

基于多社交媒体的个体信息融合 关键技术研究

学生姓名：詹 辉

指导老师：崔晓晖 教授

内容提要

一、论文背景：

1. 社交媒体的使用现状；
 2. 本阶段的社交媒体选择。
-

二、系统需求与设计：

1. 本系统的需求分析；
 2. 本系统的模块设计。
-

三、系统实现：

介绍本系统关键部分的实现：

1. 数据采集；
 2. 实体相关词抽取；
 3. 信息融合。
-

四、处理结果：

显示本系统的处理结果。

一、论文背景

1. 融合的信息是什么？

(P9. 1.1.1选题的背景)

2. 为什么选择 脸谱网 与 推特 作为现阶段的目标社交媒体？

(P16. 2.3数据来源)



二、系统需求与设计

1. 总体需求

获取一个个体在多个社交媒体中补充后的较为完整的信息。

2. 功能需求

- 1) 数据采集;
- 2) 文本预处理;
- 3) 文本实体抽取;
- 4) 实体相关词抽取;
- 5) 信息融合;
- 6) 处理结果读取;
- 7) 处理结果显示。

(P18. 3.1系统总体功能需求)

3. 模块设计

数据采集模块

采集推特数据

采集脸谱网
数据

批量采集短社
交文本语料库
数据

文本处理模块

批量文本
预处理

文本
预处理

文本实
体识别

实体相关
词抽取

信息
融合

处理结果显示模块

读取处理结果

显示处理结果

(P23. 4.1.2系统模块设计)

三、系统实现

数据采集

1. 推特数据采集

采集地址: `mobile.twitter.com`

使用工具: `requests`

2. 脸谱网数据采集

采集地址: `m.facebook.com`

使用工具: `selenium`

(P35. 5.2数据采集模块)

文本预处理
文本实体识别
实体相关词抽取
信息融合
处理结果读取
处理结果显示



三、系统实现

文本预处理

1. 文本切词

使用工具：Noah's ARK 切词工具

词切分示例：[词] [切] [分] [词] [切] [分] [词]

2. 单词标准化

使用工具：NLTK 词形还原工具 (Lemmatization)

词形还原示例：[词] [形] [词] [形] [词] [形] [词]

3. 词性标注

使用工具：NLTK 词性标注工具

词性标注示例：PR VB RB NN JJ NN

(P38. 5.3.1 预处理文本)

文本实体识别

文本实体识别

使用工具：MITIE 命名实体识别工具、NLTK 实体识别工具

命名实体识别示例：PR VB RB NN JJ NN

(P39. 5.3.2.1 文本实体识别)

实体相关词抽取

信息融合

处理结果读取

处理结果显示

三、系统实现

实体相关词 抽取

1. 基于文本词性句型的相关词抽取方法

举例说明：以（句型）NN NN VB JJ NN NN NN NN NN 为例

已训练
 My daughter @lucy1995 won the special prize of this year's
 annual painting competition.

待抽取
 Jonathon Fury sent me a cute ted from Las Vegas
 Roling-Madam souvenir shop!

Fury 相关词: sent, cute, ted, souvenir, shop.

信息融合

处理结果读取

处理结果显示

（P39. 5.3.2.2实体相关词抽取（1）句型方法）

三、系统实现

实体相关词
抽取

1. 基于文本词性句型的相关词抽取方法

对500个常见的词形句型进行手动标记，得到每一种句型对应的实体词抽取办法。（使用这500个常见句型的句子占整个语料库的10%）

人工标记：

```
STAT: Start marking 23 pattern.
STAT: Pattern: NN NN NN VB NN NN NN NN JJ.
STAT: Mark 7 text of the pattern.
      0          1      2      3      4
      NN          NN      NN      VB      NN
A      Glenview      man      wa      found      shot      to
5
      NN          6      NN          7      8
      NN          NN          NN          JJ
death in Chicago over the weekend : trib.in/1GZv0rW
The indice of relevance: 1 3 5 7
```

（P39. 5.3.2.2实体相关词抽取（1）句型方法）

信息融合

处理结果读取

处理结果显示

三、系统实现

实体相关词
抽取

2. 基于卡方统计量的相关词抽取方法

卡方统计量常运用在检验两个事件的相互独立性。利用卡方统计量计算实体词与实体词所在文本中的每一个词的相互独立性，得到的结果中与实体词不相互独立的词，即为该实体的相关词。

举例说明：



在整个语料库中统计“实体词”与“词A”一起出现、一起不出现的频率，如果“词A”在语料库中明显呈现出与“实体词”同时出现以及同时不出现，则可断定“词A”与“实体词”不是相互独立的。

