

清 华 大 学

综 合 论 文 训 练

题目：社交数据中的活动挖掘

系 别：计算机科学与技术系

专 业：计算机科学与技术

姓 名：王凝枰

指导教师：唐杰副教授

2014 年 6 月 22 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名：王凝桦 导师签名：张 艺 日 期：2014年6月17日

中文摘要

随着社交媒体和移动互联网日益渗透用户的日常生活，用户越来越多的在社交媒体上发布和自己的日常生活相关的内容。社交数据中对用户日常活动信息的挖掘，可以在用户行为建模，个性化推荐等领域有着诸多潜在应用。然而，活动挖掘这个领域并没有得到足够的重视，这方面仅由很少量的工作。本文基于微博这一流行的社交网络，研究了在社交数据中进行活动挖掘的算法和框架。

本文将活动信息挖掘分解为多个子问题，首先从微博文本中抽取活动概念；其次，抽取活动相关属性，包括时间、地点、情感极性等，构成活动实例。基于概念抽取和实例抽取的结果，对活动的相关度和序列关系进行挖掘。

此外，本文构建了关于日常活动的知识库系统 ActivityNet，对于用户的查询请求，通过地图、图表等形式，将上述工作得到的活动相关知识进行直观的展示。

关键词：活动挖掘；自然语言处理；信息抽取；社交媒体；知识库

ABSTRACT

As social media and mobile network gain popularity in people's everyday life, people are posting more and more activity-related content in social media. Activity mining in social data offers tons of potential applications, such as user behavior modeling and personalized recommendation. However, this area has not attracted much attention so far. Based on Weibo, one of the most popular social network in China, the thesis studies the algorithms and frameworks of activity mining in social data.

This work treats activity mining as a series of sub-problems. First, extracting abstract activity concept in social content. Then, extracting activity attributes including place, time, sentiment polarity and building activity instances. Based on concept extraction and instance extraction result, mining similarity and sequential relations between activities.

Besides, a knowledge-base system Activity is built to demonstrate our work. In response to user's query, a visualized representation of the knowledge about the activity is shown, using maps and charts.

Keywords: Activity Mining; Natural Language Processing; Information Extraction; Social Media; Knowledge Base

目 录

第 1 章 绪论	1
1.1 研究背景和意义	1
1.2 研究问题与挑战	3
1.2.1 研究问题	3
1.2.2 挑战	4
1.3 论文组织	5
第 2 章 相关工作	6
2.1 信息抽取	6
2.2 本体学习和现有知识库系统概况	8
2.2.1 本体学习	8
2.2.2 现有知识库系统概况	9
2.3 情感分析	10
2.4 活动挖掘	12
第 3 章 活动概念抽取	14
3.1 概述	14
3.2 数据来源	14
3.3 分词与词性标注	15
3.4 概念候选集抽取	16
3.5 训练语义特征向量表示	18
3.5.1 背景和目标	18
3.5.2 神经网络语言模型 (NNLM)	19
3.5.3 案例研究	20
3.6 训练集选取	22
3.6.1 动机与目标	22
3.6.2 NP-Hardness 证明	23

3.6.3 子模性及近似求解	25
3.7 模型训练与实验结果	27
3.8 本章小结	28
第 4 章 实例抽取与关系挖掘	29
4.1 目标	29
4.2 活动类别抽取	29
4.3 情感极性分析	31
4.3.1 方法概述	31
4.3.2 实验结果与分析	32
4.3.3 进一步工作	33
4.4 地点、时间抽取	34
4.5 序列关系挖掘	35
4.6 本章小结	36
第 5 章 ActivityNet 系统设计	38
5.1 设计目标	38
5.2 底层架构	38
5.3 功能实现	39
5.3.1 首页设计	39
5.3.2 活动搜索	40
5.3.3 地点、情感、时间分布	41
5.4 本章小结	42
第 6 章 结论	43
6.1 工作总结	43
6.2 进一步工作	43
插图索引	45
表格索引	46
参考文献	47

致 谢	49
声 明	50
附录 A 外文资料的调研阅读报告	51
A.1 Opinion Mining	51
A.2 Evolutionary Topic Models	53
A.3 Event Tracking and Burstiness detection	54
参考文献	55
在学期间参加课题的研究成果	57

第 1 章 绪论

1.1 研究背景和意义

在过去几年中，Facebook，Twitter，以及国内新浪微博、人人网为代表的社交网络迅速崛起。Facebook 全球已经有超过 10 亿用户，新浪微博全球注册用户也已经达到 5.6 亿。据报道，在美国，人们 16% 的上网时间停留在 Facebook 上，超过了传统搜索引擎 Google 的 10%。社交媒体中保存了大量的用户资料，用户间的社交关系以及海量文本信息，这些社交数据有着巨大的研究价值，在广告和推荐系统方面，也有着广阔的应用前景。

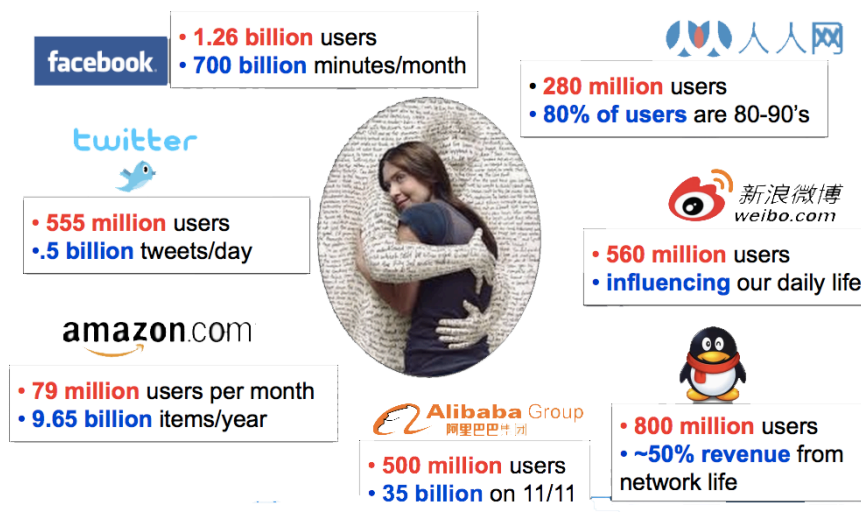


图 1.1 社交网络的崛起

社交网络为用户提供了一个信息传播，结识好友的社交平台，用户发布的信息最初也通常是交流、沟通以及对社会事件的看法。现有的社交网络研究，如网络结构和演化分析^[1]，链接预测^[2]，影响力分析^[3]，社会纽带关系推断^[4]等，更多地考虑社交网络本身的特性和用户的线上行为。而随着移动的互联网的迅猛发展，除了传统的个人计算机以外，用户越来越多地通过移动终端访问社交网络。同时，社交网络也深刻地改变了用户的社会圈，也影响了用户的行为。朋友

聚会，家庭出游等活动中，经常许多人到达目的地的第一件事是拿出手机，在微博或 Foursquare 等社交网络上签到，告诉朋友们自己现在在做什么，心情如何，社交网络和线下的日常生活以前所未有的方式深入地结合在一起。

用户为什么会在社交网络中发布自己日常活动的信息？在社会学中，马克思·韦伯提出的社会行动理论 (Social Action Theory)^[5] 指出，人的行为是社会性的，需要考虑到社交圈中其他人的行为以及他产生的结果。他将人的行动分为三类

- 理性行动 (Rational action)。它可以达成有价值的目标，但人们并没有充分考虑行为的结果，以及达成目标的方式。
- 工具行动 (Instrumental action)。人在充分考虑它与目标的关系，以及达成目标的种种方式后，采取的动作。
- 情感行动 (Affectional action)。人由于表达感情的需要，进行的行动。

用户的这种行为，并没有明确的目的性，可以归因于用户的情感需求，即第三类行动。人们希望表达自己参加活动的心情，渴望被他人了解。这也促使我们关注这样一个问题，是否可以从社交媒体的海量数据中挖掘出人们日常活动中的一些知识，比如一项活动通常在什么时间、什么地点发生，人们在参加这项活动时，心情通常是怎样的；进一步地，两个活动是否存在联系，用户在进行 A 活动后，最可能去做的下一件事是什么。最终，将我们得到的数据进行整理、分析，构建一个关于人类日常活动的知识库，将很有意义。

相比于其他的 Web 数据，如点评、论坛、百科、问答平台，社交媒体有不可替代的优势。大众点评、Yelp 等点评类网站，同样可以挖掘用户的活动信息，但它以商家为中心，局限于用户消费行为，如就餐、住宿、电影，收集的只是用户对于商家的点评信息。而用户的日常活动，如散步、睡觉等等，用户是不会发布在这些平台上的。用户在类似微博、人人网的社交媒体上，更愿意发布私人化、生活化的内容；同时，得益于移动互联网，用户通过随时随地方便地访问社交平台，获取信息的实时性更好，也更容易获取用户精确的地理信息。同时，社交网络的好友机制，可以帮助我们推断用户在活动中的互动。例如，用户 u_i 在时刻 t 发布了一条活动信息的微博并提到了用户 u_j ，或者 B 在相近的地点进行了签到，我们可以推测， u_j 参加了 u_i 提到的那一项活动，即使 B 并没有特意发微博表明这一事实。

对社交网络中活动信息的挖掘，可以预想，有许多潜在的应用，尤其是在

推荐系统的应用:

1. **用户行为模式建模** 给定用户 t_0 时刻前的活动历史 $history = \{a_i | t < t_0\}$, 预测用户在下一时刻 t_0+1 , 在地点 p 进行活动 c 的概率, 即 $P(p_i, a_j | t+1, history)$ 。
2. **基于兴趣的好友推荐** 现有的社交平台, 如 Facebook, 人人网等的好友推荐, 主要对用户是否认识进行链接预测; 而微博的推荐关注, 则考虑到被推荐用户的影响力, 以及和用户选定兴趣标签的匹配程度。如果我们能对用户参加活动的类别和地点有足够多的了解, 就能更精确地向用户推荐兴趣相投的朋友, 并且更容易转化为线下的好友关系。
3. **个性化的活动推荐** 当用户来到一个陌生的城市, 或是有闲暇的时间而不知道如何休闲娱乐, 我们可以基于活动的知识, 向其推荐活动。比如, 一个用户经常在微博中发布吃火锅类似的信息, 那么, 如果用户来北京旅游, 就可以向其推荐东来顺等有特色的涮肉火锅, 这也会产生潜在的商业价值。

1.2 研究问题与挑战

1.2.1 研究问题

在提出研究问题之前, 我们首先对活动 (Activity) 进行定义。我们在两个层面上研究日常活动。其中, 活动的概念是抽象意义上的活动类别, 定义为:

