high level, the NLP core engine reads sentences and extracts sentiment insights to support our product for market research. Our focus is on the workings of the NLP core engine that parses and indexes social media text in scale.

The NLP indexing engine is a two-component system. It forms a highly modular processing pipeline, from shallow levels of linguistic processing to deep levels of sentiment extraction, using an expanded version of the seasoned NLP-oriented formalism named Finite State Automata (FSA, a high-level formal language that supports the encoding of finite state grammars for rule based NLP, Roche and Schabes 1997) . Similar formalism and platform for NLP and information extraction include Silberztein (1999), Hobbs (1993) and Srihari et al. (2006).

The first component is a dependency parser based on basic phrase structures generated by chunking. The parser outputs a dependency-based hybrid tree structure involving basic phrases (Li 1989), representing the system's understanding of each incoming sentence. The hybrid tree is a system-internal, linguistic representation, much like diagramming taught in grammar school. The system parses text in a number of passes (modules), starting from a shallow level and moving on to a deep level. The tree output provides a logic-semantic basis for the next level of extraction modules for more generalized coverage of the sentiment phenomena.

The second component is an extractor, sitting on top of the parser and outputs a table that directly meets the needs of products. This is where extraction rules, based on sub-tree matching, set to apply, including our deep sentiment extraction for social media customer insights. Considering the combination possibilities of surface structures, extraction rules built at logical level, on top of deep parsing tree structures, are often as powerful as hundreds of, or even thousands of, low-level linguistic rules in terms of covering the relevant surface patterns of language expressions for sentiments.

3.2. Dependency Tree Structure and Frames

An insight extractor is defined by extraction frames. A frame is a table or template that defines the name of each column (called event role) for the target information (or insights). The

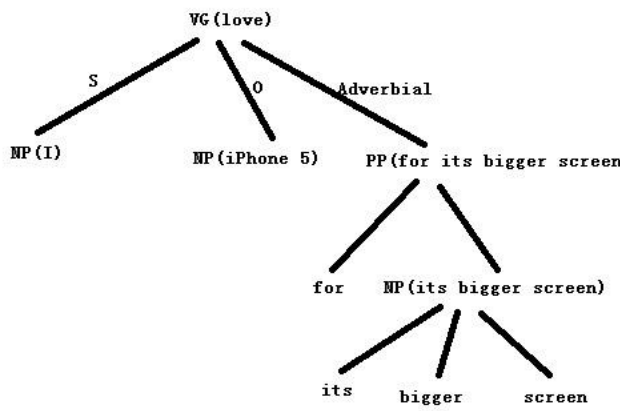
purpose of the extraction component is to fill in the blanks of the frame and use such extracted information to support a product.

The frames for objective events define things like *who did what when and where* etc. with a specific domain or use scenario in mind. The frames for sentiments or subjective evaluations contain information first to determine whether a comment is positive or negative (or neutral, in a process traditionally called sentiment classification). It also defines additional, more detailed columns on *who made the comment on what to what degree* (passion intensity) *in which aspects* (details) and *why*. It distinguishes an insight that is objective (for example, "cost-effective" or "expensive") from subjective insight (for example, "terrific", "ugly" or "awful").

Sentiment extraction is based on parsing by sub-tree matching in extraction grammars. For example:

I love iPhone 5 for its bigger screen.

The parser first decodes the linguistic tree structure, determining that the subject is "I", the predicate verb is "love", the object is "iPhone 5"and the prepositional phrase (PP) "for its bigger screen" is an adverbial. The system extracts these types of phrases to fill in the linguistic tree structure as follows.



Based on the above linguistic analysis, the second component extracts a sentiment frame as shown below:

<Sentiment Frame>

<Type="POSITIVE">
<Agent ="I">
<Object ="iPhone 5">
<Emotion="love">
<Reason="for its bigger screen">

3.3. Concept of deep sentiments

The concept of deep, fine-grained sentiments is proposed in contrast to the dominant practice of shallow, course-grained sentiment analysis, thumbs-up and down (or plus neutral) classification, coupled with sentiment association based on proximity. This concept is inspired by the needs from the real world market analysts told us they need more actionable insights and hope we can answer the why questions with regards to sentiments. To shed some light in the definition, a deep sentiment system should be able to extract insights that can answer these questions:

- About which brand is this sentiment about? (association insight)
- Who made the sentiment comment? (customer background insight)
- How intense is the sentiment? (passion intensity insight)
- Can the system associate sentiments not only with a brand such as iPhone, but also with a feature of the brand, say, screen? (granularity insight)
- In addition to sentiments related to emotions about a customer (love/hate/happy/annoy etc.), can system identify the agent's positive or negative

evaluations of the brand (e.g. cost-effective/poorly-designed/high-definition)? (customer evaluation insight)

- How about the agents' needs or wish list for brands? (market needs insight)
- How about agents' positive or negative action towards a brand (including consumers' purchase intent such as "will buy"; negative actions such as "abandon", "stop using")? (customer action insights)
- What are pros of a brand, including specs, features, functionality (designed to do what)? (pros insights)
- What are the cons of a brand, including weakness, loopholes and issues? (cons insight)
- Can system identify comparisons between brands (iPhone is better than Blackberry)? (competition insight)
- Finally, most important of all, what is the reason of the sentiment? (why insight)

Systems that can answer such questions provide invaluable actionable insights to businesses. For example, it is much more insightful to know that consumers love the speed of iPhone 4s but are very annoyed by the lack of support to flash. This is an actionable insight, one that a company could use to redirect resources to address issues or drive a product's development.

A sentiment analysis system that can answer most or all of the above questions is what we call the deep sentiment system. Apparently, some of the insights listed above, especially the actionable insight why, are beyond what can be attempted using the existing keyword-based models because it requires considerable natural language understanding to decode such elaborate and fine-grained insights. A deep parser is called for to facilitate the deep sentiment extraction in order to meet the business users' real world needs.

Since our sentiment extraction must be object centered, meaning that the sentiment extracted must be about a topic or an object, which could be a brand, a person name, etc. We typically select 10-20 brands to evaluate the system. For instance, in our most recent release the brands we use are:

iPhone, Walmart, Listerine, Costco, Olive Garden, Taco Bell, Tylenol, Camaro, Prius, Ikea, JetBlue, Skype, Yoplait, Playstation, fish oil, Pepsi.

We use CrowdFlower's anonymous annotation service to annotate the testing data, and at least 75% inter-annotator agreement of at least four judges is used for gold standard. The overall precision for the extracted sentiments is 87% on average across brands. Due to the fact that sentiment is generally sparse in randomly selected data, we measure the total number of extracted sentiment mentions and calculate its percentage given a certain amount data processed by the system. We call this sentiment coverage metric Relative Recall. The relative recall is 8.5% across brands. Although this is not absolute recall, it is good for comparing the recall performance between different product releases, and thus this metric serves our purpose well.

IV. Emoticon Frequency and Distribution

This section investigates the general statistics of emoticons as used in brand-oriented social media data. It needs to be noted that in our approach to brand-focused, post-level sentiment

analysis, the sentiment analysis must be targeted at a specific object (usually a brand).

Identifying sentiments without an object is of little value in sentiment analysis of customers' insights. It is with the brand focus that we evaluate how emoticon's use as a clue will impact the precision and recall of our customer insight system. To the best of our knowledge, most other machine learning approaches to sentiment classification, either at document (paragraph) or sentence level, will not take the object into consideration in their algorithm and evaluation of sentiment classification (at a later stage some use heuristics such as co-occurrence to associate sentiment with objects). That makes the research and scoring much simpler, but will not simulate how our users experience and evaluate the results in the real world.

Identification of emoticons and labeling them as positive or negative (and sometimes neutral if needed) are a precondition for investigating and making use of emoticons in the sentiment analysis. An approach of combining emoticon lexicon and emoticon pattern rules is taken to do this.

Our emoticon lexicon houses over 1000 emoticon symbols, which are manually collected and reviewed. Each entry is marked as positive, negative or neutral by hand. Fortunately, most of the western emoticon symbols in the lexicon are not ambiguous and the identification of them from text is a simple matter of lexical look-up.

A number of specific pattern rules are also implemented to disambiguate a handful of ambiguous trouble makers based on pattern matching. The precision and recall of emoticon identification through the combination of emoticon lexicon lookup plus emoticon pattern matching yield almost perfect precision and recall for identification after several rounds of tune-up and debugging.

A representative social media corpus is collected to help analyze emoticon's various frequencies and distribution. This corpus is made up of randomly collected data, altogether 440,000 social media posts, including tweets, Facebook posts and forum posts. For our purpose of brand-oriented sentiment analysis, the only requirement for the collection is that each post selected must contain at least one term in the represented brand set = {Pepsi, Walmart, Groupon, Taco Bell, Amazon}. For each brand, there are roughly 88,000 posts.

Table I lists the emoticon richness distribution. The emoticon richness, calculated as 3.27%, is defined as the ratio of the number of posts containing at least one emoticon to the total number of posts in the corpus. This metric informs us of the overall frequency of emoticons as used in social media and hence the suggestion of their maximum possible impact on a corpus.

TABLE I. EMOTICON RICHNESS AND DISTRIBUTION

Brand	Total Posts	Posts with emoticon	Percent
Pepsi	84,483	5,006	5.93%
Taco Bell	85,790	3,727	4.34%
Walmart	94,339	3,067	3.25%
Groupon	82,205	900	1.09%
Amazon	92,223	1,653	1.79%
Total	439,040	14,353	3.27%

Given the over 1000 western emoticons in our emoticon lexicon, we are also curious about how these emoticons are actually used in social media. Table II displays the distribution of

TABLE II. RATIO BETWEEN POSITIVE AND NEGATIVE EMOTICONS

Brand	Posts w Emo	Positive Emoticon		Negative Emoticon		Pos & Neg Mix	
Pepsi	4,549	3,058	67.2%	1,469	32.3%	22	0.5%
Taco Bell	3,112	2,141	68.8%	959	30..8%	12	0.4%
Walmart	2,661	1,764	66.3%	887	33.3%	10	0.4%
Groupon	759	619	81.6%	138	18.2%	2	0.3%
Amazon	1,491	1,210	81.2%	275	18.4%	6	0.4%
Total	12,572	8,792	69.9%	3,728	29.7%	52	0.4%

Positive (Pos) emoticons and Negative (Neg) emoticons (excluding the 1,781 neutral emoticons as they are not involved in sentiment analysis). It is shown that the positive emoticons are used much more frequently (about 70%) than the negative emoticons (about 30%). This does not necessarily imply that social media involve more positive comments than complaints. One explanation could be that people tend to use the simple popular positive emoticons such as :) very heavily (Table III), not necessarily directed towards a specific object, but more to make the post light, friendly or humorous. The default of emoticon use is found to be a positive smiley face instead of a sad face or a neutral one.

Given the large number of various expressions of emoticons, which emoticons are most popular in social media? The top 10 most frequently used emoticons in the corpus are listed in Table III. It demonstrates that although the number of emoticons is big, only a handful of emoticons are used far more frequently than all others. Usage of emoticons is skewed toward certain simple emoticons such as :) or :(. In addition, the positive emoticons are dominating,

TABLE III. TOP 10 EMOTICONS

Polarity	Emoticon	Count	Percent
P: happy	:)	4,833	33.67%
P: love	<3	1,670	11.64%
N: sad	:(	1,406	9.80%
P: laugh	:D	1,270	8.85%
P: happy	(:	1,174	8.18%
P: wink	;)	1,160	8.08%
P: happy	:-)	727	5.07%
P: kidding	:P	719	5.01%
P: grin	XD	554	3.86%
P: happy	=)	339	2.36%
	Total	9,019	62.84%

with only one negative emoticon making into the top 10 list.

As this study is based on brand-oriented data, the results are a bit different from, but still largely consistent with, a similar study reported

by others. That study on emoticon frequency is based on much bigger data sets from Twitter (see http://datagenetics.com/blog/october52012/index.html). They report, "he popular emoticons dominate the usage patterns. Of the 96,269,892 Tweets that contained emoticons, the top 20 smileys accounted for 90% of all occurrences. " This indicates that a lexicon approach to emoticon identification, by just listing the top n emoticons without using rules, can also be a viable approach with only a small loss in recall.

Our next question is, how do people actually use these emoticons in online text? One way to get an insight of this is to count the unique emoticons used by a particular individual. In addition to text body, other meta information of the text, including user id, is also available in our content store. Thanks to this, we are able to find out the fact regarding users' usage of unique emoticons. As seen in Table IV, despite all the varieties of emotions in the web, majority of users (95+%) only use a couple of unique forms of emotions.

TABLE IV. USER COUNTS OF UNIQUE EMOTICON

Unique Emoticons	Number of Users	Percentile
1	238000	95.02%
2	6325	2.53%
3	4350	1.74%
4	1265	0.55%
5	400	0.16%

V. Emoticon Precision Study

Precision and recall are key measures that are used to benchmark the quality of a sentiment analysis system. This section studies how the introduction of emoticons can help enhance the precision in brand-oriented sentiment analysis. Based on the community standards, precision used in this study is defined as: correct / (correct + wrong + spurious).