实体词与词A同时出现	56	实体词出现、词A不出现	7
实体词不出现、词A出现	21	实体词与词A同时不出现	19450

实体词与词B同时出现	1	实体词出现、词B不出现	58
实体词不出现、词B出现	41	实体词与词B同时不出现	19436

(P41. 5.3.2.2 实体相关词抽取 (2) 卡方方法)

信息融合

处理结果读取

处理结果显示

三、系统实现

实体相关词
抽取

2. 基于卡方统计量的相关词抽取方法

卡方统计量则是计算两个词之间是否相互独立的量化数值，若卡方统计量超过10.83，则该二词相互独立的概率将低于0.001，近似与小概率事件。公式如下：

e_{11} 计算公式 (e_{10} 、 e_{01} 、 e_{00} 计算公式以此类推)

$$e_{11} = n * \frac{n_{11} + n_{10}}{n} * \frac{n_{11} + n_{01}}{n}, \quad n = n_{11} + n_{10} + n_{01} + n_{00}$$

χ^2 计算公式

$$\chi^2 = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(n_{e_t e_c} - e_{e_t e_c})^2}{e_{e_t e_c}}$$

在上例中，通过计算可得：

词A与实体词的卡方统计量为：12606.48

词B与实体词的卡方统计量为：6.04

(P41. 5.3.2.2 实体相关词抽取 (2) 卡方方法)

信息融合

处理结果读取

处理结果显示

三、系统实现

信息融合

信息融合

在两个社交媒体中补全该个体所提到的信息：

1. 在两个社交媒体中提到的相同实体的不同相关词；
2. 在两个社交媒体中提到的不同实体。

(P42. 5.3.3信息融合)

处理结果
读取

处理结果读取与显示

处理结果
显示

使用框架：flask

(P43. 5.4处理结果显示模块)

四、处理结果

1. 原始文本

User Information Fusion

★ Home

☰ Text

☰ Token

☰ POS

☰ Extractor

☰ Complement

Text

Select Sample User

To protect users' rights, we only provide few number of sample users and never leak the username or other info rather than tweets and status of each peer.

User 2.

✓ Submit

Raw Texts

TwitterFacebook

Tweets	Post time	Flag
@d_lazarin if I'd known it was coming so soon I would have had an answer prepared! They should make picture books to help with this.	2015-05-21	11110
I thought we had more time before this conversation, but my 4yo son just asked The Big Question: "Do people make art for money, or for fun?"	2015-05-21	11110
@SarahEvonne please do!	2015-05-21	11110
@5redpandas 🤔	2015-05-21	11110
@5redpandas tweet the lines to me! It'll be like we are in bed together. You know, in a book friends way. (PS I snore but quietly)	2015-05-21	11110
@5redpandas I'm dying to read it!	2015-05-21	11110
@SarahEvonne @IsaacFitzgerald Me too! It's so complex & gorgeous & GREAT! (And I'm an	2015-05-21	11110

(P44. 5.4.4.1原始社交文本)

四、处理结果

2. 中间结果：文本切词与单词标准化

User Information Fusion

★ Home

Text

Token

POS

Extractor

Complement

Splitied Tokens

Twitter Facebook

Tweet Tokens

@d_lazarin if i'd known it wa coming so soon i would have had an answer prepared ! they should make picture book to help with this .

i thought we had more time before this conversation , but my 4yo son just asked the big question : " do people make art for money , or for fun ? "

@sarahevonne please do !

@5redpandas 🤔

@5redpandas tweet the line to me ! it'll be like we are in bed together . you know , in a book friend way . (ps i snore but quietly)

@5redpandas i'm dying to read it !

@sarahevonne @isaacfitzgerald me too ! it's so complex & gorgeous & great ! (and i'm an old married mom-lady .) hope to see you tues , sarah !

@isaacfitzgerald @sarahevonne thanks for the rts ! <3 <3 <3

park slope stoop wa kind enough to ask me a few question about my new interview series bookish . fb.me/27bx78ysa

next tuesday , 5/26 , at the brooklyn public library-park slope , i'll be chatting with kate bolick about her new ... fb.me/3dxy2bno1

i'm so looking forward to hostina this event ! anv question vou want me to ask @katebolick ?

(P45. 5. 4. 4. 2文本切词与单词标准化)

四、处理结果

2. 中间结果：文本词性标注

User Information Fusion

★ Home

☰ Text

☰ Token

☰ POS

☰ Extractor

☰ Complement

POS Tags

TwitterFacebook

Tweet Tags

@wfsuper it tie all of our school together .

@wfsuper The first one wa hand painted . The second one wa a picture of the painting with a fancy printer .