定义 1.1 (活动概念 (Activity Concept)): 活动的概念 c 是表示人类日常活动的动词短语, 即二元组 $\langle \text{动作 (action)}, \text{目标 (object)} \rangle$ 。其中, 动作为一动词, 目标为一名词。对于某些活动, 目标可以为空, 以 O 表示。例如, 以下均为合法的活动。

- $\langle \text{吃, 烤鸭} \rangle$
- $\langle \text{参加, 会议} \rangle$
- $\langle \text{游泳, } O \rangle$
- $\langle \text{旅游, } O \rangle$

而活动的实例是活动概念的具体化, 除了其对应的活动概念外, 还具有相关属性, 定义为

定义 1.2 (活动实例 (Activity Instance)): 活动实例为五元组 $a = \langle c, u, s, p, t \rangle$, 其中 c 表示活动的概念, u 为活动的执行者, p 为发生的地点, s 为用户的情感倾向, t 为发生的时间。

在研究中, 我们将社交数据中的活动挖掘分解为若干个子任务, 如下

问题 1.1 (活动抽取与关系挖掘): 输入给定社交数据, 包括: 用户集合 $U = \{u_i\}, i = 1, 2, \dots, |U|$; 对于每个用户 $u_i \in U$, 其发布的微博集合 $M_i = \{m_{ij}\}, j = 1, 2, \dots, |M_i|$, 解决以下子任务, 并输出:

1. **概念抽取:** 活动概念集合 $C = \{c_i\}, i = 1, 2, \dots, |C|$ 。
2. **实例抽取:** 对于微博 m_{ij} , 若其包含一项活动, 抽取其地点 p 、时间 t 、情感极性 s , 构建活动实例 a 。
3. **活动关系挖掘:** 本文关注活动之间时间上的序列关系, 即对于活动概念 c_i 和 c_j , $c_j \text{ follows } c_i$ 意味着 c_j 经常作为 c_i 的下一个活动发生。本文目标在于挖掘出所有活动间的序列关系及其强度。

基于以上研究, 本文希望构建一个平台 ActivityNet, 对研究结果进行可视化:

1. 响应用户的查询, 呈现该活动的相关知识, 如活动的类别, 地域、时间分布, 类似活动等等。
2. 对于特定的社交网络用户 (本文工作主要基于新浪微博), 分析其活动偏好。
3. 对于给定的地点, 推荐当地最热门的活动。

1.2.2 挑战

虽然社交数据相比其他 Web 数据有诸多优势, 其本身的特点也给我们的工作提出了许多挑战:

1. 问答、论坛等 Web 平台, 其自身的分类系统限定了话题, 人们谈论的内容比较单一。而社交媒体中, 人们谈论的话题不收限制, 可以对自己感兴趣的任何问题发表观点, 不仅仅局限在日常活动, 不可避免了噪声和稀疏性问题。同时, 人们在社交媒体上发布的消息格式高度自由, 长度很短, 同时常带有俚语和语法、词法的错误。这对活动信息的抽取带来挑战。

2. 社交媒体中对消息长度有严格限制，传统的基于统计的方法不易应用于挖掘活动之间的概念联系。
3. 用户参加某项活动的信息可能并不是显式表达的，并且可能有信息缺失。如何利用社交网络的结构信息，对未明确表达的参与关系进行预测和推断。

在本文中，主要解决前两个挑战。利用社交网络结构对的活动参与信息进行推断有比较大的难度，在将来的工作中会进一步研究。

1.3 论文组织

本文的章节安排如下：

第二章 介绍了与本文工作活动挖掘相关的研究领域和工作，如信息抽取等。

第三章 介绍了基于神经网络语言模型和分类算法的概念抽取模型，并对结果进行了分析。

第四章 介绍了抽取活动相关属性，如情感极性、地点等，构建活动实例的方法。并基于实例抽取结果，进行活动关系挖掘。

第五章 介绍了本文研究中构建的系统 ActivityNet，从系统的基础架构和可视化方面对该系统进行详细的介绍。

第六章 总结本文的工作和不足之处，并提出进一步的研究方向。

第 2 章 相关工作

本文实验的系统，与信息抽取、本体学习等领域有较强的相关性，同时用到了自然语言处理、机器学习的一些模型和方法。下面对本文主要的相关工作作简单的介绍。

2.1 信息抽取

信息抽取 (Information Extraction) 是文本挖掘中的一个重要研究领域，它的目标在于从非结构化的文本中抽取特定信息，如人名、组织名、地点等命名实体，或以此为基础的更复杂的抽取任务如事件、关系抽取，并进一步对抽取出的信息进行结构化的组织。

信息抽取中早期的工作通常基于规则^[6]。首先，手工设计或者自动生成一系列规则，文本中的每个符号都被表示为特征的集合。一条规则包含表达式和动作两个部分。例如，一个识别人名的规则如下：

(第一个词：“Mr.”, 第二个词：首字母大写) → 人名

将文本和规则匹配，如果匹配成功，则执行对应的动作。如果有多条规则同时匹配成功，则出现冲突，需要定义规则执行的顺序，例如顺序执行等。

在特定的任务上基于规则的方法可以达到很好的性能，但是设计良好的抽取规则依赖于领域专家大量的工作，费时费力，并且针对一个目标的工作无法应用到其他目标中。因此，研究者之后将统计机器学习应用到信息抽取中，将信息抽取分解为不同的子问题，这些子问题可以转化为分类问题，可以使用支持向量机、最大熵模型等解决。在信息抽取中，除特定目标的文本以外，有时需要辨别抽取结果中不同部分的语义角色，例如词性标注，序列标注的方法可以用于解决这类问题。在序列标注问题中，句子中的每个词作为观察值，用 BIO 表示法标记文本块的边界。BIO 标注法为，对每个实体类型 T，类标 B-T 表示一个类型为 T 的实体名称开始，I-T 表示位于 T 类型实体名称内部，O 表示在任何一个名称外部。序列标注问题即给出一个词序列，求出统计意义下最可能的标注序列，定义如下：

问题 2.1 (序列标注): 已知观察值 $\mathbf{x} = (x_1, x_2, \dots, x_n)$, 求最优序列标注 $\mathbf{y}^* = (y_1, y_2, \dots, y_n)$, 使得 $\mathbf{y}^* = \arg \max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})$ 。

隐马尔科夫模型 (Hidden Markov Model, HMM) 是解决序列标注问题的常用模型。在 HMM, 对序列的生成过程做了马尔科夫假设和输出独立性假设。马尔科夫假设即每个类标 y_i 仅由前一个类标 y_{i-1} 生成; 输出独立性假设即输出观察值之间严格独立。状态的转移概率可以在语料中进行频度估计, 如

$$p(y_i = c | y_{i-1} = c, x_{i-1} = w) = \frac{n(c_1, c_2, w)}{n(c_2, w)}$$

为了避免数据稀疏性的问题, 可以加入必要的光滑。

HMM 是一个生成模型, 但它的两个假设有时并不合理。当数据较充足时, 直接对 $P(\mathbf{y}|\mathbf{x})$ 进行建模的判别模型可能有更低的预测误差。McCallum 等人将最大熵马尔科夫模型 (MEMM) 应用于信息抽取^[7], 修正了观察值之间严格独立的假设。在 MEMM 中, 同样有马尔科夫假设, 类标 y_i 依赖于附近的观察值 $x_{i-l}^{i+l} = (x_{i-l}, x_{i-l+1}, \dots, x_{i+l})$ 和之前若干个类标 $y_{i-k}^{i-1} = (y_{i-k}, y_{i-k+1}, \dots, y_{i-1})$:

$$p(\mathbf{y}|\mathbf{x}) = \prod_i p(y_i | y_{i-k}^{i-1}, x_{i-l}^{i+l})$$

其中

$$p(y_i | y_{i-k}^{i-1}, x_{i-l}^{i+l}) = \frac{\exp\left(\sum_j \lambda_j f_j(y_i, y_{i-k}^{i-1}, x_{i-l}^{i+l})\right)}{\sum_{y'} \exp\left(\sum_j \lambda_j f_j(y', y_{i-k}^{i-1}, x_{i-l}^{i+l})\right)} \quad (2-1)$$

$f()$ 是定义在类标和观察值上的特征函数。

MEMM 虽然不需要输出独立性假设, 但由于 MEMM 只在局部做归一化²⁻¹, 它存在标注偏置问题 (label bias problem)。条件随机场 (CRF)^[8] 解决了这一问题。CRF 和 MEMM 的区别有: 1, y_i 不仅和 $y_{j<i}$ 有关, 而且和 $y_{j>i}$ 有关; 2, CRF 是一个无向图模型, 而 MEMM 是有向图模型; 3, CRF 做全局归一化, 而不是局部归一。在一阶线性 CRF 中,

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_i \sum_j \lambda_j f_j(y_i, y_{i-1}, \mathbf{x}, i)\right)$$

Z 为归一化常数,

$$Z = \sum_{\mathbf{y}'} \exp\left(\sum_i \sum_j \lambda_j f_j(y_i, y_{i-1}, \mathbf{x}, i)\right)$$

CRF 是当前使用最广泛的信息抽取模型, 并且在此基础上有许多工作对其进行发展。例如, 一般的 CRF 难以引入长程特征, 因为 CRF 的计算复杂度随着阶数指数增加, 为了解决这一问题, Sarawagi^[9] 提出了 Semi-Markov CRF, 它可以在较低的计算消耗下, 达到和高阶 CRF 相似的预测能力。

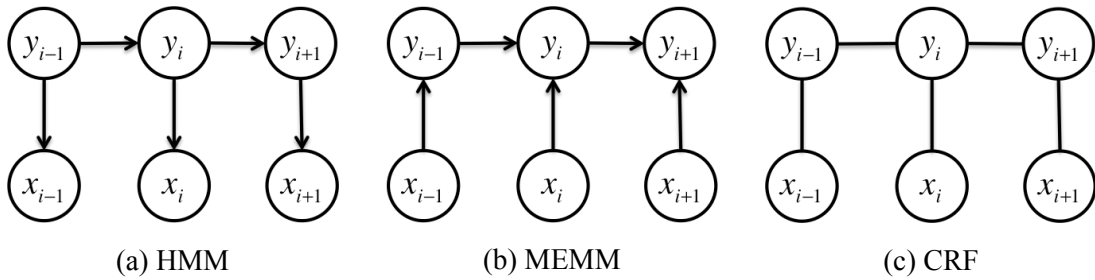


图 2.1 序列标注模型

2.2 本体学习和现有知识库系统概况

2.2.1 本体学习

本体学习 (Ontology Learning) 目标在于自然语言文本中自动地抽取概念, 并构建概念之间的关系的的过程, 它是知识库构建的基础。本体学习包括本体的构建以及本体关系的建模, 其中, 将抽取出的概念依据上下位组织为层次化的分类系统 (hierarchy induction) 是常见的手段。概念层次是对数据不同粒度上的组织, 有许多重要应用, 如基于本体的个性化 Web 搜索和浏览^[10]。对概念进行层次化的建模, 可以使用基于统计或基于词法、语法模式的方法。当前大多数的工作集中于挖掘上下位关系, 即 is-a 关系。Snow 等人^[11] 提出了一个上下位关系的学习模型。对于给定的两个词 c_i, c_j , 他首先找到语料中包含这两个词的所有句子, 并进行句法解析得到依赖树, 随后基于语法特征训练分类器, 预测两个词之间是否存在上下位关系。Navigli 等人^[12] 提出了一种新颖的基于图的方法。他首先将不同的概念组织成有向图, 这张图可能是稠密、带环甚至是不连通的, 图的边表示概念的关系, 可以用之前的方法生成。随后从最一般的概念