Human annotation is necessary to support in-depth sentiment analysis study. From the random corpus of 440,000 posts as used above, 2,000 posts are randomly selected for human annotation, based on the 14,353 posts that carry positive or negative emoticons only (i.e. excluding posts of neutral emoticons and the few positive-negative mixed posts). Each post is annotated by two annotators; and their agreement is used as the gold standard. Each post is annotated in two ways: (i) the general tone of the post as positive, negative or neutral, irrespective of an object; (ii) a sentiment choice of positive, negative, or neutral towards the corresponding brand associated with the sentiment. Of course, our focus is on the gold standard as defined in (ii), but it is insightful to perform a comparison of benchmarks between (i) and (ii), presented below in Table V and Table VI respectively.

TABLE V. EMOTICON PRECISION IRRESPECTIVE OF BRAND

Human judge (I)	Positive Emoticon	Negative Emoticon
Positive	1230 (correct)	10 (wrong)
Neutral	384 (spurious)	38 (spurious)
Negative	16 (wrong)	322 (correct)
	Precision = 77.6% (1230+322)/(1230+322+384+38+10+16)	

Table V shows 77.6% as the overall precision for sentiment irrespective of brand. 77.6% is decent for precision, showing that

emoticons are indeed important clues for determining the sentiment tone of a post. But 77.6% is apparently not good enough to be used to collect data as gold standard for training a sentiment analysis system as earlier researchers did before.

Emoticon expresses emotions of the poster, but that does not necessarily mean that there is a positive emotion for a brand mentioned in the post. For instance, in the post "*Someone asking a greeter at walmart to watch their child ---- JOKE :) ha ha ha ha*", even though there is a positive emoticon indicating the poster is happy, there is no indication that the sentiment is towards the brand "Walmart". This shows that the sentiment as representing a general tone of a post may be different from the sentiment as associated with a specific brand. Only the latter is the brand-oriented sentiment analysis required by the businesses in understanding customer insights, as shown in Table VI.

Table V and Table VI show that there is a sharp drop, for almost 50 percentage points (77.6 – 28.5), in precision from the object-free sentiment analysis to the desired brand-oriented sentiment analysis. A precision of less than 30% is really too low for any uses. This means that for brand-oriented sentiments, emoticons alone are too weak to be useful. The future research should explore combining emoticon evidence with other evidence in enhancing the sentiment analysis.

TABLE VI. EMOTICON PRECISION OF BRAND-ORIENTED SENTIMENT

Human judge (II)	Positive Emoticon	Negative Emoticon
Positive	458 (correct)	76 (wrong)
Neutral	1104 (spurious)	186 (spurious)
Negative	64 (wrong)	112 (correct)
	Precision = 28.5% (458+112)/(458+112+1104+186+76+64)	

Note that there is a large number (1104+186) of spurious cases where the human judges tag as neutral. As discussed before, positive emoticons are default ways of making the posts light, it does not have to carry strong sentiment towards a brand, especially for the simple common smiley faces like :) or :D. Hence, there are lots of gray area cases which are neutral but associated with positive emoticons, for example,

> *@tonymophoto It was American Balloons in Land O Lakes. I bought a Groupon for it. :).*
>
> *Did you know there are Amazon links to all the books we discuss at http://t.co/rQZ8Uo0E? Buying thru them supports the show! :D.*

As for spurious cases of negative emoticon, it seems that some simple negative emoticons could mean a weak "sorry" or "that's all I can do", while the post is still fairly neutral, e.g.

> *to be honest i can't really taste a difference between Coke and Pepsi ;-;.*
>
> *I need to go to Walmart to buy my letters and my mirrors. :(.*

The last example is a need statement which demonstrates some peculiar properties in terms of sentiment analysis. A need statement often goes hand-in-hand with negative emoticons (i.e. sad that there is a need not met) while the need statement is neutral from post sentiment perspective and it is often regarded as positive from brand perspective (because a needed product is a positive mention: e.g. I badly need a new iPhone).

As expected, wrong cases are very few (1%-3%) when the emoticon polarity mismatches

human key (i.e. a positive emoticon is used with a negative post or vice versa). It can be considered as random noise, caused by mistyping, and in some cases, it might involve a degree of sarcasm.

Absolutely pissed that Taco Bell isn't open.:-).

RT @Jyoti_More: @rehannaaaaaa It's like Taco Bell but it seems so much better ;(.

RT @AfsahB: When drinking Pepsi/Coke out of a thela was so cool :'(. #PuranaPakistan (this looks like sarcasm)

Table VII benchmarks the precision of the Netbase sentiment analysis system in this test corpus.

TABLE VII. NETBASE PRECISION OF BRAND-ORIENTED SENTIMENT

Human judge (II)	Positive Emoticon	Negative Emoticon
Positive	580 (correct)	24 (wrong)
Neutral	86 (spurious)	26 (spurious)
Negative	52 (wrong)	168 (correct)
	Precision = 79.9% (580+168)/(580+168+86+26+52+24)	

Remember that the Crowdsource benchmarks of Netbase system is 87% (Section 3.3) which is greater than 79.9% on this relatively small corpus. Eight percentage points are statistically meaningful, so a reasonable explanation is called for. As the testing corpora are all collected using brands as trigger words from the same social media sources, the only major difference is that posts in this study all involve emotion mentions, the drop of precision seems to indicate that the Netbase system's data quality on emoticon-involved data is not as good as the quality on random data. This is fairly understandable as posts involving emoticons tend to be more casual and degraded, involving more social media jargons and ungrammatical fragments. Therefore, adequately making use of the emoticon evidence in sentiment analysis in such data is more important.

There is another more significant finding in interpreting the relevant data and analyses. Although emoticons alone may not be reliable evidence for sentiment analysis and it requires more research on balancing them with other evidence, emoticons seem to be very good indicators as **counter-evidence** to block incorrect sentiment classification. This is especially meaningful as all incorrect classification in tagging positive posts as negative, or the other way around, involves embarrassing glaring errors. This is unlike the distinction between neutral and strong sentiments (whether positive or negative) where the mismatching between the system's tagging and the human judgment is not a big issue as there is a possible gray area involved. But the sentiment polarity error is a fatal mistake made in black and white area where human judges have no problems but a system often has trouble in sentiment identification (e.g. in tricky cases such as double negation). Extra-linguistic evidence such as emoticons can save these critical cases with high confidence. The implementation along this line is straightforward: no sentiment classification is allowed to be in conflict with the polarity of the emoticons. Based on the careful data analysis of 140 cases (i.e. all the wrong cases in Table VI), three restrictions apply as exceptions to the above rule: (i) the default positive emoticon :) should not be involved as a sentiment tagging blocker as it tends to be over-used; (ii) the negative emoticon should not be a blocker in a need statement the Netbase system can identify need statements), such as I badly need a new iPhone :(where the

system tags iPhone as a positive mention correctly; (iii) an emoticon should not block sentiment tagging in a preference statement (e.g. rather A than B; again, identifying the preference statements is within the Netbase system's tagset), a mixed case (e.g. love A but hate B), or a long post (threshold can be set at 10-15 words in length). These restrictions are not difficult to enforce in the Netbase system as many of the capabilities have already been built-in. As a result, some eye-catching and otherwise difficult-to-catch precision errors are avoided when the above heuristic is in effect.

VI. Emoticon Recall Study

We would like to get a sense of the recall improvement if we add emoticon as a sentiment indicator. The richness of emoticons listed in Table I is summarized 3.27%. This tells us that the maximum contribution to the system recall would be 3.27%, assuming: i) none of the sentences containing an emoticon have been correctly classified by the system; ii) emoticon's precision as a sentiment indicator is 100%. In reality, neither of these two assumptions hold true, of course. As a result, the actual recall contribution would be lower than these numbers. In order to get a precise idea of the recall impact to our existing system, we use an existing Crowd-source-annotated corpus used by our QA department for evaluation. The results are presented in Table VIII. The first column shows emoticons' agreement with the annotation, Column 2 is for the emoticons' agreement with annotation missed by the system, Column 3 is the recall improvement and Column 4 shows the overall recall improvement emoticons can possibly make.

TABLE VIII. EMOTICON'S IMPACT ON SYSTEM RECALL

	agree w key	system miss	Recall up	Overall up
pos	1052	20	1.94%	2.72%
neg	88	3	3.52%	

So how to assess the 2.72% recall improvement? The number does not seem to be impressive and significant from a research perspective. However, from the practical system development's point of view, this improvement is useful and meaningful. In our English rule system for sentiment extraction, we have a total of 409 rules, which consist of thousands of linguistic patterns built upon a semantic parser. However, the top ten mostly fired rules contribute to nearly 50% of all sentiments extracted. The vast majority of all other rules account for the long tail of the remaining 50% sentiments. There is a very long tail of individual rules that contribute to less than 1% of sentiments, but without them the errors are very eye-catching once they do occur. Table IX lists the top ten fired rules in our system. In fact, it is these

TABLE IX. TOP RULES CONTRIBUTING TO SENTIMENT EXTRACTION

Firing Frequency	Positive	Negative
# 1	10.16%	9.89%
#2	9.59%	7.12%
#3	5.15%	6.11%
#4	4.96%	5.22%
#5	3.49%	3.93%
#6	3.95%	3.42%
#7	3.60%	3.11%
#8	3.13%	2.73%
#9	2.88%	2.50%
10#	2.85%	2.25%
Sum	49.76%	46.37%

corner cases that make a difference between a high-precision system and a mediocre system as the easy ones can be fairly quickly captured by all systems using whatever approaches. Individual rules that contribute to modest recall but can correct glaring errors cannot be ignored in a real life system. In this sense, the emoticon provides a low-hanging fruit for enhancing the data quality which should not be ignored either.

Considering the fact that most rules contribute to the system recall marginally, the 2.72% contribution from emoticon is fairly significant. We would also like to point out that the recall contribution also varies depending on the data source and the maturity of the sentiment system. Since our English sentiment system is the most developed of all languages, it leaves less room for the emoticons to enhance the net recall. But in sources with heavy use of emoticons such as school students' Facebook updates, the impact would be much greater.

VII. Conclusion and Future Work

We have presented a comprehensive data analysis of the role of emoticons in sentiment classification, focusing on its impact in the context of consumer's sentiment for a brand because that is what the field really needs from the business clients. The study shows that emoticons alone without considering other linguistic evidence are not sufficient to dictate a sentiment towards an object. On the recall side, emoticons' potential overall contribution is limited but fairly meaningful: the upper boundary for absolute recall contribution is 4.2%. A reasonable 2%-3% recall enhancement can be expected if balanced properly with other evidence.

For future work, we plan to investigate various ways of using emoticons with other types of evidence to maximize the enhancement of the sentiment classification. For example, in the following sample sentences,

Skype messed up us talking lol smh ;(.

iPhone is coming to sprint, my dad asked at the store today :-D."

I have my Taco Bell :-) yum.

"Skype", "iPhone" and "Taco Bell" are the objects we would like to extract sentiment for. We can see that there are no any emotional words or phrases in the sentences. In the first case, the combination of negative emoticon ;(and the verb "mess up" indicates that the speaker is unhappy. In the second and third cases, there are not any emotional triggers linguistically, except for the emoticons. However, the fact that the speaker is waiting for something to come in the second case and the speaker gets something in the third case, coupled with an emoticon in each case, indicates exciting, happy emotion.

How do we linguistically understand and make sense of these facts from the sentences? These sample sentences show that emoticons mainly help extract implicit sentiment from factual sentences, not subjective sentiment. This is important because a traditional sentiment system is generally good at analyzing data with explicit sentiment triggers, such as words like "hate", "love", and "terrific", etc. Without emoticons, it is hard to judge whether there is an emotion or not in cases without explicit triggers.

In a nutshell, our NLP system provides these resources listed below, which we can utilize to balance with emoticons as evidence for sentiment classification:

- An ontology-based rich feature system
- A rich set of lexical resources with different types

- A thorough NLP analysis of each token in the sentence
- A semantic analysis and representation of the sentence by a deep parser
- An efficient finite state automata based rule engine which identifies, among others, a rich tagset such as statements for benefit, problem, need and preference

All these facilities working together allow us to make use of rich set of evidence including emoticons, and write either generic or fine grained rules that can effectively and accurately extract sentiment based on balancing different types of evidence. In particular, an effective use of emoticon as counter-evidence to fix the glaring errors such as sentiment polarity upside-down mistakes helps enhance the users' experience and trust in the system quality when the emoticon evidence is coupled with the system's existing capabilities. That is the work currently being implemented in our system.

We also plan to involve the hash-tagged sentiment words in twits such as #sad and #excited as they seem to demonstrate the similar role as emoticons in the popular twitter community. The detailed study of language dependent part and the universal part of emoticons, especially the Eastern vs. Western distinction, is also interesting and beneficial to our multilingual program. Finally, the study of the language-dependent part of emoticons, especially for the Eastern vs. Western distinction, is also interesting and would be beneficial to the multilingual program.

References

[1] N. Chinchor and E. Marsh 1998. MUC-7 information extraction task definition (version 5.1). Proceedings of MUC-7.