Nice new addition to the Board room . #MOISD pic.twitter.com/GNe46umS2N

Could Ice Cream Save Michigan's Rocky Roads ? | Oakland Township-Lake Orion , MI Patch patch.com/michigan/oakla ...

Our editorial : Choose Menzel a state's school chief detne.ws/1beqoA6 via @detroitnews

Via @nprnews : Common Core Means 3 Tests In 3 Years For Michigan Kids n.pr/1DIWFM9

Board : Snyder's school reform plan violates constitution on.freep.com/1bdLkqQ via @freep

ed.gov/blog/2015/03/n ...

March is here ! Check out this month's article on the #ReadingNow Network @ReadingNowNet moisd.org/downloads/supe ...

Tom Watkins : The sad state of child in Michigan bcene.ws/1bbHwXp via @bcenquirer

Did Gov . Snyder violate the Michigan Constitution with executive order ? Expert say no s.mlive.com/Qse4y1k via @mlive

Detroit group upset with Snyder's school reform move detne.ws/1CdvaZ0 via @detroitnews

(P45. 5.4.4.3文本词性标注)

四、处理结果

2. 中间结果：文本实体识别与实体相关词抽取

User Information Fusion

★ Home

Text

Token

POS

Extractor

Complement

Entities & Relevance Words

TwitterFacebook

Tweet Entities

Michigan school chief : Tackle poverty to boost learning detne.ws/1ADhNvB via @detroitnews

' Preschool for 3-year-olds - high cost , higher reward ' from @bridgemichigan : bridgemi.com/2015/03/presch ...

How New Autoworkers Became Second-Class Employees bloomberg.com/news/articles/ ...

Zip-lining with the wife for 25th Anniversary in California #stillcrazycouple pic.twitter.com/qoLvnl6VD

Rocking the #ThatMetalShow t-shirt on #CatalinaIsland for @EddieTrunk pic.twitter.com/HKZc3vxkT1

Roads proposal ' ha a lot of pothole , ' say Michigan Attorney General Bill Schuette s.mlive.com/2TaIXIf via @mlive

Michigan governor and lieutenant governor talk \$48 million 3rd-grade reading push s.mlive.com/WVXfO8J via @mlive

Equal funding between school district to be big talking point in debate on education budget s.mlive.com/k8JmY2M via @mlive

Michigan's teacher of the year blast standardized testing , say educator want more input on policy s.mlive.com/UCw8YPI via @mlive

Christianity no longer required for next McBain school chief petoskeynews.com/news/state-reg ... via @petoskeynews

State Board of Education identifies 6 candidate to publicly interview for superintendent job | News - Home clickondetroit.com/news/state-boa ...

Bolger : Why I'm voting yes on Prop 1 detne.ws/1FZVtkn via @detroitnews

(P46. 5. 4. 4. 4信息抽取)

四、处理结果

3. 最终结果：信息融合

- 1) 是否能够表现融合的两个账户属于统一个个体；
- 2) 是否能够补全一个个体在两个社交媒体中的不同信息。

(P20. 3. 6信息融合功能)

四、处理结果

3. 最终结果：信息融合 User1（不同个体的账号）

User Information Fusion

★ Home

☰ Text

☰ Token

☰ POS

☰ Extractor

☰ Complement

Complement

Select Sample User

User 1.

✓ Submit

Information

MatchTwitterFacebook

Match Entities	Sources	Relevance Words
Lincoln	twitterfacebook	CareerCenterTechaLIVE
Extra	twitterfacebook	8LeadershipMakeEmployeesVideo
David	twitterfacebook	NovakacoachCEOMark
Social	twitterfacebook	aExpertNetworkHard
Miami	twitterfacebook	SuperwinOhiohomeMississippiplaybeat

PreviousNext

(P46. 5. 4. 4. 5信息融合)

四、处理结果

3. 最终结果：信息融合 User1（不同个体的账号）

User Information Fusion

★ Home

☰ Text

☰ Token

☰ POS

☰ Extractor

☰ Complement

Select Sample User

User 1.

✓ Submit

To protect users' rights, we only provide few number of sample users and never leak the username or other info rather than tweets and status of each peer.