开始，寻找最优分支，修剪有噪声的边，得到分类层次。

这些传统的层次构建方法均需要一个领域相关的语料。但不足在于，

1. 与目标领域高度相关的语料常常是难以获得的，例如，“计算机科学”相关的语料库很容易获得，但仅和“模式识别”相关的语料就不容易了；并且我们常常关注崭新的、不断变化的领域，相关语料更难以获得。
2. 为了构建概念之间的关系，常常要从中找出一些语言模式。而高质量的模式是稀疏的，尤其是语料不足时。

Liu 等人^[13]提出了一种基于关键词集合的层次构建方法。对于每个关键词，从知识库中获取相关概念，并从搜索引擎中获取上下文信息，然后进行多分支的层次聚类方法 Bayesian Rose Tree (BRT) 进行聚类，得到概念层次。BRT 的构建过程中，首先初始化 $T_i = \{\mathbf{x}_i\}$ ，即每棵树仅包含一个节点，构成一个森林；随后进行迭代，使用贪心策略，从森林中每次选取两棵树 T_i 和 T_j ，将其组合为一棵新树 T_m 。与二叉的层次聚类不同，对树 T_i 和 T_j ，BRT 有三种可能的操作：

- 合并 (Join)。 $T_m = \{T_i, T_j\}$ ，即 T_i, T_j 作为 T_m 的子树。
- 吸收 (Absord)。 $T_m = \{\text{children}(T_i) \cup T_j\}$ ， T_m 有 $|T_i| + 1$ 个子结点。
- 坍缩 (Collapse)。 $T_m = \{\text{children}(T_i) \cup \text{children}(T_j)\}$ ， T_m 有 $|T_i| + |T_j|$ 个子节点。

选取 T_i, T_j 和合适的操作，使得最大化

$$\frac{p(D_m|T_m)}{p(D_i|T_i)p(D_j|T_j)}$$

其中 $p(D_m|T_m)$ 是 T_m 下所有叶节点数据点， $D_m = \{D_i \cup D_j\}$ 。 $p(D_m|T_m)$ 递归定义为

$$p(D_m|T_m) = \pi_{T_m} f(D_m) + (1 - \pi_{T_m}) \prod_{T_i \in \text{Children}(T_m)} p(D_i|T_i)$$

2.2.2 现有知识库系统概况

知识库是对人类知识的结构化表示，它在搜索、智能系统中有着日益重要的应用，Google, Microsoft 以及国内百度、搜狗等互联网企业均有自己的知识图谱计划。本文的工作目标就在于构建一个关于人类日常活动的知识库。表2.1对研究工作中常见的知识库的概况进行了总结。

表 2.1 现有知识库概况

名称	开发者	概念数量	isA 关系数量	概述
SenticNet	南洋理工大学、新加坡国立大学	14,244	NA	借助图挖掘和数据降维，提供了概念级的情感分析资源。
Freebase	社区	1450	24,483,434	常用的开放知识库系统。对不同领域的知名人物、地点、事物根据 Topic 组织归类，提供搜索和 API 查询
WordNet ^[14]	普林斯顿大学	25,229	283,070	英语词汇的知识库，根据同义语将英语词汇进行组织，并且提供词汇之间的多种语义关系。
WikiTaxonomy ^[15]	HITS	127,325	105,418	基于维基百科 (Wikipedia) 的语料，将类别按照 is-A 关系构建为一个大规模分类系统
Probase	Microsoft	2,653,872	20,757,545	借助数十亿的网页和多年的搜索记录，构建了比传统知识库庞大许多的概念。并且对不确定性进行建模，以此为基础实现概率推理的功能。

从表2.1中可以看到，现有的知识库种类繁多，其中 **Probase** 在系统规模和功能上达到了很高的水准，但现有的知识库系统主要关注客观实体如人物、地点、物体及其关系建模，没有面向人类日常活动的系统。希望我的工作能在这方面对现有的知识库的一个补充。

2.3 情感分析

本文的工作中，需要对用户参与一项活动的情感状态进行分析。情感分析 (Sentiment Analysis) 和观点挖掘 (Opinion Mining) 是信息检索的一个重要研究领

域，在过去十多年中，学界已有许多工作，也有许多商业公司提供观点挖掘的服务。一个完整的观点挖掘任务的目标在于，给定带有观点的文档集 D ，发现其中包含的所有观点，每个观点可用五元组 $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ 表示，其中

- e_i : 观点相关的实体名称
- a_{ij} : e_i 的一个方面 (aspect)
- oo_{ijkl} : 观点的情感极性，如正面，负面，中性，或者一系列情感强度
- h_k : 观点的持有者
- t_l : 观点的发表时间

情感分析可以在不同的粒度进行。文档级的情感分类假定文档 d 仅对一个实体 e 发表观点，且观点仅来自一个用户 h ，仅判断其情感极性，它是情感分析工作的基础。它是一个分类问题，现有的有监督学习方法均可应用在情感分类中。Bo Pang 等人^[16] 使用 unigram 作为特征对文档进行二分类，发现朴素贝叶斯和 SVM 均有很好的表现。随后的工作^[17] 利用了更多的特征，如

- 词组及词频
- 词性标签。形容词是观点的重要标志，因此作为单独的特征
- 观点词。收集人们经常使用的表达观点的词，如动词“喜欢”，“讨厌”，形容词“不错”，“差”作为特征，对于提高精度有很大帮助。
- 否定词。句子中的否定词可能会改变情感极性，因此也作为特征。

非监督方法也可应用于情感分类。Turney 等人^[18] 提出了一个非监督的框架。首先抽取出包含形容词的短语集合 $\{p_i\}$ ，对每个短语 p_i ，计算与已知情感极性的常见词的 PMI，作为情感倾向

$$SO(p_i) = PMI(p_i, \text{“excellent”}) - PMI(p_i, \text{“poor”})$$

最后，对于给定的文档，计算所有形容词短语的平均情感倾向得分 SO ，若为正，则归类为正面，否则为负面。

尽管从文档级别对情感进行分类大多数情况下是有用的，但在很多应用中，它没有提供足够的细节。一个关于特定实体 e 的正面文档并不意味着作者对 e 的所有方面都有正面的评价。因此，需要进行更细粒度的基于方面的情感分析 (Aspect-based Sentiment Analysis)，它可以分为两个子任务：

- 方面抽取
- 情感分类

两个子任务可以依次进行，分别利用信息抽取相关工具和文档/句级情感分类工具完成相关工作。Topic Modeling 的方法也被尝试用于非监督的观点抽取，许多工作在 LDA 的基础上进行扩展，增加表示情感和方面的隐层，同时对 Aspect 和 Opinion 词进行建模，Wayne Zhao 等人提出的 MaxEnt-LDA 是其中比较有代表性的工作^[19]。如图2.2, 文档中的每个词 $w_{d,s,n}$ 由多个 Mixture 产生: 背景词 Φ^B , 全局 Aspect $\Phi^{A,g}$, 全局 Opinion $\Phi^{O,g}$, 某个特定 Aspect A_t , 某个特定 Opinion O_t , 并由 Bernoulli 变量 $y_{d,s,n}$ 和 $u_{d,s,n}$ 抽样得到由哪一个 Mixture 产生。他们的工作借助了语法特征，帮助将 Aspect 词和 Opinion 词区分开，解决了此前工作的问题^[20]。

本文的工作中，将进行文档级的情感分析，判断活动实例的情感极性。

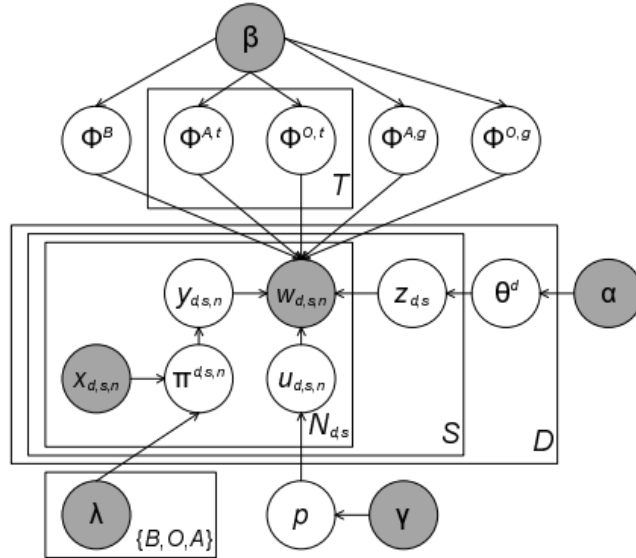


图 2.2 MaxEnt-LDA

2.4 活动挖掘

现阶段，对社交数据中的用户活动进行挖掘的工作不多，可以找到的工作有 Nguyen 等人的工作以及 Park 等人的工作。Nguyen^{[21][22]} 提出了使用了自监督条件随机场 (self-supervised CRF) 从博客中挖掘用户活动的系统，目标在于抽取活动的基本属性：参与者 (actor)，动作 (action)，对象 (object)，时间 (time)，地点 (location)。Nguyen 工作的特点在于，借助了语法正确，易于解析的维基百