[2] D. Davidov, O. Tsur and A. Rappoport 2010, "Enhanced sentiment learning using twitter hashtags and smileys," Proceedings of COLING 2010: Poster Volume, 241–249, Beijing, August 2010.

[3] J. R. Hobbs 1993. FASTUS: A system for extracting information from text. Proceedings of the DARPA workshop on Human Language Technology, pp. 133–137. Princeton, NJ.

[4] A. Hogenboom1, D. Bal1, F. Frasincar1, M. Bal1, F. Jong, U. Kaymak 2013. "Exploiting Emoticons in Sentiment Analysis", Proceedings of the 28th Annual ACM Symposium on Applied Computing: 703-710

[5] A. Khan, B. Baharudin and K. Khan 2010. "Sentence based sentiment classification from online customer reviews", Proceedings of the 8th International Conference on Frontiers of Information Technology

[6] K. Liu, W. Li and M. Guo 2012. "Emoticon Smoothed Language Models for Twitter Sentiment Analysis". In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012

[7] R. McDonald, K. Hannan, T. Neylon, M. Wells and J. Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In ACL 2007.

[8] B. Pang and L. Lee 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1): 1-135, 2008

[9] M. Min, T. Lee and R. Hsu 2013. Role of Emoticons in Sentence-level Sentiment Classification. Proceedings of First International Symposium on Natural Language Processing based on Naturally Annotated Big Data (NLP-NABD) 2013, Suzhou, October 2013. Lecture Notes in Artificial Intelligence (LNAI), Published by Springer ISSN 0302-9743

[10] B. Pang, and L. Lee 2005, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," Proceedings of ACL 2005. Ann Arbor, MI, USA. 2005:115-124.

[11] B. Pang, L. Lee and S. Vaithyanathan 2002 Thumbs up? sentiment classification using machine learning techniques. In EMNLP 2002, 79–86.

[12] M. Ptaszynski, R. Rzepka, K. Araki and Y. Momouchi 2011. "Research on Emoticons: Review of the Field and Proposal of Research Framework," 言語処理学会 第 17 回年次大会 発表論文集 (2011 年 3 月)

[13] M. Ptaszynski, J. Maciejewski, P. Dybala, R. Rzepka and K. Araki 2010. "CAO: a fully automatic emoticon analysis system based on theory of kinesics", In IEEE

Transactions on Affaective Computing. Vol. 1, No. 1: 46-59

[14] J. Read 2005. "Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification ," Proceedings of the ACL Student Research Workshop: 43–48

[15] E. Roche and Y.Schabes 1997. Finite-State Language Processing. MIT Press.

[16] M. Silberztein 1999. INTEX: a Finite State Transducer toolbox. Theoretical Computer Science 231(1). Elsevier Science.

[17] R. Srihari, W. Li, T.Cornell and C. Niu 2006. InfoXtract: A Customizable Intermediate Level Information Extraction Engine. Journal of Natural Language Engineering, 12(4), 1-37

[18] C. Strapparava and R. Mihalcea 2008. Learning to identify emotions in text. In SAC, 2008

[19] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede 2011. "Lexicon-Based Methods for Sentiment Analysis," Computational Linguistics, 2011,37(2): 267-307.

[20] I. Titov and R. McDonald 2008. Modeling online reviews with multi-grain topic models. In WWW, New York, NY, USA. ACM: 111-120

[21] P. D. Turney 2002, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. USA.2002:417-424.

[22] T. Wilson, J. Wiebe and P. Hoffmann 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics, 35(3):399–433.

[23] C. Yang, K. H. Lin and H. Chen 2007. "Building Emoticon Lexicon from AWeblog Corpora", Proceedings of the ACL 2007. Demo and Poster Sessions: 133-136

[24] H. Yu and V. Hatzivassiloglou 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. EMNLP

[25] J. Zhao, L. Dong, J. Wu, K. Xu 2012. "MoodLens: an emoticon-based sentiment analysis system for chinese tweets", KDD 2012: 1528-1531

社交媒体中的粤语情报挖掘
Mining Public Opinions from Social Media in Cantonese

Lei Li, Tanya Lee, Tian Tang,Min Martin
NLP Department
Netbase Solutions, Inc.
Mountain View, USA
lli,tlee,ttang,cmin@netbase.com

***Abstract*—**Sentiment extraction from social media has been an applied research area of Natural Language Processing (NLP) that has drawn considerable attention in both academia and industry. However, the treatment of dialects in an NLP-based sentiment mining system has not yet been extensively studied. This paper explores the topic of dialect processing in NLP as applied to Cantonese in the context of Chinese sentiment system. Preprocessing a dialect using machine translation belongs to the so-called 'restoration' approach in the NLP community. Compared with the approach of natively developing a Cantonese parser from end to end or via re-training, the restoration approach has the major benefit in saving the development time as well as linguistic resources. Benchmarks using crowd-sourcing human judges show that the data quality for sentiment extraction undergoes limited degradation (less than 5%) with this translation-based approach, which can be extended to NLP of most dialects.

***Keywords*-**NLP; Chinese parsing; social media; sentiment extraction; Cantonese; dialect

摘要—以新浪微博、腾讯微信为代表的中文社交媒体在近几年呈爆炸式增长，也包括为数不少的粤语的帖子。本文介绍我们实际应用开发的中文社交媒体舆情挖掘系统，重点描述如何用尽可能简单而有效的前处理，把粤语自动翻译成普通话，然后进行鲁棒性(Robustness)中文分析和舆情抽取。由于机器翻译的不规范，鲁棒性成为中文分析器得以应对从粤语翻译过来的普通话的关键。本文介绍了中文系统鲁棒性开发的实践，使得分析器成功投入了粤语的舆情抽取应用。质量检测表明，简单机器翻译再行中文处理的方法非常有效，对于最终舆情抽取的质量影响小于 5%。我们在粤语处理方面的做法，为在实际应用中对方言的处理闯出了一条路子。

关键词：NLP； 中文自动分析； 社会媒体；舆情挖掘；情感抽取；粤语；鲁棒性；方言

I. 引言

以新浪微博、腾讯微信为代表的中文社交媒体在近几年呈爆炸式增长。社交媒体的影响力越来越大，它们成为普通百姓表达意见的一个重要平台，是公众舆论的窗口。在研发社交媒体舆情挖掘的中文系统过程中，遇到为数不少的粤语的帖子。从机器处理角度来看，方言与行业用语类似，算是一种子语言（sublanguage），可以看作是普通话的变体。其中绝大多数的区别在词汇层，句子结构则大同小异。于是决定采用先把粤语文句自动翻译成普通话的方法，然后再由中文主体系统去做分析和挖掘，这样比较简便可行，系统接口清晰，主体中文系统无需改变即可识得粤语文句。

社交媒体的突出特点之一是充满了不规范的字词和表达法。这就要求研发的系统，必须注重分析系统的鲁棒性（robustness）。我们的工作在鲁棒性上有突破性进展，足以应对极不规范的社交媒体。这样的鲁棒性中文自动分析器刚好可以应对机器翻译译文的不规范。因此，我们的粤语工作基本局限于前处理阶段的词对词的轻装机器翻译，中文核心系统无需改变。这比为粤语单单开发一套分析抽取系统，省工省力，更好维护。

本文第二节综述相关工作。第三节讨论粤语自动翻译的过程、挑战和对策。第四节描述核心中文分析器的鲁棒性。第五节展示粤语情报挖掘示例，讨论系统的测试结果以及本研究对方言处理的意义。

II. 相关研究与讨论

社交媒体舆情挖掘的基础是情感抽取（sentiment extraction）。情感抽取的主流是利用机器学习基于所谓“一袋子词”（bag of word）模型的的情感分类（sentiment classification）。通常的做法非常粗线条，就是把要处理的语言单位（通常是文章 document，或帖子 post）分类为正面（positive）和负面（negative），叫做 thumbs up and down classification (Pang, Lee and Vaithyanathan 2002；Turney 2002)。这种做法流行快捷，在某个特定领域（譬如影评论坛），分类质量可以很高 (Titov and McDonald 2008)。这是因为在一个狭窄的领域里面，评论用语相当固定有限，正面负面的评价用词及其分布密度不同，界限清晰，通过一袋子关键词的不同密度来识别褒贬自然不难。开始的情感分类系统多以情感词典为主（Wiebe 2000；Turney 2002；Riloff 2003）。有情感分类结合词典和上下文规则的计算（Ding 2008；Taboada 2011）。

对汉语句子做浅层情感抽取的工作近年开始引起研究者的关注 (Tsou et al. 2005、朱嫣岚等 2006、李钝等 2008、王素格等 2009、党蕾等 2010、赵妍妍等 2010）。特别值得一提的是 Yang and Hou (2012) 基于规则的工作。他们从人工词典标注出发，利用规则系统，对汉语语句进行自动情感分类，步骤清晰，工作做得比较踏实。该系统设计的输出是一个表示趋向的数值。这个人为的设计带来了操作上的困难，包括人工标注词典的困难和人工标注语句的困难。这两个困难的叠加使得测量系统的准确度以及指导系统的进一步发展，变得不太可靠。系统的测试量很小，与大多数中文情感系统一样，它依然是一个离开实际语料和应用非常遥远的实验室系统。

这些研究的局限在于：（1）情感分析很粗疏，以分类为主；（2）分析主要依赖词典和非常有限的上下文，没有句法分析的支持；（3）离开实用距离还远。我们研发的中文系统在这三方面均有突破，尤其突出的是句法分析的鲁棒性使得处理不规范的社交媒体成为可能。

在这样的背景下看粤语的舆情挖掘，有两条路子。一条是重起炉灶，把粤语当作一门新的语言来从头开发一套分析和抽取系统（native development approach）。另一条是把作为方言的粤语当作国语的一种变体，先把它恢复（或翻译）成国语（restoration approach），然后进行中文的分析和抽取，是两步走的方案。方言处理以外，NLP 输入端处理其他不规范文本中对这两条路子也有过很多尝试，譬如输入文本大小写不分的情况。Niu et al（2004）详细比较了两种方法在大小写不分的文本处理中对于最终信息抽取质量的影响。从抽取质量看，恢复大小写（case restoration）再行分析和抽取的两步走方法，明显好过直接在大小写不做区分的文本上从头训练一套分析抽取系统的方法。在方言处理上，两步走的路线可以大大简化研发的投入，前提是我们需

要做到以下两点。首先是粤语到国语自动翻译应该是轻装前处理系统：容易开发、加强和维护，否则还不如从头开发一套粤语系统。第二，中文处理器必须足够鲁棒，可以接受粤语到国语粗糙的翻译。实践表明我们做到了这两点，开拓了一条成功处理方言的道路，质量检测的结果令人满意。

III. 作为前处理的粤语到国语的自动翻译

研发一个句对句自动翻译系统作为前处理太过繁难，维护也不容易，对于子语言而言，是牛刀宰鸡。于是决定退回到上个世纪 50 年代的词对词第一代机器翻译（word-for-word MT）的思路上来（Hutchins and Somers 1992）。我们不追求完美的翻译，只要翻译过来的普通话大体可读，基本不影响抽取就差不多了。传统的词对词翻译容易实现，但最大问题有两点：一是不能解决词汇歧义问题，二是不能应对源语与目标语的句式差异。由于方言与国语之间的句式差别很小，主要在词汇差异，因此第二个问题不构成挑战。解决第一个问题从而成为这项研究的重点。研究表明，采用词典为基础的有限语境消歧（local context disambiguation），可以解决绝大多数的歧义问题，基本满足后续的分析抽取要求。除此以外，我们还采用一项所谓远距离全局放大（global propagation）的技术来帮助消歧一些特别的难点。

词对词机器翻译的主要工作就是一部源语到目标语的转换大辞典。理想的状况是粤语词汇针对国语的对应物没有歧义，一一对应，或者多对一。

吃醋 ← 呷醋
吃饱 ← 食饱
吃香蕉 ← 呔蕉
各就各位 ← 埋位
不是 ← 唔系/唔喺/吾系/5 喺/吾喺/5 系/吴喺

实际情况自然不是这样，歧义表现在词汇转换中普遍存在的一对多现象。譬如：

食/吃 ← 食
孩子/娃娃/小 ← 仔
亲吻/咀 ← 咀
点/怎么 ← 点
虾/欺负 ← 虾
野/东西 ← 野
很/棒 ← 劲

消歧因此是不可避免的任务。

很多时候，歧义的字词很容易在上下文的字词中消解，并不需要抽象的规则或远距离的条件。对策就是编制一部大词典，不怕信息冗余，由数据制导，让熟悉粤语的学生用它来测试大量数据来不断调整上下文的字词条件。这部词典的粤语词条可以是词（如果没有歧义），也可以大于词（如果上下文的字可以帮助消歧）。换句话说，就是增加字符串的长度达到消歧的效果。目前这部词典大约有一万粤语词条，对应与约 5000 条国语词条，基本涵盖了粤语中常用的词汇及其消歧条件。例示如下（国语在前，粤语随后）：

调羹：瓷庚 瓷羹 //22
孩子：个仔　//1
不就得了：米得咯 咪得啰　//22
别抽烟：米食烟　//2
别跑：米走　//1
来过：黎过　//1
没来过：未黎过　//2
晚一些：遲d　//1
这样：噭 咁 哩样 0甘 o甘　//22211
一丁点：鸡碎 斗零　//11