Information

Match Twitter Facebook

Match Entities	Sources	Relevance Words
Park	twitter facebook	Craft Day BR A State
Paul	twitter facebook	Super show Nice 25 fan give a
Ohio	twitter facebook	service dog odds National store central I found PM !! wa a great u Grad Dayton city time good Delaware " State " columbus wait minute day show luck Jim Lewis Center teacher Middle School school friend gas Columbus Miami Hey 8th fan student @WHSRowe Visit !!!! Powell absolutely watch #Pittsburgh Change lived inch hr coming -- A Black Gold today tonight storm love column twitter
Miller	twitter facebook	free parent Conference Center pm Call William #crazykaz making
Well	twitter facebook	Sports Michigan deserved morning I

Previous

Next

(P46. 5.4.4.5信息融合)

四、处理结果

3. 最终结果：信息融合 User1（不同个体的账号）

User Information Fusion

★ Home

☰ Text

☰ Token

☰ POS

☰ Extractor

☰ Complement

Complement

Select Sample User

User 1.

✓ Submit

Information

Match Twitter Facebook

Match Entities	Sources	Relevance Words
Way	twitter facebook	!!!! win a Greg ?? kid made 🎵 Music
China	twitter facebook	history time follow a family #Westerville kid Check
Ready	twitter facebook	Circa Start flying 20
Gruden	twitter facebook	John wa Lions a division title call
Old	twitter facebook	a great band Kind Hall school " met " student

Previous Next

(P46. 5. 4. 4. 5信息融合)

四、处理结果

3. 最终结果：信息融合 User2（同一个个体的账号）

User Information Fusion

Information

Match Twitter Facebook

Match Entities	Sources	Relevance Words
BOOKISH	twitter facebook	Park Slope wa kind question interview series Brooklyn Public Library excited library chat wait Kate 26 Friends author branch completely -- u happy book join
Brooklyn	twitter facebook	Public Slope Kate coffee shop ?? Talking opening line hate time people u resident street -- tonight teen live Pennsylvania police murder feel Hey today report traffic school doe join great part :) kid folk Green Market wa Magazine event series man Manhattan rock Library understand year cc love Year's !!! !! thing moved Matthew coach Read Black Balloon Book Festival beautiful Park interview 6pm author book Tough Questions hair Books Friends happy wait chat coming field trip Torah Animal Street museum house animal Hold public amazing King Jr. MLK
Kate	twitter facebook	Public Slope Happy book birthday good Congrats !! heck excited talk chat Magazine happen Park Library interview series wait 26 6pm author Tough Questions people hair Books Friends happy join u
I'm	twitter facebook	Teresa bed sleeping missed World thing dream Yeah pant good Simple book deal DR lot MAN Good hate read white making feel doe ?? feeling middle thinking luck radio talking Um time k hope Days genius Genius !! celebrate day school talk people today's buddy PE today Years give 5 red idea event Bar guy top :) man Brooklyn street hell tweet 4 daughter surgery close Ben Marcus important game Starts crazy year wa writing Cantor flipped picture hard life James awww mom Life

(P46. 5. 4. 4. 5信息融合)

四、处理结果

3. 最终结果：信息融合 User2（同一个个体的账号）

User Information Fusion

★ Home

☰ Text

☰ Token

☰ POS

☰ Extractor

☰ Complement

Select Sample User

To protect users' rights, we only provide few number of sample users and never leak the username or other info rather than tweets and status of each peer.

User 2.

✓ Submit

Information

MatchTwitterFacebook

Match Entities	Sources	Relevance Words
Melanie	twitterfacebook	Harvey Hope Ted read book art Adam lot nice Lit week Kids great Day March Melissa run pop-up
Betsy	twitterfacebook	Harvey Melanie Hope Ted read book art Adam lot nice Lit week Kids wa photo WORD Greenberg great Day March Melissa run pop-up
Ted	twitterfacebook	Harvey Melanie Hope read book art great Kids Day guy make kid picture project WORD pop-up
Lit	twitterfacebook	Great series Time Back Conference Check lineup talking join Magazine great Brooklyn excited cafe 16th TV Night funny David Gordon Nicole Sara !! true Park author Amy Avenue E 10th Street Road —
Thanks	twitterfacebook	great Twitter wait !! Year's roundup upcoming NYC folk Park Slope Brooklyn Public Library series BOOKISH

PreviousNext

(P46. 5.4.4.5信息融合)

四、处理结果

3. 最终结果：信息融合 User2（同一个个体的账号）

User Information Fusion

★ Home

☰ Text

☰ Token

☰ POS

☰ Extractor

☰ Complement

Complement

Select Sample User

User 2.

✓ Submit

Information

MatchTwitterFacebook

Twitter Entities	Sources	Relevance Words
@PeninaRoth	twitter	
Loneliness	twitter	feeling shit center totally !! Man
@Peter_v_Aguero	twitter	life
Plath	twitter	s version Happy feel Sylvia
Sylvia	twitter	feel wild pink listening :)

PreviousNext

(P46. 5. 4. 4. 5信息融合)

结束

致 谢
THE END