科中人物类目的语料，通过语法模板进行自动标注，作为训练数据，训练 CRF 抽取模型。这个系统包含两个模块

1. 自监督学习模块
2. 活动抽取模块

如图2.3所示

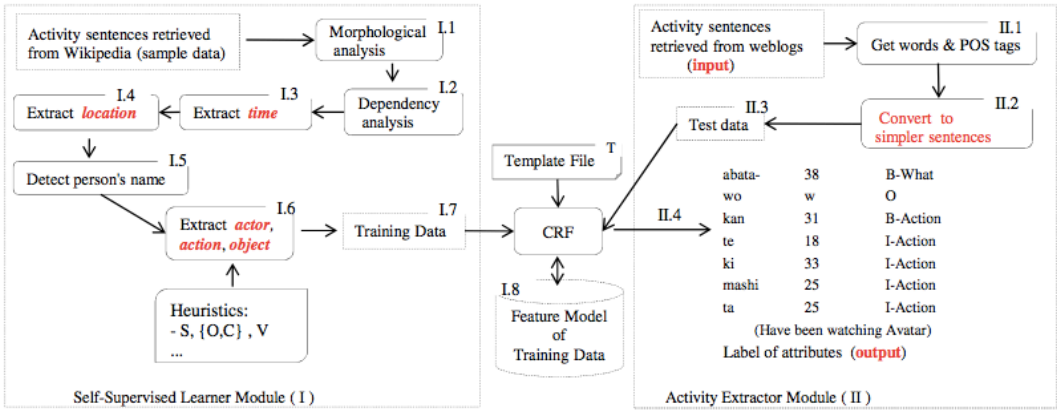


图 2.3 Nguyen 的活动抽取框架

主要步骤如下

- 对维基百科的语料进行分词、词性标注和句法解析，得到词性标签和短语依赖关系, 如动词短语 (VP), 名词短语 (NP), 命名实体。
- 使用 Google Map API 进行地点标注。
- 手动设计一系列的语法模板，如 “S, {O,C}, V”, “{O, C}, V, S”, 其中 S 表示主语，O 表示对象，V 表示动作，C 表示补语。根据这些语法模板，可以获得包含活动的句子，并标注活动的参与者，动作和对象。
- 使用以上步骤得到的语料，训练 CRF，获得一系列特征函数。
- 在活动抽取模块中，使用训练的到的 CRF 模型解析，获得抽取结果。

但是 Nguyen 的工作关注的是博客 (Weblog) 内容的抽取，博客和微博的形式有比较大的区别，一方面，博客长度没有限制，一般是对一个活动的完整叙述，信息完整；另一方面，博客的写作也更加严谨，语法结构比较正式。此外，这个系统仅仅关注信息的抽取，并没有对活动进行更进一步的挖掘。因此，他的工作难以直接应用于本文的工作。本文从短语着手，并充分利用了微博的元信息，克服了这些问题。

第 3 章 活动概念抽取

3.1 概述

本章主要解决如何从社交数据中抽取抽象的活动概念。概念抽取是一个典型的信息抽取问题，这类问题可以通过 CRF 解决。但 CRF 一方面需要比较繁重的标注工作，另一方面，它常常有精度较高而召回率较低，因此本文采取了不同的策略，将其转化为一个二分类问题。

本文首先从全部微博中抽取出较频繁的短语，包括一元和二元语法，并提取短语特征；其次，根据数据分布选取并标注训练集，训练分类模型，对其余数据进行预测。其中，在特征提取中，为了充分利用词语之间的语义关系，本文使用神经网络语言模型工具 Word2Vec 获取了语义向量表示；在训练集选取中，提出并解决了一个组合优化问题，使得训练集能够更好地代表数据集合，在同等标注代价下获得更好的分类效果。概念抽取的框架如图3.1

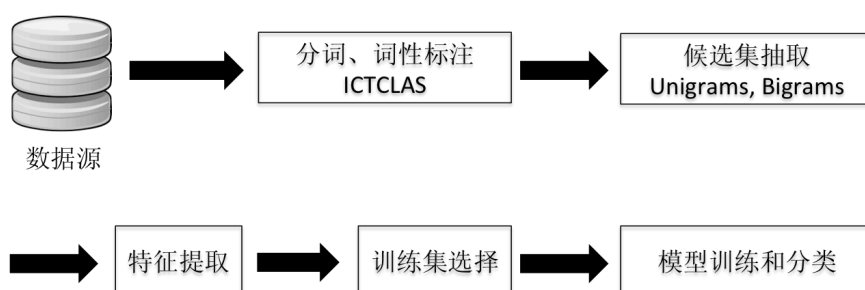


图 3.1 概念抽取框架

3.2 数据来源

本文的工作基于新浪微博的数据。我们随机选取了 100 个用户作为种子，并获取他们的关注者和关注他们的人，以此为核心进行广度优先搜索。对每个用户，抓取其最近的 1000 条微博，包括用户发布和转发的微博。对于一条微博，

可抓取到发表时间、评论数、地理信息等元数据。本文还抓取了用户的个人档案，包含姓名，性别等。表3.1是微博数据的一些信息：

表 3.1 微博数据概况

用户数	1,787, 443
微博数	约 10 亿
关注关系	40 亿
用户信息	名字、性别、关注者，被关注者，创建时间
微博信息	发表时间、评论数、转发数、点赞数、原始微博、地理信息

3.3 分词与词性标注

由于中文没有天然的词边界，因此在中文文本处理中，通常要先进行分词和词性标注。当前常用的中文分词工具如中国科学院开发的 ICTCLAS^①，Stanford Word Segmenter^② 和 Pos Tagger^③，微软研究院的 S-MSRSeg^④ 等。在选择分词工具时，着重关注以下方面：

- 分词和词性标注精度。
- 处理速度。本文要处理千万量级的微博数据，因此分词速度非常重要。
- 扩展性和新词处理能力。微博数据中常常会有新词和惯用语，一般的分词工具难以处理，需要人工添加用户词典，并具有新词发现能力。

综合比较之后，我选择了 ICTCLAS 作为分词工具。

ICTCLAS 接收原始文本作为输入，分词结果中单词以空格隔开。同时，ICTCLAS 提供了细致的词性标注功能。对于每个单词，除主词性外，还有二级标注，如动词 (v) 可进一步分为不及物动词 (vi)、名词性动词 (vn) 等。ICTCLAS 进行分词的结果如表3.2：

此外，对于一些常用的新词和惯用语，本文其加入用户词典中提高分词精度。

① <http://ictclas.nlpir.org/>

② <http://nlp.stanford.edu/software/segmenter.shtml>

③ <http://nlp.stanford.edu/software/tagger.shtml>

④ <http://research.microsoft.com/en-us/downloads/7a2bb7ee-35e6-40d7-a3f1-0b743a56b424/default.aspx>

表 3.2 ICTCLAS 分词结果示例

序号	分词结果
1	等/v 人/n 永远/d 是/vshi 那么/rz 的/ude1 无聊/a 我/rr 在/p :/wp http://t.cn/zj2nzBu/url
2	晚宴/n 要/v 开始/v 咯/y , /wd 金色/n 礼服/n 穿/v 起来/vf
3	酒足饭饱/vl 好/a 开心/a http://t.cn/Sa37Bk/url

3.4 概念候选集抽取

我们首先给出 n 元语法的定义:

定义 3.1 (n 元语法 (n -gram)): 设所有不同的词组成集合 V , n 元语法表示文本中相邻的 n 的词的序列, 即 $\langle w_1, w_2, \dots, w_n \rangle, w_i \in V, i = 1, 2, \dots, n$ 。特别地, 一元语法 (**unigram**) 为单个单词 w , 二元语法为相邻的两个单词 $\langle w_1, w_2 \rangle$ 。

根据我们对活动概念的定义1.1, 活动概念是单独的动词或一动宾结构的短语, 因此, 活动概念为 **unigram** w , w 为动词, 或 **bigram** $\langle w_1, w_2 \rangle$, w_1 为动词, w_2 为名词。为此, 可以遍历分词文本中所有的 **unigram** 和 **bigram**, 同时为了去除噪声, 只保留出现频度大于一个阈值的短语。但这种方法有一些问题

- 受停用词影响大。同英语语言一样, 汉语中也存在一些停用词, 如“是”, “在”, 一些停用词在进行词性标注时被标注为动词。另一些虽然不是传统意义上的停用词, 一般不作为一个动作, 而是副词, 如“完”, “过”但经常出现在动词短语中, 如“做完作业”, “吃过午饭”等。这些词常常词频很高, 它们组成的短语频度很高。
- 表示活动的短语在文本中可能不是连续的。除了上面举过的例子之外, 动词和名词之间可能包含一些更复杂的修饰成分, 如“买了一件衣服”, “参加学校举办的比赛”。对于这些情况, 简单抽取 **bigram** 就会产生遗漏。
- 倾向于选取包含高频词的短语, 易受噪声影响。一个 **bigram**, 不一定构成一个常用短语, 有可能仅仅是随机出现在一起。虽然我们用频度滤除一部分噪声, 但高频词会更多地和其他词随机共同出现, 依据频度的过滤方法会倾向于选择高频词的词组, 无法将这种情况区分出来。

对于问题一, 我们使用一个中文停用词词典, 包含 1220 个停用词, 对抽取结果进行过滤。对于问题二, 我们的目标在于抽取常见的动词短语, 而并不关注句法的精确解析, 因此不需要很高的召回率, 我们根据词性标签和词本身设计一些抽取规则, 可以处理简单的包含数词、量词的情况。

表 3.3 χ^2 检验

	w_j	\bar{w}_j	sum
w_i	a	b	$a + b$
\bar{w}_i	c	d	$c + d$
sum	$a + c$	$b + d$	$n = a + b + c + d$

对于问题三，要识别一个 bigram 是否构成短语，一方面，它应该出现足够的次数，否则，它更有可能是随机噪声；另一方面，如果我们把词 w_i 是否出现作为随机变量 x_i ，那么，如果 $\langle w_i, w_j \rangle$ 构成短语，那么 x_i, x_j 应该有比较强的相关性。相关性检验可以借助统计中的 χ^2 检验。令 $a = n(\langle w_i, w_j \rangle), b = n(\langle w_i, \bar{w}_j \rangle), c = n(\langle \bar{w}_i, w_j \rangle), d = n(\langle \bar{w}_i, \bar{w}_j \rangle), n = a + b + c + d$ ，如表3.3, χ^2 计算如下：

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} \quad (3-1)$$

据此，可得到短语抽取算法1

算法 1: Phrase Extraction

Input: 已分词的句集 $S = \{s_i\}$ ，规则集合 $R = r_i$ ，频度阈值 λ ，结果集大小 K

Output: 短语集合 $P = \{p_i\}$

foreach $s \in S$ **do**

 应用每条规则 $r_i \in R$ ，获得短语列表 $P_s = [\langle w_i, w_j \rangle]$;

foreach $p = \langle w_i, w_j \rangle \in P_s$ **do**

 更新 $n(\langle w_i, w_j \rangle), n(w_i), n(w_j)$;

令 $P_{raw} = \{\langle w_i, w_j \rangle \mid n(\langle w_i, w_j \rangle) > \lambda\}$

foreach $p \in P_{raw}$ **do**

 依照式3-1, 计算 $\chi^2(p)$

以 $\chi^2(p)$ 为键值，以递减序排序 P_{raw} 得到 P_{sorted}

返回 P_{sorted} 的前 K 个元素

最终，我们得到了 98520 条动词短语，作为活动概念的候选集。

3.5 训练语义特征向量表示

3.5.1 背景和目标

在第3.4节中，我们得到了一系列动词短语，作为活动概念的候选集合，但其中大多数并不表示一个日常活动，需要将活动概念加以区分。这是一个分类问题，可以使用现有的分类模型，关键在于特征的选取。