一定会：实会　//1
一定来：实嚟　//2
不敢：唔敢 吴敢　//11
不是：唔系 唔喺 吾系 5喺 吾喺 5系 吴喺　//2212212
不知：唔知 吴知　//11
不行：唔得 吴得　//11
中国菜：唐餐　//1
买东西：买野 买嘢　//22
人家：人地 人哋　//12
什么东西：乜水　//2
什么消息：咩料　//1 //睇睇咩料先
什麽都：乜都　//1
何必：使咩 使乜 駛乜　//222 //駛乜用電筒
你好：哈佬　//1 //英文HELLO直译
挑食：择食 拣饮择吃 拣饮择食　//122
储钱罐：钱罌 扑满　//11
去死：仆街　//2 //广东话的脏话，意思是摔倒在街上
哪一天：边日　//2
哪个：边个　//21
哪儿知道：点知　//2
哪种：边种　//1
唠叨：吖吱吖咗 噏 长气　//201 //（妳唔好再噏我啦！）
不会吧：不系吖嘛 唔系呀话 唔喺啩 唔系挂 唔会啩 唔系啩 唔係啩　//2222222
不太：5太 唔系几　//12
不大理会：好少理　//2
乱说：亂UP 乱UP　//22 //所以你只好係到亂UP扮正義
乱说话：乱噏嘢 山草药　//22
亲一下：咀一啖　//2
亲近：行得埋　//2
什么样：点咖样　//2
什么设计：咩设计　//2
什麽东西：乜东东　//2
办得到：揽得掂　//2
动不了：唔啷得　//2
动不动就：啷啲就 啷亲就　//22
动作快点：快手啲　//2
不用：五驶 五洗 唔驶 吾驶 唔洗 唔使 五使　//1122221
一些什么：d咩　//2
较好：几劲　//1
吧：吖嘛　//1
好棒：好正 好鬼正　//12
好正常：好正常　//2 // exception rule for 好正
好吃：好食　//1
搪塞：回塘 回搪　//11
不用急：唔使急 吾使急 5使急 无使急 5洗急　//22222
学生那：学生果　//2 // 我做學生果時
看来：睇怕 睇黎　//11 //睇怕你都冇論據喎
没法做：无得做　//2
有些：有D 有d 有嘀 有滴 有啲　//11111
无非：冇非　//1
他的：距的 距既　//11
那里：果度　//1
真的：真噶 真咖 真嘎 真嫁 真架 真驾　//111111
不够：唔够喉 唔够　//22 //唔夠錢買咪做下part-time，唱少d k，買少d 名牌law
少一点：少d 少啲 少嘀 少滴　//2211
的吗：噶咩 嫁咩 ga咩 嘎咩 咖咩 嘅乜 架咩　//1121111
难道不是：乜五系 乜唔係 乜唔系 乜五係　//2222
那就：咁就 甘就 敢就　//111
不讲道理：番鬼 打横嚟 啷箕　//121
不对劲：唔系路　//2
乱来：冇谱 乱咁囃 乱黎　//121
臭男人：麻笠佬 麻甩佬 死八公　//222
这还不算：咁都唔算　//2
死了：钉咗 钉左 拉柴 香左 瓜左 直左　//111111

没有头绪：盲摸摸 无纹路　//22
操你妈：屌娜星 屌你老鼠　//22
是这样的啦：喺感价啦 系咁咖啦　//22
这些东西：呢地野　//2
好久没见到你:好耐没见你　//2
评头论足：阿之阿咗 多多声　//22
招摇撞騙：偷呃拐騙　//2
太过聪明：精过冇尾蛇　//2

上述大辞典穷举的办法能解决相当一部分歧义问题，但还有一些常见字词的遗留问题无法穷举消歧，需要某种上下文条件的规则。譬如，“不好虾弟弟”里面的“虾”是“欺负”的意思，它的条件是后面跟的是表示人的宾语。由于表示“人”的词汇很多，这样的条件词典里面收不胜收。因此，我们利用由词条驱动的专家词典（expert lexicon， Liu，Fu and Li 1989；Li et al 2003）的机制，用这些条目的词典规则来消除歧义。专家词典的本质就是词典化的个性规则集合，以词驱动，例示如下。

虾 // 虾 before person noun is 骗：不好虾弟弟
NEXT:human → *欺负*
D // D before noun is 这： 裹面d蛋都幾香； 你D员工
NEXT:noun → *这*
咩 // 咩 → 什么 unless it is sentence final
!sentenceFinal → *什么*
既 // 既+标点 or sentence final → 的 （=嘅）
NEXT:punctuation → *的*
sentenceFinal → *的*
劲 // as 棒：e.g. 好劲；还有什么叫拍照勁?
PREV:intensifier → *棒*
PREV:verb → *棒*
sentenInitial NEXT:punctuation → *棒*
比 // 比 for passive 被: 东西都比(你)给整没了
NEXT:human? NEXT:verb → *被* //? is optional
极 // V+极+都 = “no matter how” (无论如何)
PREV:verb NEXT:都|有→ *无论如何*

以上转换词典加专家词典的做法足以应对来自报纸和新闻管道的正规粤语文句的翻译了，但是社交媒体不同。由于我们的系统面对社交媒体，粤语处理的难度比传统媒体（譬如香港、广东的地方报纸上的文字）大很多。不少粤语特用的汉字在社会媒体上都是别字。很多人为了打字方便，就胡乱用别字代替（如 【既】 代替 【嘅】，【系】代替【喺】，【左】代【咗】，【比】代【俾】，等等），只要读起来差不多音的，怎么方便怎么来，简直是随心所欲，没有什么约束。最头疼的是这些别字往往是普通话里也常用的汉字，人为造成了很多歧义。譬如，“野”（“嘢”的社交媒体中的变体或别字）在粤语社交媒体中使用极为广泛，它的消歧的有限上下文条件很难尽举。然而，如果已知该文句是粤语，把它翻译成“东西”就相当可靠。

沿着这个思路，我们利用无歧义的粤语作为一个全局杠杆，来帮助消解有限上下文难以解决的歧义（在专名识别领域，利用无歧义的种子通过放大效应（propagation effect）来应对有歧义案例的过程，原理是类似的， 例见 Niu et al 2003）。这本质上是把超出局部上下文的消歧条件以两步走的方法实现。具体的做法是把粤语处理分成两个阶段，首先是识别粤语文句（sublanguage classification）。这个过程利用的是粤语识别词典的信息和一些粤语常用词有关的 heuristics，包括：

（i）词典中标注的粤语专用的字或者词组（标注为加权的 2），一旦出现，该文句分类为粤语：粤语常用的专用字词例举如下；

不就得了：米得咯 咪得啰　//22
别抽烟：米食烟　//2
不是：唔系 唔喺 5喺 吾喺 吴喺　//22222

买东西：买野 买嘢　//22

（ii）对词典中有粤语倾向的字词（标注为加权的 1），表明文句中至少还要有一个权级相同的案例，该文句才能分类为粤语，例如：

别跑：米走　//1
来过：黎过　//1
一丁点：鸡碎 斗零　//11
一定会：实会　//1

还有少数标注为 0 的粤语条目，表明其转换只能在已经分类为粤语的文句中进行，如：

所用的：所用既　//0

0、1、2 的加权标注由粤语词典编纂者根据词条对粤语的指向程度初定，然后通过数据制导的测试来验证和调整。

为了便于词典维护，同时保证随着转换词典的加大及其覆盖面的增加，粤语分类质量也可同步提高，我们把粤语分类词典与粤语转换词典合二为一（前文所有词典示例都是合二为一的词典格式）。分类模块只利用粤语词条及其加权信息（1 和 2，忽略标注为 0 的词条）。翻译模块只利用转换信息，包括标注为 0 的词条：标注为 0 的词条的转换隐含了利用粤语分类的放大效应作为前提条件，这就完全突破了词对词一代机器翻译系统无法利用全局信息的局限。

粤语分类与翻译分为两步还有下列好处：只有当文句被分类为粤语，系统才会去调用粤语转换词典和粤语专家词典去施行到国语的翻译。非粤语的文句就跳过这两个针对粤语方言的过程。（另外一个技术细节值得一提，中文文句的繁简转换安排在粤语到国语的翻译之前进行，因为这样可以避免粤语词典列举繁简组合各变体的负担。）

下面演示一些粤语帖子的自动翻译实例。

點解學生就要平比佢？佢地咩資格去玩器材
→ 为什么學生就要平比他？他们什么資格去玩器材
佢呃人你又知 → 他骗人你又知
搏大霧哩樣野真係唔要得、有欺詐成份
→ 混水摸魚这样东西真是要不得、有欺詐成份
佢話你知? → 他告诉你?
反正你咁八又要投訴， 點解唔問下當時人
→ 反正你这样八又要投訴，为什么不問下當時人
4s冇啦啦壞左 → 4s无端壞了
係咪架, 睇下先 → 是不是架，看下先
好似幾靚, 快啲俾我睇 → 好像蛮漂亮，快一些给我看
琴日係唔知邊到早就見到，我仲以為又係fake
→ 昨日是不知哪里早就見到, 我还以為又是fake
官網既可信性好高 → 官網的可信性好高
而且 Apple 不嬲都唔會一早更新定網站
→ 而且 Apple 向来都不會早就更新定網站
作為iphone的fans明知是呃錢也會買的
→ 作為iphone的fans明知是骗钱也會買的
估佢唔到 → 估他不到
真係唔講都唔知→ 真是不講都不知
有冇多d → 有没有多d
所以~佢應該食埋煙 → 所以~他應該吃完煙
咁樣佢會變成炸蛋人架喎 → 这样樣他會變成炸蛋人架喔
香港迪士尼真係睇唔到有咩前景
→ 香港迪士尼真是看不到有什么前景

初期的粤语分类质量检测用社交媒体测试文句库 5500，粤语文句 3000，国语文句 2500，结果是 99.1% 的精确度（precision），71.2% 召回率（recall）。在把粤语转换大辞典加权并入分类词典以后，综合分类质量检测的结果显著提高，召回率攀升到 86.1%，而精确度只轻微下降到 97.6%。

true positive： 788
false positive： 19
true negative： 4706
false negative： 127

Precision: 0.9764
Recall: 0.8612

这样的质量已经满足支持翻译和抽取的基本要求，我们的初步测试表明粤语文句经过词对词翻译以后，对舆情抽取的精度影响小于 5%，而粤语文句如果不经词对词自动翻译的前处理，舆情精度则会下降 20% 以上。测试用 iPhone 作为品牌条件，主要情感关键词作为共现条件，在系统索引中随机搜集 1000 个粤语的帖子，1000 个国语的帖子，结果如 TABLE 1。

TABLE 1

Sentiment	粤语 (未经前处理)	粤语 (经过前处理)	国语
Positive	48.2%	76.4%	78.3%
Negative	56.8%	81.7%	89.5%
Overall	57.5%	79.1%	83.9%

上述设计和实现的用粤语轻装增强版词对词自动翻译来支持语言分析和舆情抽取已经投入实际应用。两步走的架构合理，基于词典的系统设计具有可扩充性，今后主要就是在维护中开发增强的问题，无需架构上的改变。其实，由于把粤语文句只是当作前处理，整个核心中文系统的分析和抽取不变，分类的误差对最终抽取质量的影响不大，对抽取的影响主要在自动翻译的质量上。而自动翻译主要决定于转换词典的规模：一万词条已经把基本现象涵盖，转换词典还会在开发和维护中逐步增强（incremental enhancement）。随着越来越多的粤语成语、熟语编入词典，越来越多的粤语词语或变体的翻译上下文条件被发现，系统分类和翻译的召回率会随之自然加强而基本不影响精度。

下面给出核心系统的质量检测结果。系统舆情挖掘的质量检测（QA）是这样进行的。由于本系统是为挖掘品牌的客户情报产品而研发，我们通常选取各主要行业有代表性的 10-20 个品牌名作为系统的输入，然后测试舆情情报结果的平均质量。我们使用的是 CrowdFlower 的问卷服务来做检测，每个结果要经过四个匿名人判断，判员之间必须达到 75%（四人中至少三人）的一致意见才计入计算。最近的这次测试选取了下列 15 个品牌作为情报搜索对象：

iPhone|中国电信|丰田|伊利|南航|可口可乐|宝马|家乐福|必胜客|携程|淘宝|苏宁|茅台|蒙牛|麦当劳

TABLE 2

Sentiment	Precision	Relative Recall
Positive	73.2%	14.5%
Negative	89.1%	5.8%
Overall	79.6%	10.2%

由于情感文句具有稀疏数据的性质，对于随机社交媒体文本施行大规模人工标注以求绝对召回率（absolute recall）很不现实。我们于是转而根据情感抽取数在文本中的密度定义了一个相对召回率（Relative Recall）的指标来监测系统对抽取目标的覆盖面的上下波动以及精度与召回率之间的相互影响。中文抽取 10.2% 的相对召回率已经很接近我们英语系统的结果了，而英文系统已经开发数年，相对成熟了，说明开发仅一年的中文系统的召回率不算太落后。当然，我们也注意到语言间的差别，单从情感词汇的统计（基于现代汉语基本词汇五万左右，英语基本词汇四万四