一种方法是，以 **unigram** 作为特征。设共有 $|V|$ 个不同的词，则 **bigram** $\langle w_i, w_j \rangle$ 的特征向量是一个 $|V|$ 维向量 $(0_1, 0_2, \dots, 1_i, 0_i + 1, \dots, 1_j, 0_j + 1, \dots, 0_{|V|})$ 。这样做有效的原因是，不同的短语可能包含相同的动词或名词，其中带有某些动词的短语，吃、喝，更有可能表示一个活动概念。如果我们进行足够的标注，覆盖足够多的情况，就可以对未知的短语进行分类。这种特征表示方法在计算语言学中称为 **One-hot Representation**。

这样做的问题有

- 无法处理单个单词。**One-hot Representation** 仅当不同短语间存在共同的单词才会有效。对于单个单词，除非已进行标注，否则无法获得任何信息。
- 没有考虑不同词语的语义联系。假如两个词之间语义相似，那么他们的类标有更大的可能相同，而 **One-hot Representation** 中，不同单词是独立的，因此需要进行大量的标注，以覆盖尽可能多的样本。

在实际测试中，以 **SVM** 作为分类模型，可以达到 82% 的分类精度，但召回率仅有 29%。为了避免大量的标注工作并提高召回率，我们需要寻找其他特征表示方法，以充分利用两个词之间的语义联系。

对语义相似度的计算，已经有许多方法。例如，基于文本片段中两个词的共现概率^[23]；基于搜索引擎查询结果条目数量^[24]；将词表示为维基百科中条目的分布并计算余弦相似度^[25]。这些方法提供了两个词之间相似度的度量，但使用并不方便。理想的情况是将每个词表示为 K 维空间的向量， K 远远小于不同词的总数，词之间的相似度可以计算两个向量的余弦相似度。这 Hinton 提出的 **Distributed Representation**^[26]。本文使用基于神经网络语言模型的工具 **Word2Vec** 获得词的语义向量表示。

3.5.2 神经网络语言模型 (NNLM)

我们在此对神经网络语言模型及词的向量表示做简要介绍^{[27][28]}。要对语言进行建模，容易想到的思路是，假如两个词经常出现的上下文 (context) 相似，那么它们的语义可能也是相似的，这是自然语言处理中统计语言模型的基础。一个语言模型是语言基本单位（如句子）的生成模型，用各个单词出现的条件概率来表示，即

$$p(\text{sentence}) = \prod_t p(w_t | \text{Context}_t)$$

上下文的选取不同，语言模型也随之变化，其中 **n-gram** 语言模型是其中常用的一种。它对语言的生成做了 $n-1$ 阶 Markov 假设，一个词的出现概率仅和前 $n-1$ 个词相关，即

$$\begin{aligned} \text{Context}_t &= w_{t-n+1}^{t-1} \\ &= (w_{t-n+1}, w_{t-n+2}, \dots, w_{t-1}) \\ p(\text{sentence}) &= \prod_t p(w_t | w_{t-n+1}^{t-1}) \end{aligned}$$

n-gram 语言模型的问题是，由于语料的限制，语言模型受制于数据的稀疏性，高阶语言模型难以训练，无法建模更远的关系，三元语法 (tri-gram) 是研究中经常使用的模型。随着互联网带来的海量数据以及计算能力的提升，更高阶的语言模型成为可能，Google 曾经公开了 5-gram 的语言模型，但体积非常大，对我们的应用来说不切实际的。其次，**n-gram** 对语义相似度建模的能力有局限性，它仅考虑词在给定上下文出现的概率，但对上下文的相似性没有考虑。比如，“房间里趴着一只狗”和“卧室里趴着一只猫”，对“猫”和“狗”建模时，**n-gram** 模型没有考虑“房间”和“卧室”的相似性，而它们的相似性是可以根据其他文本得到的。我们需要找到一种可以将上下文相似性一并考虑的方法模型。

换个角度思考，我们要求的是 $p(w_t | w_{t-n+1}^{t-1})$ ，实际可以看做是一个函数 $f(w_{t-n+1}, w_{t-n+2}, \dots, w_{t-1}, w_t)$ 。Hornik 等人证明了，带有隐含层的多层神经网络，可以近似 R^n 上任意连续函数 (universal approximation theorem)^[29]。因此，我们也能用神经网络来逼近 $f()$ 。我们把 $w_{t-n+1}, w_{t-n+2}, \dots, w_{t-1}$ 的 One-hot Representation 作为输入，希望输出 $output_i = P(w_t = i | w_{t-n+1}, w_{t-n+2}, \dots, w_{t-1})$ ，这样就是 **n-gram** 语言模型的神经网络近似。如果我们把输入换成每个词对应的 K 维向量

表示，那么上下文的相似性可以自然体现在向量的相似性中，这个模型就是神经网络语言模型（NNLM）^{[27][28]}。与一般神经网络不同的是，它的输入，即每个词的向量表示，是未知的，需要和模型参数一同优化。如图3.2所示。

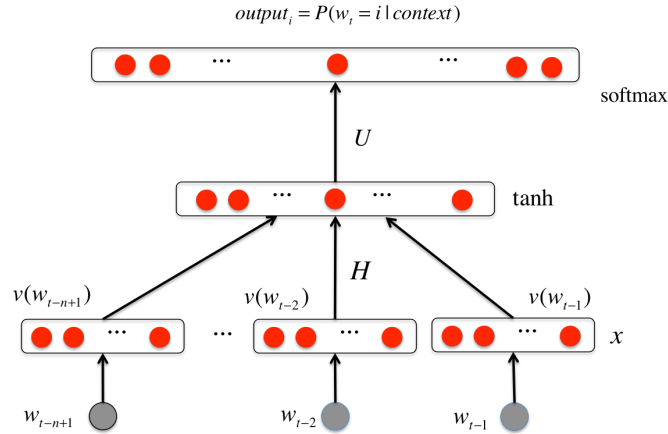


图 3.2 神经网络语言模型

此网络输入为词序列 w_{t-n+1}^{t-1} ， $v(w_i)$ 为词 w_i 的向量表示，输出经过 softmax 归一化为词出现的概率，如式3-2。可用随机梯度优化网络参数。由于输出层为 softmax，不会出现概率为 0 的情况，因此不需 n-gram 模型中复杂的平滑方法，并且可以取得更好的效果。

$$y = b + U \tanh(d + Hx)$$

$$output_t = P(w_t | w_{t-n+1}^{t-1}) \quad (3-2)$$

$$= \frac{e^{y_i}}{\sum_i e^{y_i}}$$

Google Word2Vec^①，提供了使用前馈神经网络对神经网络语言模型进行训练的高效工具。

3.5.3 案例研究

我们在 1300 万条新浪微博上训练了 NNLM，向量维度一般取 50 到 200 维，本文取 100 维。本节将对得到的词向量做案例研究，以检验模型的有效性。

① <https://code.google.com/p/word2vec>

首先，我们给出一个活动短语，用余弦相似度衡量语义上的相似性，找出和其最相似的 10 个词组，如表3.4。可见，根据 NNLM 得到的向量表示，很好地反映了语义的相似性。

表 3.4 语义相似度样例

短语	相似度	短语	相似度	短语	相似度
打_篮球	-	打扫	-	吃_晚饭	-
打球	0.800	打扫_卫生	0.805	吃_午饭	0.920
踢_足球	0.790	收拾	0.757	晚饭	0.844
打_网球	0.789	洗_衣服	0.747	吃_中饭	0.831
打_羽毛球	0.797	家里_打扫	0.699	吃饭	0.816
踢球	0.727	收拾_屋子	0.693	吃_夜宵	0.795
踢_球	0.697	搞_卫生	0.666	吃_早饭	0.780
练_球	0.686	捣鼓	0.659	午饭	0.777
游泳	0.667	打扫_打扫	0.659	出去_觅食	0.755
排球	0.651	洗_床单	0.654	中午_饭	0.746
打_排球	0.649	拖_地板	0.652	喝_早茶	0.736

并且词向量具有可加性^[28]。例如，可以用向量加法来表达语义关系的组合，此处以：

$$v(\text{济南}) - v(\text{山东}) + v(\text{浙江}) \rightarrow v(\text{杭州})$$

$$v(\text{北京}) - v(\text{中国}) + v(\text{法国}) \rightarrow v(\text{巴黎})$$

这里的 \rightarrow 表示, 在所有的词中, 约等号右边的词向量与左边运算的结果余弦相似度最大. 这两个式子的含义是, 杭州对浙江的关系, 与济南对山东的关系相似 (都是省会); 东京对日本的关系, 与北京对中国的关系相似 (都是首都)。

再如

$$v(\text{喝}) - v(\text{渴}) + v(\text{饿}) \rightarrow v(\text{吃})$$

$$v(\text{校长}) - v(\text{学校}) + v(\text{公司}) \rightarrow v(\text{总裁})$$

$$v(\text{胡锦涛}) - v(\text{中国}) + v(\text{日本}) \rightarrow v(\text{首相})$$

这些式子都有很明确自然的语义关系。虽然这种语义组合的关系并不是总能成立，但反映出，词的向量表示在某种层面上反映了词的语义特征。因此，我们使用词的向量表示作为短语特征。对于 unigram w , 特征向量就是 $v(w)$ 本身；对于 bigram $\langle w_i, w_j \rangle$, 特征向量为 $\frac{v(w_i)+v(w_j)}{2}$ 。

3.6 训练集选取

3.6.1 动机与目标

在3.5中，我们得到了每个词对应的向量表示。对于 unigram, 词向量本身就是短语的向量；对于 bigram, 由于向量可加，将动词和名词的向量相加后归一化，如此，每个动词短语可以看做 K 维语义向量空间中的一个向量，更确切来说，每个向量的模长均为 1，它们分布在一个 K 维超球面上。

在对短语进行分类前，需要手工构建一个标注集，类标“1”表示是活动，“0”表示非活动，作为训练数据训练分类模型，确定模型参数。传统上，训练样本是从待分类数据中随机抽样得到。由于语义空间很大，待分类的数据也比较多，使用随机抽样得到的训练样本训练 SVM 分类器，精度为 69%, 召回率相对 One-hot Representation 有所提升，但依然只有 43%。更多的标注可以改善精度和召回率的情况，但大量标注需要耗费很常时间。本文希望根据数据分布的特性，希望设计训练样本抽样方法，在标注数据量一定的情况下，能够尽可能提高训练的效果。假设标注集为 L ，我们需要恰当定义其效用函数 $Q(L)$ 以判断它的有效性，并且限制 L 的大小为 M ，找出最优的 L , 也就是

$$L^* = \arg \max_{L, |L|=M} Q(L)$$

这个问题与 Active Learning 有相似之处，但 Active Learning 一般是基于一个已有的分类算法，寻找使分类器性能提高最多的样本。而本节希望找到一种抽样方法，使得抽样出的集合等尽可能全面地代表数据集，独立于特定的分类算法。目标相似，但途径不同。

通过第3.5.3案例研究，可以发现和一项活动有较高相似度的短语，基本上也是一个活动，从而标注一个训练样本后，和此训练样本相似的样本，我们有较高的概率将其正确分类。例如，在 K 近邻分类器中，对于每个未知样本，寻

找训练集中和其距离最近的 K 个样本，由这 K 个近邻投票确定此样本的类标；在 SVM 中，若 x_i 与 x_j 相似，则 $w x_i + b$ 与 $w x_j + b$ 也比较相似。

在我们的问题中，样本间的相似性 $\text{sim}(x_i, x_j)$ 使用向量的余弦相似度来表示 $\langle x_i, x_j \rangle = \frac{x_i \cdot x_j}{|x_i| |x_j|}$ ，据此可以定义一个样本 x 和集合 S 的相似性：

定义 3.2:

$$\text{sim}(x, S) = \max_{u \in S} \text{sim}(x, u)$$

即，样本 x 与集合 S 的相似度为 S 中与 x 最相似的元素与 x 的相似度。

设全部数据样本的集合为 U ，标注集为 L ，我们希望对于每个未知样本，在 L 中，都能找到一个与其相似度高的元素，即标注集能够更好地代表所有数据。根据这个想法，我们的效用函数可以定义成

$$Q(L) = \min_{v_i} \text{sim}(v_i, L)$$

意义是，所有样本与 L 的最小相似度。 $Q(L)$ 越大，说明 L 越好地代表了数据集。我们的目标是，限制 L 的大小为 M ，找到最优的标注集 L^* 。问题可以形式定义为：

问题 3.1: 给定集合 U ，对任意元素 $x_i, x_j \in U$ ，有相似度度量 $\text{sim}(x_i, x_j)$ 。元素与集合相似度 $\text{sim}(x, S)$ 和效用函数 Q 如前定义。要求最优子集

$$L^* = \arg \max_{L \subseteq U, |L|=M} Q(L)$$

3.6.2 NP-Hardness 证明

问题3.1是一个组合优化问题，这类问题通常是 NP-Hard 的，下面我们将集合覆盖问题可以归约到此问题来证明其 NP-Hardness。

集合覆盖问题的判定版本是 Richard Karp 在 1971 年提出的 21 个 NP 完全问题之一。问题定义为：

问题 3.2 (集合覆盖): 给定全集 U ，一族子集 $S = \{S_i\}$, $\cup S_i = U$ ，以及整数 M ，判定是否存在覆盖 $[C \subseteq S, |C| = M]$ ，使得 $\cup \{S_i | S_i \in C\} = U$

问题3.1的判定问题是

问题 3.3 (问题3.1的判定问题): 给定阈值 θ , 判定是否存在大小为 M 的集合 L , 使得

$$\theta \leq \min_{x_i \in U} \max_{x_j \in L} \{sim(x_i, x_j)\}$$

如果能证明此判定问题的 NP 完全性, 那么原问题的 NP-Hardness 就得证。显然, 如果能解决集合覆盖问题, 问题3.3也可得到解决。但为了证明3.3的 NP 完全性, 需要进行相反方向的归约, 即证明, 如果能在多项式时间解决问题3.3, 则集合覆盖问题也可在多项式时间解决。证明如下。

证明 对于集合覆盖问题3.2, 如果存在一个元素 $x \in U$ 不在任何一个 $S_i \in S$ 中, 那么覆盖显然是不存在的。下面仅考虑每个元素都至少被一个 $S_i \in S$ 覆盖的情况。

设 $|U| = N, |S| = P$, 我们构造一个包含 $N + P$ 个结点的图 $G = (V, E)$ 。其中, 结点集 V 为:

- 集合 U 中的每个元素 $x_i, 1 \leq i \leq N$ 都对应 G 中的一个结点 v_i , 称为元素结点 (element nodes)。
- 对每个 $S_i \in S$, 对应节点 $v_{i+N}, 1 \leq i \leq P$, 称作集合结点 (set nodes)。

边集 $E = \{(v_i, v_{j+N}) | v_i \in S_j, 1 \leq i \leq N, 1 \leq j \leq P\} \cup \{(v_i + N, v_j + N) | 1 \leq i, j \leq P\}$ 。即, 每个集合都和它包含的元素连边, 集合之间两两连边。所有边赋权为 θ , 不存在边的赋权负无穷。如图3.3所示。

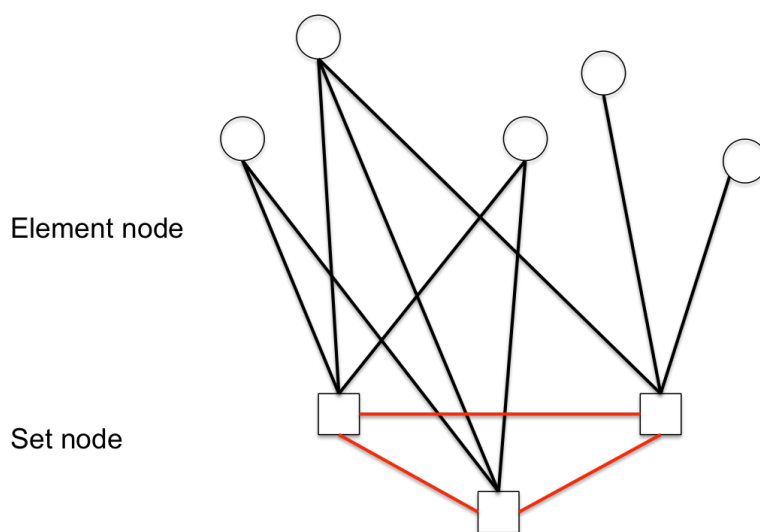


图 3.3 问题归约

假如我们能在多项式时间内求解问题3.3，则分两种情况。

1, 如果存在满足条件 L ，可在多项式时间构造集合覆盖问题的解 C ：对任意 $v_i \in L$ ，若 v_i 为集合结点 (即 $N + 1 \leq i \leq N + P$)，则 $S_i \in C$ ；若 v_i 为元素结点 ($N + 1 \leq i \leq N + P$)，取任意包含 v_i 的任意集合 $x_i \in S_j$ ，使得 $S_j \in C$ 。这样， C 构成了 U 的一个覆盖。

2, 如果 L 不存在，则集合覆盖也不存在。否则，若存在 U 的一个覆盖 C ，选择 C 中集合对应的结点构成 L ，就可以得到问题3.3的解，矛盾。

这样，集合覆盖问题就在多项式时间内归约到了问题3.3。又由于它也可以归约到集合覆盖问题，因此，问题3.3是一个 NP 完全问题。其优化版本，即我们要求解的问题3.1，是一个 NP-Hard 问题。不存在已知的多项式时间解。 \square

3.6.3 子模性及近似求解

节3.6.2已经证明最优化 $Q(L)$ 是一个 NP-Hard 问题，但是下面我们将证明其单调性和子模性，并据此得到一个贪心算法 (greedy algorithm)，在多项式时间内得到近似解 $Q(L')$ ，并且保证

$$Q(L') \geq (1 - \frac{1}{e})Q(L^*)$$

首先给出子模函数 (submodular function) 的定义。

定义 3.3 (子模函数)： 从幂集 $2^\Omega \rightarrow R$ 的一个函数 f 称为子模函数，如果对任意 $X, Y \subseteq \Omega, X \subseteq Y, \forall x \notin Y$ ，有 $f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$ 。

下面我们分别证明 $Q(L)$ 的单调性和子模性。单调性是显然的。

证明 (单调性) 令集合 $Y = X \cup x$ 。对任意元素 $y \in U$ ，设 $\text{sim}(y, X) = \text{sim}(y, t), t \in X$ ，即对任意 $x \in X, \text{sim}(y, x) \leq \text{sim}(y, t)$ 。

由于 $X \subset Y$ ，故有

$$\begin{aligned} \text{sim}(y, Y) &= \max_{x \in Y} \text{sim}(y, x) \\ &\geq \max_{x \in X} \text{sim}(y, x) \\ &\geq \text{sim}(y, X) \end{aligned}$$

从而, $\text{sim}(y, Y) \geq \text{sim}(y, X)$ 。由 y 的任意性, 得到

$$\begin{aligned} Q(Y) &= \min_{y \in U} \text{sim}(y, Y) \\ &\geq Q(X) \end{aligned}$$

单调性证完。 □

证明 (子模性) 设集合 $X \subset Y$, 对任意 $x \in U - Y$, 设 $X' = X \cup x, Y' = Y \cup x$ 。对任意 $y \in U$, 由单调性有

$$\begin{aligned} \text{sim}(y, Y') &\geq \text{sim}(y, X') \geq \text{sim}(y, X) \\ \text{sim}(y, Y') &\geq \text{sim}(y, Y) \geq \text{sim}(y, X) \end{aligned}$$

对于新加入的元素 x , 有

$$\begin{aligned} \text{sim}(y, X') &= \max\{\text{sim}(y, x), \text{sim}(y, X)\} \\ \text{sim}(y, Y') &= \max\{\text{sim}(y, x), \text{sim}(y, Y)\} \end{aligned}$$

于是

$$\begin{aligned} \text{sim}(y, Y') - \text{sim}(y, Y) &= \max\{\text{sim}(y, x) - \text{sim}(y, Y), 0\} \\ &\leq \max\{\text{sim}(y, x) - \text{sim}(y, X), 0\} \\ &= \max\{\text{sim}(y, x), \text{sim}(y, X)\} - \text{sim}(y, X) \\ &= \text{sim}(y, X') - \text{sim}(y, X) \end{aligned}$$

由 y 的任意性, 子模性得证。 □

对于单调非减的子模函数, 有以下定理

定理 3.1: 对于一个单调增的子模函数 Q , 从空集 L_0 开始, 使用贪心策略进行迭代, 即第 k 次迭代选取元素 v_k , 使得

$$v_k = \arg \max_{v_k \notin L_{k-1}} Q(L_{k-1} \cup \{v_k\})$$

$$L_k = L_{k-1} \cup \{v_k\}$$

那么 K 次迭代之后, 对任意 $L, |L| \leq K$, 有

$$Q(L_k) \geq (1 - \frac{1}{e})Q(L),$$

基于此定理, 可以得到了一个保证下界的贪心算法2。

算法 2: 问题3.1贪心算法

Input: 样本集合 U , 目标大小 M

Output: $L^* = \arg \max_{L, |L|=K, L \subseteq U} Q(L)$

$L_0 = \emptyset$;

for $k = 1; k \leq M; k+ = 1$ **do**

foreach $x_i \in U - L_{k-1}$ **do**

$q(x_i) = \min_{t \in U} \max\{sim(t, L_{k-1}), sim(t, x_i)\}$

$x_{selected} = \arg \max_{x \in U - L_{k-1}} q(x)$

$L_k = L_{k-1} \cup \{x_{selected}\}$

return L_M ;