千），现代汉语褒义词是英语褒义词数量的三倍，汉语贬义词也是英语的两倍以上，这从侧面反映了汉语的情感语句的密度可能远远高于英语。因此，汉语情感抽取的覆盖面在今后的进一步开发中应该还有不少余地。

IV. 鲁棒性中文分析和舆情抽取

前面提到，方言轻装翻译前处理的方案可以奏效的前提是，后续的核心分析系统必须具有可以包容机器粗糙译文的鲁棒性。由于我们的中文舆情抽取挖掘系统从一开始就是针对充满了错别字和不规范用法的社交媒体，当初设计开发系统的主要指标就是鲁棒性。不规范的译文不过是在输入端增加了另一种不规范文句而已。因此，我们无需对核心系统做任何改变，只要在管式处理流程中增加粤语翻译的前处理即可运行。本节简要描述后续核心中文系统的分析与抽取，展示粤语前处理输出的粗糙译文如何达成分析和抽取，进而支持粤语社交媒体的舆情挖掘。

情感抽取核心引擎的总体设计是以自浅而深、层层推进的句法分析来支持社交媒体的深度舆情抽取。引擎由两大构建组成。底层是一个鲁棒性中文语法分析器（Chinese parser），顶层是情感抽取器（sentiment extractor）。语法分析器旨在把中文线性字符串（文句）从语言学角度层层处理，条分缕析，生成相应的结构句法树。这样的句法树为下一步的褒贬信息抽取打下了逻辑语义的抽象基础，使得抽取规则对语言现象更具有概括性。引擎通过一连串有限状态文法（finite state grammars, Roche and Schabes 1997）编制的管式规则系统（pipeline rule system）实现，类似的系统形式和架构见（Silberztein 1999；Hobbs 1993；Srihari et al. 2006b）。

中文分析器对社交媒体的文句进行自底而上的层层模式匹配（pattern matching），由浅层分析到深层分析，最终生成带有逻辑语义依从关系的句法结构树（syntactic dependency tree，Li 1989）。下面是粤语帖子“作為 iphone 的 fans 明知是呃錢也會買的”大体转化为国语以后的自动句法分析而得的树形图演示，如图 1。

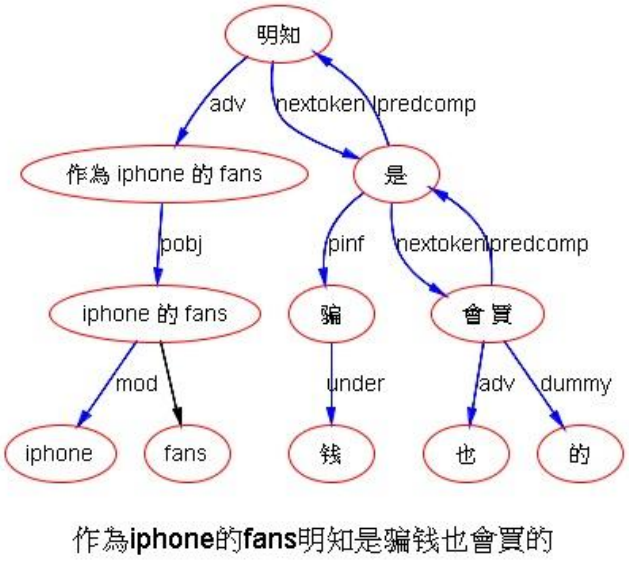


图 1 语帖子“作為 iphone 的 fans 明知是呃錢也會買的”大体转化为国语以后的自动句法分析而得的树形图演示

下图展示另一粤语帖子自动翻译后的句法分析以及在此基础上抽取出来的关于产品（橄榄油）和品牌（DHC）的客户评价，如图 2。

整个舆情挖掘系统由前后两个子系统组成。核心引擎是后台子系统（back-end indexing engine），用于对社交媒体大数据做自动分析和抽取。分析和抽取结果用开源的 Apache Lucene 文本搜索引擎 (lucene.apache.org) 存储。生成后台索引的过程基于 Map-Reduce 框架，利用计算云 (computing cloud) 中 200 台服务器进行分布式索引。对于过往一年的社会媒体大数据存档（约 300 亿文档跨越 40 多种语言），后台索引系统可以在 7 天左右完成全部索引。

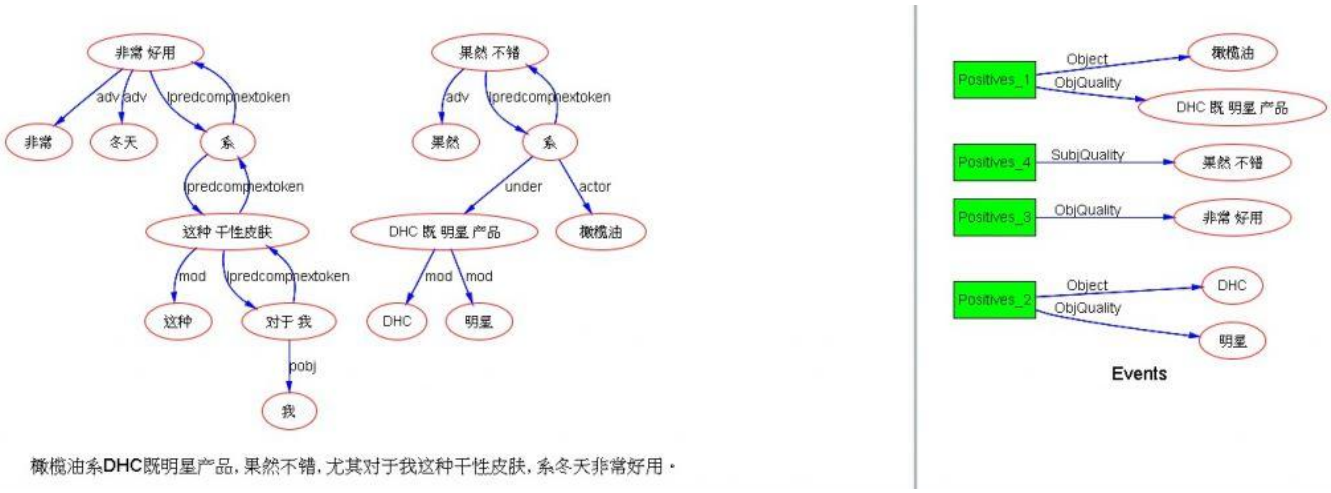


图 2 关于产品（橄榄油）和品牌（DHC）的客户评价

前台子系统（front-end app）是基于 SaaS 的一种类似搜索的应用。用户通过浏览器登录应用服务器，输入一个感兴趣的话题，应用服务器对后台索引进行分布式搜索，搜索的结果在应用服务器经过整合，以用户可以预设（configurable）的方式呈现给用户。这一过程立等可取，响应时间不过三四秒。前台系统负责搜索、挖掘、整合和表达，设计成一个三层的混合后备式模型（hybrid back-off model），以求最大程度地满足不同用户对精确度和召回率平衡的不同的需求。

V. 粤语社交媒体挖掘示例

本节给出围绕热点话题在应用层测试系统的一个实例，说明本文报告的粤语处理技术已经投入实用。下面的社交媒体挖掘，来自中文世界社交媒体过往一年的档案中被系统识别为粤语的部分。我们选择香港娱乐圈的两位名人鐘欣桐（阿嬌）和陳冠希作为挖掘对象，结果如 TABLE 3。从褒贬指标净情绪（NET SENTIMENT）来看，两位的社交媒体形象都不好，陳冠希更是低到-22%，说明网民对他的评论明显贬多于褒。

TABLE 3 鐘欣桐（阿嬌）和陳冠希挖掘信息

鐘欣桐和陳冠希 Summary 2012/6/18 - 2013/6/18					
话 题	提 及	净情绪	强 度	正 面	负 面
鐘欣桐	89,126	-10%	40	3,471	4,242
陳冠希	31,849	-22%	40	1,100	1,714

钟欣桐褒贬指数不高主要是受到以前负面新闻之累，其实粤语地区喜欢阿嬌的粉丝并不少，主要是她长得年轻甜美（年轻/甜美/甜蜜：17.1%）。有意思的是，喜欢她的人很多提到她漂亮的手（18/9%）、眼睛和脸，可见她确实以外形条件见长，如图 3。

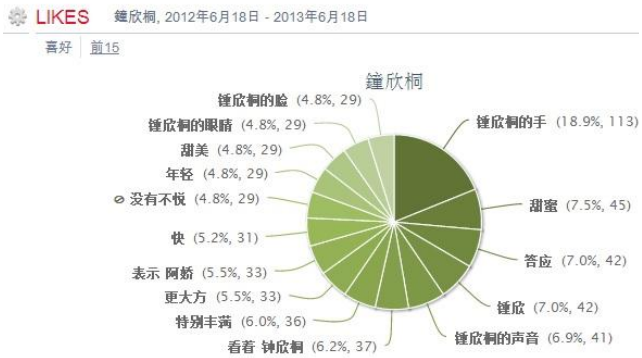


图 3 钟欣桐褒贬指标

至于陈冠希，不管多少负面新闻缠身，女粉丝喜欢（爱）他的仍然不少，见下面的【情绪云图】。具体的不满集中在【褒贬云图】中

大大的那个“搅”字。我们把部分网友议论陈冠　希的粤语帖子附在最后，如图 4。

图 4 陈冠希的粤语帖子

VI. 结语

本文描述了利用改良版的词对词一代机器翻译的方法，把粤语在前处理阶段转换成国语，然后送给核心中文处理器施行社交媒体的舆情抽取。本方案的创新表现在突破了一代机器翻译难以消解歧义的瓶颈，表现在三个方面：（1）利用转换大辞典紧邻上下文消歧，解决了一代系统难以解决的大部分歧义问题；（2）利用专家词典解决了需要上下文类属条件的高难歧义问题；（3）利用先分类后转换两步走的设计解决了需要全局条件才可以解决的歧义问题。本系统的设计思想是把方言看成是社交媒体中已经存在各种不规范表达法之一种，只要把这种变体规范化，主体核心分析器无需改变即可施行对于方言的 NLP 分析和应用。保证这一方案可行性的前提是核心分析器必须具有鲁棒性，否则一代系统的粗糙译文将会导致错误放大（error propagation）效应。这一要求与社交媒体对语言自动分析提出的要求完全一致。我们的测试显示，这条管式系统的方言处理路线是可行和有效的。它为方言的机器处理开辟了一条道路。

致谢

感谢 NLP 组 Ray、Cheng-Ying、Lao Jiang 在中文核心系统研发中的贡献，感谢质量检测组 Xin 为中文情感抽取提供的质量测试报告。还要感谢 Yuan Tian、Qianni、Sophia 在粤语系统初期开发阶段对于数据、词典以及分类做出的贡献。词典资源亦用到“知网”（HowNet、董强、郝长伶、董振东，2003）。

参考文献

[1] X. Ding, B. Liu, and P.S. Yu 2008, “A holistic lexicon-based approach to opinion mining,” Proceedings of the international conference on Web search and web data mining (WSDM ’08). NY, USA. 2008:231-240.

[2] J. R. Hobbs 1993. FASTUS: A system for extracting information from text. Proceedings of the DARPA workshop on Human Language Technology, pp. 133–137. Princeton, NJ.

[3] W.J. Hutchins and H.L. Somers 1992. An introduction to machine translation. London: Academic Press, 1992. [ISBN: 0-12-362830-X]

[4] W. Li 1989. A Dependency Syntax of Contemporary Chinese, BSO/DLT Research Report, the Netherlands.

[5] W. Li, R. Srihari, C. Niu and X. Zhang 2003. “An Expert Lexicon Approach to Identifying English Phrasal Verbs,” Proceedings of ACL 2003. Sapporo, Japan. pp. 513-520.

[6] Z. Liu, A. Fu and W. Li 1989. “JFY-IV Machine Translation System,” Proceedings of Machine Translation SUMMIT II. pp. 88-93, Munich.

[7] C. Niu, W. Li, J. Ding and R. Srihari 2003. A Bootstrapping Approach to Named Entity Classification Using Successive Learners. In Proceedings of ACL 2003. Sapporo, Japan. pp. 335-342.

[8] C. Niu, W. Li, R. Srihari and J. Ding 2004. Orthographic Case Restoration Using Supervised Learning Without Manual Annotation. International Journal of Artificial Intelligence Tools, Vol. 13, No. 1, 2004.

[9] B. Pang, L. Lee and S. Vaithyanathan 2002 Thumbs up? sentiment classification using machine learning techniques. In EMNLP 2002, 79–86.

[10] E. Riloff 2003. Learning extraction patterns for subjective expressions. In EMNLP.

[11] E.Roche and Y.Schabes 1997. Finite-State Language Processing. MIT Press.

[12] M. Silberztein 1999. INTEX: a Finite State Transducer toolbox. Theoretical Computer Science 231(1). Elsevier Science.