3.7 模型训练与实验结果

为了检验我们标注集选取和分类的效果, 我们从数据集中随机抽取了 5000 个短语进行标注, 作为客观事实 (Ground Truth) 检验我们的方法。

首先, 我们使用随机抽样 (Random) 和通过最大化 $Q(L)$ (MaxSim) 的两种训练集选取方法分别抽取 500、1000、2000 个训练样本, 训练 SVM 分类器, 在剩余测试数据中的表现。对于随机抽样, 进行了 5 次试验, 记录每次试验的结果, 并标记最大值、最小值、平均值。

图3.4中发现, 我们在节3.6中提出的训练集选取算法, 在 Recall 和 Precision 方面, 都稳定的高于随机随机抽样, 尤其是在标注量较小时, MaxSim 选取的标注集在 Recall 方面的优势更加明显 (61% 对 52%)。

对结果进行错误分析, 我们发现, 有些相似度高的词, 类标并不相同, 如“天亮”和“起床”, “开场”和“看演出”。为了解决这一问题, 我们引入了其他特征帮助分类。

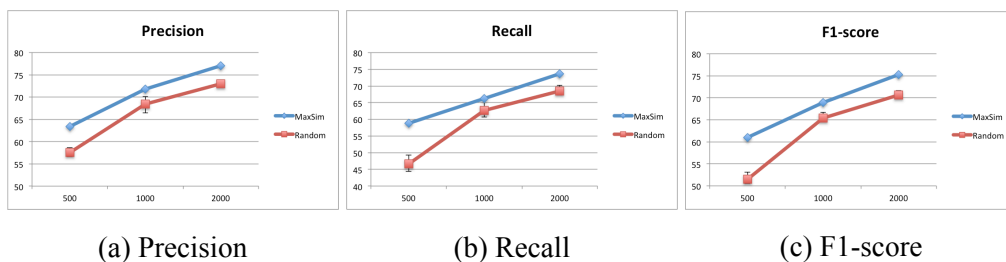


图 3.4 概念抽取初步结果

表 3.5 概念抽取实验结果

	Accuracy	Precision	Recall	F1 score
结果	85.1%	85.6%	81.3%	0.834%

- 签到信息。
- 时间词，如“今天”，“上午”，“晚上”等。
- 同一句子中是否出现人称代词。
- 短语的时间分布。

加入这些特征后，我们的抽取方法得到了更好的结果，如表3.5。最终，我们从微博中抽取出 13470 个活动概念。

3.8 本章小结

本章着力解决活动概念抽取问题。从候选集抽取、语义向量获取、测试集选取、模型选择到最后的分类，得到了较好的分类精度。方法的核心在于通过神经网络语言模型训练词的向量表示，把语义相似度的计算简化为计算余弦相似度。在此基础上提出的基于最优化效用函数的训练集选取方法，使得以较小的标注成本获取更好的分类效果。时间、地理信息特征的引入，进一步提高了分类精度，取得了较好的结果。

第 4 章 实例抽取与关系挖掘

4.1 目标

对于一项活动，除了抽象的活动概念以外，还应有相应的属性，如情感状态、时间、地点。为此，我们定义了活动实例(定义1.1)。本章的目标在于，输入一条特定微博 m ，从中抽取活动相关属性，构建活动的实例 a ；并基于实例抽取的结果，进一步构建活动之间的关系。

4.2 活动类别抽取

在第3.1中，我们已经抽取出一系列抽象的活动概念。在构建活动实例时，我们首先要获取微博 m 对应的活动概念，即活动类别抽取。

类似于第3.1章的方法，本文首先使用了规则匹配，首先抽取微博中所有的动词短语，并判断是不是对应一个活动概念，这一步只需判断它是否在抽取出的活动概念集合即可。如果不存在，则认为这条微博包含一个活动。这样做的结果精度为 74%，但召回率仅为 56%。对结果的分析发现，大约有 11% 的缺失是由于语法结构的复杂性。用户在表述一项活动时，动作和目标有时并不构成一个简单的动宾短语，例如以下情况

- 宾语可以前置，如“找本书读了一下午”，
- 加入复杂的修饰成分，如“陪父母看了一场周星驰拍的很有意思的电影”
- 以主谓、定中等形式出现。如“练了一天的琴”。

这些较复杂的情况，通过简单的词法分析和规则匹配是难以处理的。为解决这个问题，本文对句子进行句法分析，将线性文本处理为依存树。句法分析在当前已有许多成熟的分析器，如 Stanford Parse^① 和 HIT-LTP^②。

Stanford Parser 是斯坦福大学开发的句法分析工具。它主要面向英语语言，也提供中文、德语、阿拉伯语等其他语种的解析，但它的处理结果并不非常直

① <http://nlp.stanford.edu/software/lex-parser.shtml>

② <http://www.ltp-cloud.com/>

表 4.1 活动类别抽取实验结果

	Precision	Recall	F1 score
Result	73.2%	68.4 %	0.707

观。最终本文选择哈尔滨工业大学开源的语言技术平台 (Language Technology Platform, LTP)^[30] 作为句法分析工具。LTP 接收原始文本作为输入，结果以 XML 的形式给出，对每个词，输出其依赖词和依赖关系，依赖关系包括主谓 (SBV)，动宾 (VOB)，定中 (ATT)，并列 (COO) 等。图4.1是 LTP 进行句法分析的结果的样例，可见，它正确分析出了“看”和“电影”的动宾关系。

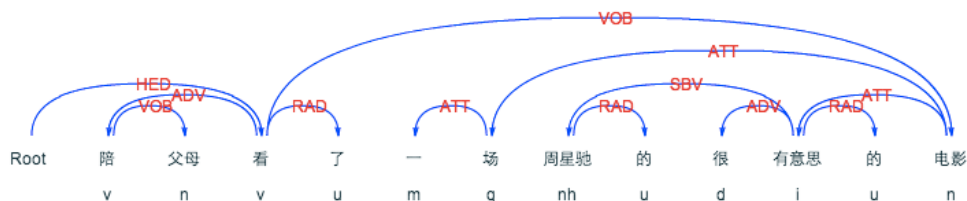


图 4.1 LTP 句法分析示例

通过 LTP 进行句法分析，得到词语之间的依赖关系，进一步抽出动词短语，判断是否是一个活动概念，结果如表4.1。

除精度、召回率以外，处理速度也是一个关键的问题。将整条微博作为 LTP 的输入，处理速度非常慢，分析工作的复杂度随文本长度指数增长。为提高速度，首先将文本分割为多个短句，再进行分析，就可以达到较快的速度。由于不同微博的分析工作是独立的，为了进一步提高处理能力，可以使用多线程、多机进行并发。图4.2中比较了不同处理策略，分析 1000 条微博的速度。

本文方法的不足在于，没有考虑特定命名实体与活动类别的关系。如，“下午去看了阿凡达”这类微博。“阿凡达”是一部电影的名称，因此这条微博应对应“看电影”这个活动类别，但本文的方法无法正确抽取这类活动。类似的情况如餐馆、电视剧、特定的地名等等。在下一步工作中，可以引入外部的知识库，以建立命名实体与活动的联系。

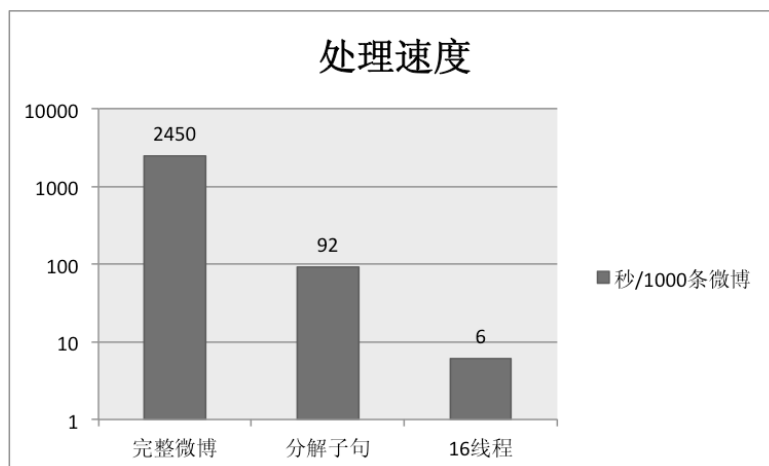


图 4.2 分析速度

4.3 情感极性分析

在活动抽取中，如果微博中包含用户参与的一项活动，我们希望了解用户参与这项活动时的心情状态如何，帮助我们发现活动本身是正面还是负面，这样帮助推荐系统进行选择。因此，需要获取微博的情感极性。

4.3.1 方法概述

本文的工作中，对微博进行三分类，正面 (positive) 表示积极的情感，如高兴、舒畅等，负面 (negative) 表示消极的情感，如愤怒、悲伤等，中性 (neutral) 表示用户没有在微博中表现出明显的情感倾向，如仅仅表述客观事实。

我们对非监督和监督学习均进行了尝试。在非监督的方法中，我们从知网 (HowNet) 等途径获取了中文的情感词典，共有 4986 个积极词汇和 4818 个消极词汇，包含动词、形容词和名词。对每一条微博，我们首先计算其中积极、消极词汇的数量 n_{pos} 和 n_{neg} ，若其差值 $|n_{pos} - n_{neg}| < \theta$ ，则认为微博是中性的，否则判别为词数多的类别。但这样简单的非监督方法只能达到 55% 的正确率 (注意到这是三分类问题，这个结果还是比随机分类 (33%) 和全部判为最多的类别 (45%) 要好)。为此，我们使用监督学习的方法，使用以下特征训练分类器

1. Bigrams 和 Unigrams 的频度。使用 bigram 的原因是，情感词之前常常会带有修饰性的前缀，如“不”，“非常”，有时会加强或者逆转情感词的极性。因此对于较频繁出现的模式，使用 bigram 作为特征。
2. 正面词、负面词出现的频度。这可以根据情感词典得到。

3. 表情符号。用户在发布微博时，常常会加入一些表情，如“高兴”，“愤怒”，有时用户选择表情并不关心这个表情具体的含义是什么，但是也可以体现出用户当时的心理状态。

同时，为了加快训练速度和减少噪声，我们将过于稀疏，即出现次数少于一个下界的特征滤除。

4.3.2 实验结果与分析

我们标注了 20815 条包含活动信息的微博，根据其情感极性标注为 5 级，-2 为很负面，-1 为一般负面，0 为中性，1 为一般积极，2 为很积极，其中分级 -2、-1 为负面，+1,+2 为中性，0 为中性。为了避免不同人倾向性的不同，每条微博会有至少两个人标注，如果出现分歧，由实验者最终决定类别。在数据中，共有 9462 条为中性,6566 条为正面，4787 条为负面。在此数据上进行交叉验证。