[13] R. Srihari, W. Li, T.Cornell and C. Niu 2006. InfoXtract: A Customizable Intermediate Level Information Extraction Engine. Journal of Natural Language Engineering, 12(4), 1-37

[14] M.Taboada, J. Brooke, et al 2011. “Lexicon-Based Methods for Sentiment Analysis,” Computational Linguistics, 2011,37(2): 267-307.

[15] I.Titov and R. McDonald 2008. Modeling online reviews with multi-grain topic models. In WWW, pages 111–120, New York, NY, USA. ACM.

[16] B.Tsou, et al 2005. “Polarity classification of celebrity coverage in the Chinese Press,” Proceeding of the 2005 International Conference on Intelligence Analysis. 2005:137-142.

[17] P.D.Turney 2002, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. USA.2002:417-424.

[18] J. Wiebe 2000. Learning subjective adjectives from corpora. In AAAI.

[19] J. Yang and M. Hou 2012, “基于规则的句子语义倾向计算,” Proceedings of the 13th Chinese Lexical Semantics Workshop (CLSW2012), Wuhan, China. 2012.

[20] 董强，郝长伶，董振东，2003. 基于《知网》的中文语块抽取器，语言计算与基于内容的文本处理（全国第七届计算语言学联合学术会议论文集），孙茂松、陈群秀主编，清华大学出版社

[21] 朱嫣岚,闵锦,周雅倩,黄萱菁,吴立德 2006, “基于HowNet 的词汇语义倾向计算,”中文信息学报, 2006,20(1):14-20.

[22] 李钝,曹付元,曹元大,万月亮 2008, “基于短语模式的文本情感分类研究，”计算机科学, 2008,35(4):132-134.

[23] 王素格，杨安娜，李德玉 2009, “基于汉语情感词表的句子情感倾向分类研究,”计算机工程与应用, 2009,45(24):153-155.

[24] 党蕾,张蕾 2010, “一种基于知网的中文句子情感倾向判别方法,”计算机应用研究, 2010,27(4):1370-1372.

[25] 赵妍妍,秦兵,车万翔,刘挺 2010, “基于句法路径的情感评价单元识别,”软件学报, 2010,21(8):1834-1848.

识别码在大数据时代的要义

胡善庆博士 [1], 丁浩 [2]

[1] 美国联邦政府退休官员，乔治•华盛顿大学教授，Jeremy.s.wu@gmail.com

[2] 华盛顿大学访问学者， edwarddh101@gmail.com

First published as a personal blog on April 30, 2013.

摘要

在 21 世纪，大数据承诺将为社会有效治理以及大众信息分享做出贡献。尽管任何数据本身都包含一定的信息与作用，但是关联和整合后的数据不仅减少收集数据的重复性，而且极度的增加它的價值和可用性。识别码在这个过程中不仅促进实际记录和数据的整合，而且是解放大数据威力的关键。如果识别码没有得到正确的使用和管理，它亦将会是系统失灵、误用和滥用、甚至欺诈及犯罪的元凶。因此，除了技术以外，合理的统计学设计，提高质量的反馈，适当的教育和培训，相关的法律法规，公众的认知，这些都将成为识别码和大数据有效和负责任应用的必要条件。

识别码的必要性

在学生入学时，会有档案存储学生的各种数据，比如：姓名、性别、年龄、家庭背景、专业等。当学生选修一门课并获得成绩时，这个结果也被记录下来。当这个学生满足了所有毕业要求，另一条记录会显示出她的加权平均分并且获得的学位。

每一条记录都是这学生的一个“快照”，随时间累积成为行政记录。这些纵向“快照”提供每个学生受教育情况的丰富信息。

当学生进入工作单位，更多的关于她工作的数据将被收集，伴随她一生，这些数据包括：她的职业及工作单位，工作表现，工资及晋升情况，保险和税的支付数额，就、失业状态等。

在同样的情形下，大量关于公司的数据也会被收集。这些数据记录了：最初注册成立，收支财政状况报告，上市情况，收购或者与其他公司合并，所缴税费，收入增长和雇员增加，公司的扩增或是公司的倒闭。

这些行政记录过去被封存于满身尘埃的文件柜里，但是在千禧年伴随着大数据时代的到来，它们大部份都已数字化。

对学生数据及时和适当的整合将会提供空前的细节，使我们更详细地了解这所学校运作情况，比如说毕业率随时间的变化。当数据整合扩展到所有学校，我们将更好的了解这个国家的教育状况，例如它对就业和经济增长的潜力和支持。这些就是 21 世纪大数据承诺将会给我们带来的变化。从分配资源，评估表现，到制定政策，社会的方方面面都可以从大数据的细节和深度中促进社会有效治理及大众信息分享。

尽管任何数据本身都包含一定的信息与作用，但是关联和整合后的数据将更为重要，因为它不仅减少收集重复的数据，而且极度的增加它的價值和可用性。识别码的重要意义是促进实际记录和数据的整合。在这个过程中，统计学家可以作出卓越的贡献，运用他们的智慧和知识创立新的统计系统。

识别码的种类

当文件例如纸质表格还未被数字化前，人名或者公司名称是被常用的识别码。通常来说，人们会用相同的名称整合记录并给他们排序，比如英文的字母，中文的笔画，或按时间顺序。

但是，使用名字的一大弊端是他们并不是独特唯一，特别是在电脑大量处理数据时，这一弊端尤其明显。据 2006 年的统计，李、王、张、刘这四个中国最大姓氏占了 3.34 亿人口[1]，超过美国人口总数。同样的中文姓名，也有繁体和简体中文的可能分别。英文名罗伯特（Robert）有至少七种不同使用方法，包括：Bert，Bo，Bob，Bobby， Rob，Robbie，以及 Robby，它在 2011 年美国出生的男性人名中的使用率排第 61 位[2,3]，而 Bert 又可以是英文名 Albert 的缩写。个人又有可能更改名字或者有不止一个名字；女性可能在结婚后改名。人为的错误又可能增加不正确的名字。在使用不同语言的情况下，引用同一名字更是特别困难。

在注册的过程中，公司的名称会被检查以确保不出现重名。公司的名称包括它的商标也会被当地，全国性以及国际性的规则和法律受到保护。但公司仍有可能使用多个名称，包括它的缩写和公司股票代码，而且它也有可能在合并，重组，被收购的时候变更名称，或者只是简单的更改品牌。

非唯一的识别码会造成不正确链接和合并数据的风险，导致不正确的结果或结论。虽然给一个名称增加辅助信息，比如说年龄，性别和地址，可以减少风险，但是并不能完全去除错误配对记录和数据的可能性，而且会增加处理数据的时间。

识别码可以由一系列的数字、字母或特殊字符(字母数字)组成。越来越多使用单纯的数字来组成识别码，应用于现代的机器排序，链接和合并电子记录。因为纯数字识别码不依赖于语言系统，受到比较少的限制。使用字母数字的识别码，可能适合使用拉丁语系的系统，但是那些非拉丁语系的系统就比较难以使用、明白或理解。同时，数字字符比较字母数字字符容易排序。

当美国在 1935 年通过社会安全法案时，履行法案遇到的第一个挑战就是创造如何“永久识别每个个体”的识别码，同时保有“以后能够有效和无限制的增加识别对应增长工人的功能”[4]。一个八位字母数字系统最

先被提出来，但很快遭到了统计机构、劳工及法律部门的反对。这个变换被描述为“机器会如何深远地映响[政府] 操作”的第一个徵兆[4,5] 。这些都是在计算机实际使用以前发生的事情。

如今，信息科技的巨大影响很明显，不但是政府，商业和个人活动的方方面面，而且影响力还在不断增强。一个识别码可以应用于个人，一家公司，一辆车，一张信用卡，一箱货物，一个电子邮箱账户，一个地方，或者是任何一个实际个体。

如果一条电子记录不包含识别码，或不能和其他记录連接，大数据中称为缺乏“结构”或者叫做“无结构”。从 21 世纪初开始，“无结构”数据比有“结构”数据出现的频率多得多，但是它们比有“结构”数据包含较少信息内容，也更难应用，特别是在社会和经济方面，我们很难得到后续、连贯和可靠的时序信息。

如何有效地使用识别码将是发挥大数据巨大作用的关键。

有效使用识别码

1. 匹配和合并记录。理想的识别码同时互斥，完全穷尽，在代码和实体间建立了明确的一一对应关系，同时也会延生到未来的记录。识别码促进对电子记录直接有效地排序、匹配及合并，具有无限扩增实体信息内容的潜能。

2. 匿名和保护身份。因为代码是实体的匿名，所以它为身份保护提供第一道防线。但随着识别码重要性的增强，以及它与其他数据链接相对容易，通过识别码伪造及盗用身份的危险性和可能性也在增加，这就需要加强对识别码的政策和负责任的管理，以使它起到保护的作用。

3. 基本描述和分类。识别码可以对数据内容提供最基本的描述，迅速从中得到简单的信息或者是总结。随着时间的推移，这个概念延伸到识别码的分类和“元数据”的发展[6,7]，这个过程包含了在数据系统中建立有效的结构以及扩展它们跨系统的应用。

4. 初部质量检查。无意的人为输入错误以及不正确的转录识别码都可能对整体数据和最终分析结果的质量造成破坏。欺诈或恶意改变识别码亦可能会造成对数据的完整性和可靠性严重的破坏。“效验码”[8,9]在早期检查中的使用可使识别码中常见错误降低 90%。

5. 促进统计学创新。通过对每个学生数据连续不断的收集和整合，可以建立一个含有所有学生和学校丰富信息的动态框架。在严格保护个体隐私和数据安全的同时，新的变数可以定义用做分析研究，描述一所学校的表现或者是一个国家教育状况的统计结论可以是定时或实时产生。在美国及中国，建立这些动态框架和纵向数据系统都已起步[10]。Data Quality Campaign[11]列“唯一州际连接学生数据和主要数据库学生识别码”为建立全美教育纵向数据系统中最关键的部分。

美国和中国的个人识别码

美国没有全国性的个人识别系统。社会安全号创立于 1936 年，还在商业使用电脑之前，用于追踪劳工的收入。在电脑大规模使用以后，社会安全号作为识别码表现出一些优势和劣势，如图 1。

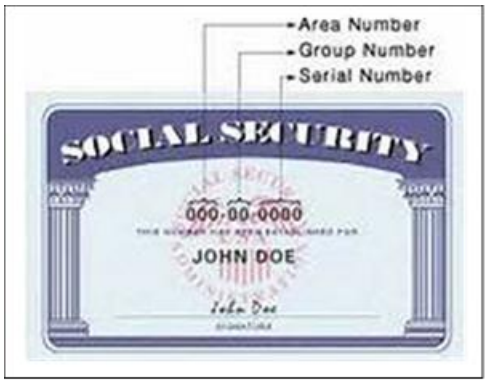


图 1 美国识别码

识别码是九位的社会安全号由三部分组成：

地区号码（三位）- 最初是发放社会安全号的地区代码，后来代表申请邮寄地址的邮政代码

组号（两位）- 代表着一个社会安全号集合被指定为一个组

系列号码（四位）- 从 0001 到 9999

社会安全号的申请过程中[12]收集人口信息，包括名字、出生地、出生日期、国籍、种族、性别、父母的名字和社会安全号、电话号码和邮政地址。美国社会安全部负责社会安全号的发放。有一些社会安全号被保留，没有使用。一旦一个社会安全号被发放，它应是唯一的，因为它不会被第二次发放。但重复的情况仍可能存在。

1938 年，一个钱包的生产厂商显示他们的产品是多么适合社会安全号卡来促销其在百货商场出售的钱包，但是他们使用一张自己员工的社会安全号卡[13]。这导致有四万人错误的使用了这个社会安全号，甚至到 1977 年还有人将这个号码做作为自己的社会安全号。

自从社会安全号的产生，它被政府部门和私有企业的使用显著增加。从 1943 年开始，总统行政命令要求各联邦政府部门必须使用社会安全号建立拥有永久账户号码的系统[5]。在 1960 年代初，政府雇员和个体报税者必须使用社会安全号。到 1960 年代末，社会安全号被作为军人的识别码。在整个七十年代，当电脑被越来越多使用后，金融活动，如开设新银行账户和申请信用卡和贷款，以及联邦福利的运行中，社会安全号成为必不可少的一部分。从 1986 年开始，如果父母想要有受抚养人的免税，就必须将其抚养人的社会安全号也列在税表里。在法律实施的第一年，这反欺诈行动减少了七百万的受抚养人数[14]。

社会安全号可以将同一个人的很多电子文件链接合并到一起，因此它本质上作为了非官方全国性识别码，但是它也可能直接造成误用或者滥用，例如身份盗用[15]。社会安全号没有效验码，它并不能有效的作为身份的認证。有学者也展示如何用公开的信息“异常精确的重建社会安全号”[16]。这些年

图 2 中国相对较晚开始使用个人识别码

在美国，识别码的这些脆弱点使得人们更加小心谨慎和负责任的使用社会安全号。1943 年要求使用社会安全号的行政命令也被废除，取而代之的是在 2008 年颁布的行政命令使社会安全号成为可以选择而非必须的。

中国相对较晚开始使用个人识别码，如图 2。在 1999 年 7 月 1 日，身份证号码由 15 位提升为 18 位，其中出生年份由两位变为四位，并且增加效验码。18 位身份证号由四部分组成[17,18]：

识别码的要义地区代码（六位）—个人住址的行政编号

生日代码（八位）—按生日的年月日顺序组成

系列代码（三位）—其中奇数代表男性，偶数代表女性

效验码（一位）—使用 ISO 7064 标志算法，基于前面 17 位数字计算得到[18,19]

居民身份证由居民常住户口所在地的县级人民政府公安机关基于未满 16 岁居民的申请签发。居民身份证登记的项目包括：姓名，性别，民族，生日以及居住地址。居民身份证有效期长至永久，也可能短至五年，取决于申请人的年龄。根据官方的声明，居民身份证号在中国电子健康档案中也用于记录个人的健康信息[20]。

中国及美国的商业识别码和工业分类码

美国企业的僱主识别码相当于个人的社会安全号[21]。但是，这里的企业包括地方，州和联邦政府，也包括无雇员的公司，亦包括需要为其雇员缴纳税款的个人公司。僱主识别码是由美国税务局负责指派的一个九位数字，它的形式是 GG-NNNNNNN，其中 GG 在 2001 年前是公司所在地的代码，而后七位数字没有特别的含义。一旦一个僱主识别码被使用，美国税务局就不会再次使用。另外，每个州亦各有自己的僱主识别码用于税务收缴和行政管理。

在联邦僱主识别码的申请过程中有以下信息被收集:正式名称，交易名称，法人姓名，责任人员，邮政地址，商业地址，公司类型，申请原因，成立时间，财政年度，未来 12 个月员工数目估计，首次工资发放日期以及公司主营业务[22]。

美国统计部门使用北美工业分类系统（以下称 NAICS）来对公司营业进行归类，以期能收集，分析以及发布美国经济的统计信息[23]。在 1997 年，北美工业分类系统（NAICS）继承取代工业标准分类系统(SIC)。

NAICS 是一个层级分类代码系统，其中可能包含有 2 到 6 位数字。最高层级的 2 位数字代表主要经济部门，例如建筑和生产。每个 2 位数字所代表的部门都包含一系列的 3 位数字子部门，而它又包含有一系列 4 位数字的工业集团。例如 31 到 33 是表示生产部门，而碾米工业在其所属层级之中：

311 食品加工制造业
3112 粮食和油菜籽加工业
31121 面粉和麦芽生产业
311212 碾米业

层级系统其中的一个优势就是它可以相当容易地链式聚集产业总值。比如说，所有代码为 311x 企业的总和就组成了代码为 311 的食品加工制造业。

持续用 NAICS 代码準确地把企业分类是一大挑战，因为在当今快速变化的动态国际经济环境下，一夜之间过时的行业会被淘汰，

新的行业也会出现成长，过去的“高科技企业”及最近的“绿色”行业就是例子之一。使用 NAICS 代码的过程中有着理解和持续性的问题，例如美国统计局和美国劳工统计局就因为数据来源和 NAICS 代码分类不同，令到各自创立和维护的商业框架有異[10]。不一致使用 NAICS 代码破坏甚至造成对时间序列和纵向数据分析无效。

中国的新企业必须向当地质量监督局申请 9 位数的国家组织机构代码，其中由 8 位数字（或大写拉丁字母）本体代码和 1 位数字（或大写拉丁字母）效验码组成。中国的组织机构代码，借鉴原 ISO6523《数据交换标识法的结构》（现 ISO6523《信息技术组织和组织各部分标识用的结构》）国际标准的基础上，根据 GB 11714—1997《全国组织机构代码编制规则》国家标准的规定，编制的全国统一的组织机构代码识别标识码[25]。可以通过网上信息核查系统基于国家组织机构代码查询组织机构的信息[26] 。

国内及国外的经济学家和其他学者十分认可中国工业企业数据库的价值。透过相当的投资，这个丰富的综合数据系统从 1998 年开始纵向描述中国差不多所有的国有和大型企业（2010 年前销售额在 500 万人民币以上及 2010 年后销售额在 2000 万人民币以上的企业）。但是，十分严重的质量问题已有所报导，而主要数据错误原因可以追溯到不正确和不连贯地使用识别码[27]。虽然中国从 1989 年就开始标准化国家组织机构代码，并且现在已经进行到了第三阶段，但是这个问题仍然存在[28]。

就在上个月，广东省宣布他们运用国家组织机构代码这个平台推动反腐败[29]。中国也有一个根据 GS-T4754-2002 文件而建立的标准工业分类系统[30]。这个层级系统有四类，其中最高层为一个字母，其余分别有 2 位、3 位、4 位数字代码表示较低层级。以前述的的碾米业为例，中国的分类系统中表示为以下层级：

C　　制造业
C13　　农副食品加工业
C131　　谷物磨制
C1312　　大米加工企业

总结

随着科技的轉变和发展，收集大规模的数字化数据的成本将更低，速度也将更快。这些是大数据时代的标志。

这些大数据包含了空前规模的信息。如果数据整合和结构化，它们的价值和功能将会暴涨，超过现有数据系统所能提供。识别码促进数据的链接和合并，是提供这些巨大机会的关键。

识别码能解放大数据的巨大能量。如果我们不能正确使用和管理识别码，它同样可以成为系统失灵，误用和滥用，甚至是欺骗和犯罪行为的罪魁祸首。

现实使用识别码的挑战是多样复杂。除了科技技术，统计学设计和质量回馈途径，适当的教育和培训，有效的政策和监控，以及公众的意识參与都是有效负责使用识别码所必须的。未来的文章中将讨论这些话题。

参考文献

[1] 360doc.com. Quantitative Ranking of Chinese Family Names (中國姓氏人口數量),November 25, 2012. Available at

http://www.360doc.com/content/12/1125/17/6264479_250155720.shtml on April 29, 2013.

[2] Wikipedia. Robert. Available at http://en.wikipedia.org/wiki/Robert on April 29, 2013.

[3] U.S. Social Security Administration. Change in Name Popularity. Available at http://www.ssa.gov/OACT/babynames/rankchange.html on April 29, 2013.

[4] U.S. Social Security Administration. Fifty Years of Operations in the Social Security Administration, by Michael A. Cronin, June 1985. Social Security Bulletin, Volume 48, Number 6. Available at http://www.ssa.gov/history///cronin.html on April 29, 2013.

[5] U.S. Social Security Administration. The Story of the Social Security Number, by Carolyn Puckett, 2009. Social Security Bulletin, Volume 69, Number 2. Available at http://www.ssa.gov/policy/docs/ssb/v69n2/v69n2p55.html on April 29, 2013.

[6] Wikipedia. Metadata. Available at http://en.wikipedia.org/wiki/Metadata on April 29, 2013.

[7] Wikipedia. 元数据. Available at http://zh.wikipedia.org/wiki/元数据 on April 29, 2013.

[8] Wikipedia. Check Digit. Available at http://en.wikipedia.org/wiki/Check_digit on April 29, 2013.

[9] Wikipedia. 效验码. Available at http://zh.wikipedia.org/wiki/校验码 on April 29, 2013.

[10] Wu, Jeremy S. 21st Century Statistical Systems, August 1, 2012. Available at http://jeremyswu.blogspot.com/2012/08/abstract-combination-of-traditional.html on April 29, 2013.

[11] Data Quality Campaign. 10 Essential Elements of a State Longitudinal Data System.Available at http://www.dataqualitycampaign.org/build/elements/1 on April 29, 2013.

[12] U.S. Social Security Administration. Application for a Social Security Card, Form SS-5.Available at http://www.ssa.gov/online/ss-5.pdf on April 29, 2013.

[13] U.S. Social Security Administration. Social Security Cards Issued by Woolworth. Available at http://www.socialsecurity.gov/history/ssn/misused.html on April 29, 2013.

[14] Wikipedia. Social Security Number. Available at http://en.wikipedia.org/wiki/Social_Security_number on April 29, 2013.

[15] President's Identity Theft Task Force. 2007. Combating Identity Theft: A Strategic Plan. Available at http://www.idtheft.gov/reports/StrategicPlan.pdf on April 29, 2013.

[16] Timmer, John. New Algorithm Guesses SSNs Using Data and Place of Birth, July 6, 2009. Available at http://arstechnica.com/science/2009/07/social-insecurity-numbers-open-to-hacking/ on April 29, 2013.

[17] baidu.com. GB11643-1999 Citizen Identity Number 公民身份号码. Available at http://wenku.baidu.com/view/4f19376348d7c1c708a14587.html on April 29, 2013.

[18] Wikipedia. Resident Identity Card. Available at http://en.wikipedia.org/wiki/Resident_Identity_Card_(PRC) on April 29, 2013.

[19] Wikipedia. ISO 7064. Available at http://en.wikipedia.org/wiki/ISO_7064:1983 on April 29, 2013.

[20] baidu.com. Electronic Health Record 电子健康档案. Available at http://wenku.baidu.com/view/348d5a18a300a6c30c229fec.html on April 29, 2013.

[21] Wikipedia. Employer Identification Number.Available at http://en.wikipedia.org/wiki/Employer_identification_number on April 29, 2013.

[22] U.S. Internal Revenue Service. Form SS-4: Application for Employer Identification Number. Available at http://www.irs.gov/pub/irs-pdf/fss4.pdf on April 29, 2013.

[23] U.S. Census Bureau. North American Industry Classification System.Available at http://www.census.gov/eos/www/naics/index.html on April 29, 2013.

[24] National Administration for Code Allocation to Organizations. Introduction to Organizational Codes, 组织机构代码简介. Available at http://www.nacao.org.cn/publish/main/65/index.html on April 29, 2013.

[25] Wikipedia. ISO/IEC 6523. Available at http://en.wikipedia.org/wiki/ISO_6523 on April 29, 2013.

[26] National Administration for Code Allocation to Organizations. National Organization Code Information Retrieval System, 全国组织机构信息核查系统. Available at http://www.nacao.org.cn/ on April 29, 2013.

[27] Nie, Huihua; Jiang, Ting; and Yang, Rudai. A Review and Reflection on the Use and Abuse of Chinese Industrial Enterprises Database. World Economics, Volume 5, 2012. Available at http://www.niehuihua.com/UploadFile/ea_201251019517.pdf on April 29, 2013.

[28] National Administration for Code Allocation to Organizations. Historical Development of National Organization Codes, 全国组织机构代码发展历程. Available at http://www.nacao.org.cn/publish/main/236/index.html on April 29, 2013.

[29] National Administration for Code Allocation to Organizations. Guangdong Aggressively Promotes the Use of identification Codes in its Campaign against Corruption, 广东积极发挥代码在反腐倡廉中的促进作用, March 7, 2013. Available at http://www.nacao.org.cn/publish/main/13/2013/20130307150216299954995/20130307150216299954995_.html on April 29, 2013.

[30] baidu.com. National Economic Industry Classification, GB-t4754-2002, 国民经济行业分类(GB-T4754-2002)(总表). Available at http://wenku.baidu.com/view/69f04af8c8d376eeaeaa31cf.html on April 29, 2013.

作者简介：

胡善庆博士,美国联邦政府退休官员。现任乔治•华盛顿大学统计系客座教授及上海华东师范大学大数据创新中心主任。邮件：jswu@gwu.edu

丁浩，现为华盛顿大学访问学者及数据治理促进会成员，毕业于华中科技大学和乔治华盛顿大学。

中国如何应对大数据时代的挑战

涂子沛

最近，我回国参加了一些大学的研讨会、政府部门的座谈会以及企业的培训活动，主题都是大数据，时下，国内各大报刊杂志也都在探讨这个热门话题，但我发现，中国社会对“大数据”的概念还存在一些不准确的认识甚至观念上的误区，特别是对大数据在国家层面上的战略意义认识不足，亟须深化。

从小数据到大数据

“大数据”是一股新的技术浪潮，也是逐步形成的历史现象，其具体是指随着信息存贮量的增多，人类在实践中逐渐认识到，通过数据的开放、整合和分析，能发现新的知识、创造新的价值，从而为社会带来“大科技”、“大利润”、“大智能”和“大发展”等新的机遇。大数据概念的提出，可以追溯到 1980 年代，但其“数据”二字却和我们传统的理解有所不同。

传统意义上的“数据”，是指“有根据的数字”，但在进入信息时代之后，“数据”二字的内涵在扩大，它不仅指代“数字”，还统称一切保存在电脑中的信息，包括文本、声音、视频等。更重要的是，随着信息技术的进步，其数量在呈指数增长特别是新媒体出现之后，数据的收集、保存、维护、使用等任务，成为横跨各个领域的现象和挑战。

大数据之“大”，并不在于其表面的“大容量”，而在于其潜在的“大价值”。有很多例子可以证明，由于新工具的出现，我们从以前的小数据当中也能发现大的价值。例如，美国把二十多年的犯罪数据和交通事故数据映射到同一张地图上后惊奇地发现，无论是交通事故和犯罪活动的高发地带，还是两者的频发时段，都有高度的重合性。这引发了美国公路安全部门与司法部门的联合执勤，通过共治数据“黑点”，交通事故率和犯罪率双双降了下来。再例如，最近有学者将白宫 200 多年总统洗衣服的记录电子化，然后进行分析，也得出了一些新的结论。这些数据，都是地道的小数据。这说明，小数据只要在纵向上有一定的时间积累，在横向上有细致的记录粒度，再和其他数据整合，就能产生大的价值。从这个角度来看，大数据也可以理解为针对某个对象在时空两个维度上的“全息”数据。这种“全息”，在大数据的时代还表现为“多源”，即有多个源头在从不同方向对同一个对象进行数据记录，数据之间可以互相印证。

另外，从全球数据技术投入的资金分布来看，传统的小数据仍然占据绝对的重头。据国际数据集团(IDG)统计，2012 年，全球对小数据分析工具的投资为 349 亿美元，而对大数据分析工具 Hadoop 的投资仅为 1.3 亿美元，不及前者的 1%。IDG 的结论是，传统的小数据软件满足了企业和组织 95%的需求。目前行业发展的最新态势，是“大”、“小”数据分析工具趋于一体化并在向“云”迁徙。

大数据的战略意义

大数据的意义，也远远不局限于我们当前众多新闻报道中所津津乐道的“啤酒和尿布”等通过数据挖掘、实现精准营销的故事。事实上，数据挖掘已经不是大数据领域的前沿，取而代之的是机器学习。数据挖掘是指通过特定的算法对大量的数据进行自动分析，从而揭示数据当中隐藏的历史规律和未来的发展趋势，为决策者提供参考。时下兴起的机器学习，凭借的也是计算机算法，但和数据挖掘相比，其算法不是固定的，而是带有自调适参数的，也就是说，它能够随着计算、运行次数的增多，即通过给机器“喂取”数据，让机器像人一样通过学习逐步自我提高改善，使挖掘和预测的功能更为准确。这也是该技术被命名为“机器学习”的原因。这也是大数据之所以被称为革命性现象的根本原因，因为从本质上来说，它标志着我们人类社会在从信息时代经由知识时代快速向智能时代迈进。

不妨举一两个例子，来说明大数据对社会形态的影响以及对国家战略的重要性。

今年以来，一股在线教育的浪潮正在席卷美国的教育领域，一种新型的智能学习平台正在成为高科技领域创新和投资的重点，其中不少公司已经获得了初步成功。如著名的在线教育公司 Coursera，已经和普林斯顿、伯克利、杜克、香港理工等全世界 30 多所大学达成协议，通过其平台免费开放课程。如今这些学校的课程可以实现全球几十万人同步学习。分布在世界各地的学习者不仅可以在同一时间听取同一位老师的授课，还和在校生一样，做同样的作业、接受同样的评分和考试。一些学校看到了这种智能学习平台的价值和潜力，甚至开始投资兴建自己的独立平台，2012 年 5 月，哈佛大学与麻省理工学院就宣布，将投入 6000 万美元开发一个类似平台，并向全世界免费开放。

这种学习平台的崛起，在美国引起了广泛的关注和激烈的讨论。其中的原因，是因为该平台已经不是一个镜头、一段视频那么简单，而能对学习者的学习行为自动进行提示、诱导和评价，从而弥补没有老师面对面交流指导的不足。例如，通过记录鼠标的点击，计算机能够记录你在一张幻灯片上停留的时间，判别你在答错一道题之后有没有回头复习，发现不同的人对不同知识点的不同反应，从而总结出哪些知识点需要重复或强调，哪种陈述方式或学习工具在哪种情况下最有效等规律。

不难发现，该平台之所以强大，正是因为大数据。单个个体学习行为的数据似乎是杂乱无章的，但当数据累积到一定程度时，群体的行为就会在数据上呈现一种秩序和规律。通过收集、分析大量的数据，就能总结出这种秩序和规律，然后有的放矢，对不同的学习者提供有针对性的帮助。哈佛大学和麻省理工学院之所以向全世界免费开放其学习平台，目的也是想让更多的学习者在上面学习，以收集更多的数据，有了数据，它们才能研究世界各国学习者的行为模式，进而打造更好的智能学习平台。

数据驱动的智能时代

前面的例子说明数据正在成为组织的财富和创新的基础，也证明大数据确实在催生一个更加智能的社会。那么，又该如何理解我们正在迈进的这个智能型社会呢？

理解这个问题的关键在于，无论是信息、知识还是智能，在我们这个时代，都是以数

据为载体存在的。数据是对客观世界的记录，当我们对数据赋予背景时，它就成为信息；信息是知识的来源，当把信息提炼出规律的时候，它就上升为知识；知识是智能的基础，当电脑、网络能够利用某种知识作出自动判别，采取行动为人类服务的时候，机器智能就产生了。目前，人类记录周围世界的范围正在不断扩大，过去，我们是决定记录什么，现在及将来，我们要进入一个决定不记录什么的时代，同时数据分析的能力不断增强，这都将加速我们迈向智能时代的步伐。智能时代的特点，是无处不在的计算机和网络将像有智商的人一样为人类工作和服务。换句话说，越来越多的工作将被计算机或者机器人所代替。此外，由于精准的计算和预测，整个社会可以像无数个大大小小的齿轮轴承一样，环环相扣，齿齿吻合，日常管理通过数据更加优化，各种任务、合作可以无缝对接，社会运行的成本可大幅降低。

回到上面的例子，不难想象，这种智能学习平台将会给教育行业带来怎样的影响。学校曾经是最重要的教育资源，好的学校更是异常稀缺，由于这种智能平台的普及，在不远的将来，名校将人人可上，也就是说，如果应对得当，中国教育资源匮乏的问题将很快得到有效缓解。对个人来说，随时随地地学习、终身学习都将成为可能，例如，高中生可以尝试大学的课程，离开了校园的人，也可以登录在线平台再和在校生一起听课。这些都是教育工作者探讨多年、孜孜以求的梦想。但硬币的另一面，是中国的教育行业要面对更加激烈的全球化竞争和挑战。过去，是学生争学校；将来，可能是学校在全球范围中争夺学生。发达国家的一流大学会挤压发展中国家普通大学的生存和发展空间，普通大学该如何来吸引生源？它们会不会因此衰落？既然最好的教学视频等学习资源都可以免费获得，教师的角色又需不需要调整？又该如何调整？这些问题，都是大数据时代催生的重大挑战。

智能学习平台只是大数据大潮在教育领域掀起的一朵浪花。毫不夸张地说，大数据将影响人类社会发展的方方面面、优化改造每一个行业，其作用难以限量。我们再以时下另外一个热门名词“智慧城市”为例。近几年来，国内外都兴起了建设智慧城市的浪潮。据国内智慧城市的领军公司神州数码董事局主席郭为介绍：目前，国内已经有 60 多个城市把建设智慧城市纳入了“十二五”规划，他相信，智慧城市将成为推动中国经济可持续发展的主动力。但从一个更高的角度来看，智慧城市的建设问题，其实是一个城市的大数据综合治理问题：一是要在以前没有收集数据的地方收集数据，这主要是利用物联网的技术；二是要让不同系统的数据有效对接起来，这是系统整合的任务；最后，还要利用数据可视化的技术把海量数据中隐藏的知识揭示、展现出来，让数据中的智慧能够以一种直观的形式流向城市的管理者、决策者和市民大众。也就是说，数据的收集、整合、分析、展现才是智慧城市的核心，未来的智能型城市，必将是数据驱动的城市，而大数据则相当于智慧城市的大脑。郭为还指出，智慧城市的建设，是在用信息技术解决社会治理中的难题，提高人民的幸福指数，这又证明，大数据的应用和价值，绝对不仅仅是在商业领域这么简单。

除了推进社会形态的跃进、加速企业创新，引领新的经济繁荣，我在《大数据：正在到来的数据革命》一书中还指出，通过开放数据，大数据还可以成为启动透明政府的利器。这对当下的中国，现实意义毋庸讳言。

也正是因为以上种种战略考量，2012 年 3 月，美国联邦政府宣布投入巨资启动大数据的研发任务，并把大数据提到了和历史上的互联网、超级计算机一样的高度，成为国家战略。

政府需要做什么

一是政府机构、行业组织和大型企业要建立专门的数据治理机构来统筹数据治理的工作，例如数据治理委员会、大数据管理局等。数据治理的重点在于数据定义的一致性和数据的质量。在大数据时代，不同系统之间的数据要进行整合，因此要有统一的元数据定义，这不仅是中国而且是全世界当下都在面临的挑战。各个领域和行业的数据标准制定得好，将会起到事半功倍的效果。就单个企业而言，要认识到，未来的竞争是知识生产率而不是劳动生产率的竞争，数据分析产生的价值可能比较碎片化，分布在商业流程的各个环节，数据挖掘的投资回报也有不确定性，但企业领导必须有眼光，把数据治理的工作尽快统筹起来，为增强企业在大数据时代的竞争力做好准备。此外，数据治理机构的首长应该由组织的高层领导担任，否则标准无法推进到全局，也改善不了整个行业或组织的情况。

二是开放数据。数据增值的关键在于整合，但自由整合的前提是数据的开放。开放数据是指将原始的数据及其相关元数据以可以下载的电子格式放在互联网上，让其他方自由使用。开放数据和公开数据是两个不同的概念，公开是信息层面的，是一条一条的；开放是数据库层面的，是一片一片的。开放也不一定代表免费，企业的数据，可以以收费的形式开放。开放也是有层次的，可以对某个群体、某个组织，也可以对整个社会开放。在大数据的时代，开放数据的意义，不仅仅是满足公民的知情权，更在于让大数据时代最重要的生产资料数据自由地流动起来，以催生创新，推动知识经济和网络经济的发展，促进中国的经济增长由粗放型向精细型转型升级。

三是鼓励、扶持基于数据的创新和创业。政策扶持的传统方法，可能是以政府为主导建立大数据产业园，对新兴企业提供办公场所等便利条件或者现金支持，这固然有效，但更有效的方式是调动全社会的力量。例如，拨款支持大数据开源社区、程序员协会等民间组织的建设，通过扶持类似的民间团体，快速推进新技术、新理念在全社会的传播和普及；再例如，以开放的数据为基础，举办应用程序开发大赛，向全社会征询数据使用、创新的意见，主办方可以是政府，也可以是企业，拿出一定的资金，奖励最优秀的应用程序，激发民间蕴藏的创新力量。

四是要在全社会弘扬数据文化。数据文化，是尊重事实、推崇理性、强调精确的文化。要承认，回望历史，中国是个数据文化匮乏的国家，就现状而言，中国数据的公信力弱、质量低，数据定义的一致性差也是不争的事实。这方面，政府应该发挥主导作用，首先在公共领域推行数据治国的理念，要认识到，在大数据时代，公共决策最重要的依据将是系统的数据，而不是个人经验和长官意志，过去深入群众、实地考察的工作方法虽然仍然有效，但对决策而言，系统采集的数据、科学分析的结果更为重要。政府应加大数据治国的舆论宣传，将数据的知识纳入公务员的常规培训体系，力争在全社会形成“用数据来说话、用数据来管理、用数据来决策、用数据来创新”的文化氛围和时代特点。

最后是要围绕个人数据安全，逐步加强隐私立法。任何技术都是双刃剑，大数据也不例外。如何在推动数据开放的同时有效地保护公民隐私，将是大数据时代的一个重大挑战。

2013 年，是全世界的大数据年。希望中国政府相关部门尽快制定和大数据相关的政策，出台具体的措施，从而抓住历史的机遇、推动中国社会的发展和进步。

作者简介：

涂子沛：知名专栏作家、信息管理专家，华南理工大学公共政策研究院副教授，中国旅美科协匹兹堡分会主席。涂子沛先后毕业于华中科技大学、中山大学和卡内基梅隆大学，现担任美国一软件公司数据中心的主任。赴美留学之前，曾在省、市、县几级政府的不同部门磨砺 10 年，做过职业程序员，担任过公安边防巡逻艇的指挥官,也从事过政府统计工作。目前为《南方都市报》、《IT 经理世界》、艾瑞网等国内多个报刊网站撰写专栏。著有《大数据：正在到来的数据革命》。


BIG
DATA
USE
ONE
SOFTWARE
USING
WORKINGS
GARTNER
MASSIVELY
BIOLOGICAL
HUNDREDS
ANALYTICS
BUSINESS
MOVING
SAN
PARALLEL
TOLERABLE
SIZE
NEEDED
PETABYTES
INTERNET
TECHNOLOGIES
PERFORMANCE
ALSO
DISK
GENOMICS
SHARED
TIME
RELATIONAL
TOOLS
SET
WITHIN
PROCESS
DEFINITION
SEARCH
INCLUDE
SYSTEMS
NETWORKS
INFORMATION
RECORDS
TENS
CAPACITY
PRACTITIONERS
PROCESSING
LOGS
DISTRIBUTED
CAPTURE
MAY
MANAGEMENT
MANAGE
GROW
SOCIAL
LARGE
STORE
EVERY
LARGER
CASE
COST
CONTINUES
CITATION
TERABYTES
MPP
RESEARCH
COMPLEX
DATABASES
STORAGE
DIFFICULTY
TARGET
ABILITY
SETS
EXAMPLES
APPLIED
ELAPSED
CURRENT
SENSOR
ARCHIVES