我们首先检验不同特征的选取对分类精度的影响，如图4.3。可以看到，我们选取的特征，对于分类结果都有明显的提升，其中情感词词典的作用最为明显。

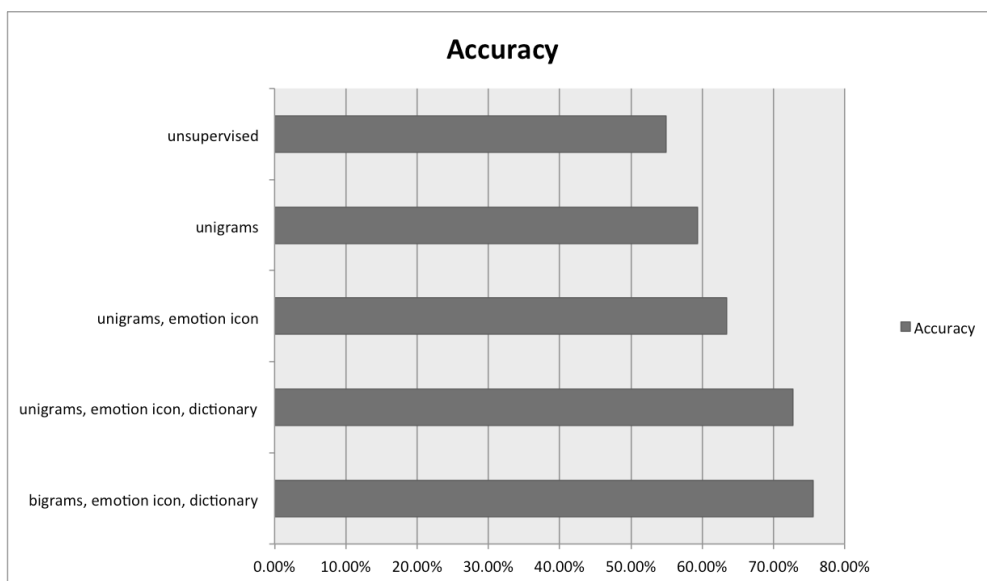


图 4.3 特征选择

本文继续尝试了不同的分类模型，包括

- 朴素贝叶斯
- 以决策树为基础的 AdaBoost

- 随机森林
- 线性核 SVM

测试结果如图4.4。线性核 SVM 表现最好，但朴素贝叶斯也有比较好的性能，与^[16] 中的结果一致。

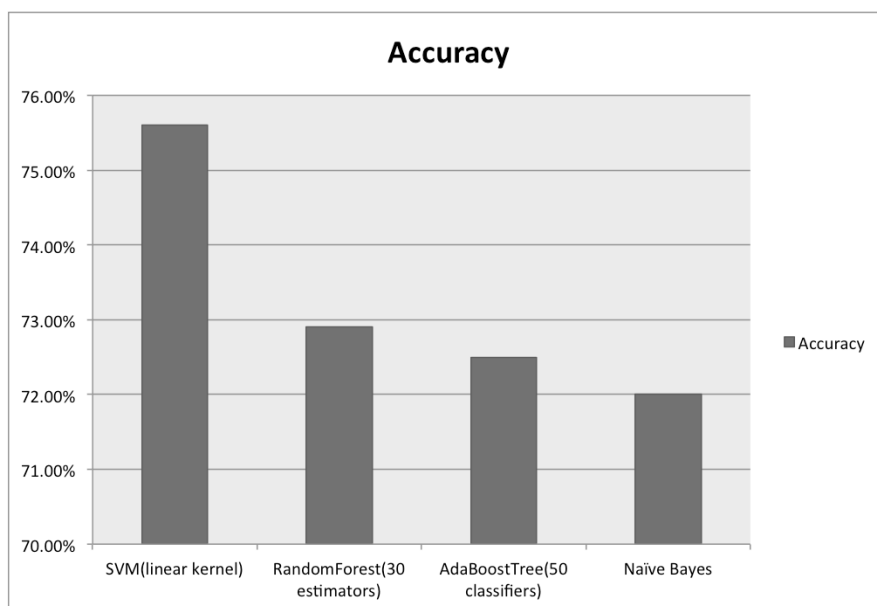


图 4.4 模型选择

最终达到的 75% 的分类精度从数值上看并不是很高，但情感极性的判断有很强的主观性，正面负面和中性并没有很明显的界限。根据之前的研究，人类对情感的判断也只能在 79% 的情况下达成一致，因此，一个准确率 70% 以上的系统在实际中是可用的。

4.3.3 进一步工作

文档级的情感分类做出了一个假设，即用户在一个微博中只提到一项活动，或者提到多项活动但情感极性是相同的，这在社交媒体短文本的限制下是合理的，在我们的统计中，只有 3% 的微博描述多个活动，并且带有相反的情感倾向。但这在实际中并不总是成立。同时，即使同为负面情感，用户的具体情感也可能非常不同，如“疲倦”、“伤心”、“生气”三者代表的情感完全不同，无法用简单的极性表达。Ekman^[31] 将人类情感分为 6 中基本情感，“高兴 (Happy)”，“激动 (Excited)”，“温和 (Tender)”，“惊吓 (Scared)”，“悲伤 (Sad)”，“生气 (Angry)”。进一步可以考虑情感的多分类问题。

第二点是，一个活动中会有不同的子方面，如旅游，用户可能会对天气、交通、饮食、住宿等方面分别表达情感。这种基于方面的细粒度的观点挖掘在商品评论中有比较多的研究，下一步工作中会加以考虑。

4.4 地点、时间抽取

新浪微博的数据，可以分为两类

1. 带有地理信息。通过移动终端发布微博，并开启定位服务，在微博元信息中会包含微博发表时的地理位置，包含经纬度；有些微博还包含对应兴趣点（POI）的信息。
2. 普通微博。仅带有发布时间，无法通过元信息直接获取地点。

对于第一类带地理信息的微博，我们可以通过经纬度，直接获取到地点。我们与搜狐合作，得到的国内主要城市的兴趣点数据，来辅助我们的地点抽取。POI 数据中包含位置，类型，名称等信息，样例如表4.2。通过经纬度在兴趣点数据中寻找最近的兴趣点，即可获取活动发生的地点，包括城市，地址，POI 类型。这些微博占有所有微博的 16%，但总数很大，共有 620 万条左右，对于我们挖掘活动信息已经比较充分了。

表 4.2 POI 样例

Attribute	Value
Name	北京密云东方商贸大厦
Address	北京市密云县新东路 40
Contact	010-69043424
Type	购物场所
Category	一般商场
Province	北京市
City	北京市
District	密云县
Longitude	13008155.004543
Latitude	4892797.007317

对于第二类微博，我们通过分析文本获取地点信息，这类信息抽取任务可以使用 CRF 来完成，为了减少标注所需的工作量，我们使用了自监督的方法。可从微博中直接抽取到地点的微博，我们从文本中定位地点的位置进行标注，

作为正例训练标注模型。地点信息通常带有明显的上下文特征，可以利用这些信息，对地点进行分类。我们使用以下特征进行分类：

- 词本身 w_i ，对于常见地点、城市，如“超市”、“北京”，词本身可以提供有效的特征。
- 词性。地点通常为名词，并且在 ICTCLAS 的工具中，对于常见地点词，会标记为地点，可以加以利用。
- 前一个词 w_{i-1} 。我们统计地点词之前一个词的词频，频度最高的词有：在、去、抵达、到达、来、来到等。这些标志词，对地点识别有很大帮助。

我们标注了 2000 条微博中进行测试，结果如表4.3。表4.4是对一些微博进行地点抽取的样例。

表 4.3 地点抽取实验结果

	Precision	Recall	F1 score
Result	62.6%	81.4%	0.717

表 4.4 地点抽取结果样例

微博	抽取结果
打车去高铁站	高铁站
奔波一天，终于回上海了	上海
在文华园吃的很 nice	文华
凌晨五点，寒冷的北京	北京
在长城上跳骑马舞最好玩了	长城
见过人海吗？快来乌镇！	乌镇
我在这里	这里
兰州军区那么多丰田越野车	兰州
就算一车切糕一万元。水果不值钱嘛，去超市看看	超市
在新疆叶城时，天黑去维吾尔族聚居区吃的宵夜。	新疆

4.5 序列关系挖掘

知识库系统还需要构建概念之间的关系。与通常的知识库系统一般建模概念之间的上下位关系不同，本文关注活动间的序列关系 (follow-up relation)，即

用户在参加一项活动后，通常进行的下一项活动是什么。这一点有助于帮助我们对用户行为进行建模，进行活动的推荐。

通过之前的工作，我们已经在微博中抽取出大量活动的实例，序列关系的抽取可以在这个基础上进行。序列关系的强弱包含两个方面

1. 用户进行活动 c_i 后，在时间窗口 T 内，进行活动 c_j 的概率
2. 用户进行活动 c_i 和 c_j 之间的期望时间

这两个方面缺一不可。由于用户行为的复杂性，第一项可以对噪声进行抑制，避免个别用户随机行为的影响，使挖掘出的活动有较高的置信度；第二项表示两项活动间隔的时间越短，它们的序列关系越密切。基于这两个考虑，我们对问题定义如下：

问题 4.1 (序列关系挖掘)： 给出

- 活动概念集合 $C = c_i, i = 1, 2, \dots, N^c, N^c$ 为活动概念的数量。
- 用户集合 $U = \{u_i\}, i = 1, 2, \dots, N^u, N^u$ 为用户数量
- 活动实例集合 $A = A_i$ 。对每个用户 u_i ，有其参加活动实例的集合 $A_i = \{a_{ij}\}, j = 1, 2, \dots, N_i^a, N_i^a$ 是用户 u_i 在给定微博语料中参与活动实例的个数，每个活动实例 a_{ij} 是活动概念、时间、地点、情感极性的四元组，即 (c, t, p, s) 。

求在给定时间窗口 T 内，

1. 用户进行活动概念 c_i 后进行活动 c_j 的概率 $P(c_i|c_j, T)$
2. 在 $P(c_i|c_j, T) > \lambda$ 的条件下， $E(t_{c_j} - t_{c_i})$

根据问题的定义，可以得到算法如下：

在我们的系统中，时间窗口取 6 小时。为了查询时的快速响应，本文对所有活动离线进行计算，对每项活动，记录序列关系最强的 10 项活动。

4.6 本章小结

本章关注活动实例中属性的抽取，包含类别、地点、时间、情感极性等。我们使用句法分析器 HIT-LTP 提高的活动抽取的召回率。在情感分类中，通过恰当选取特征，本文训练了分类模型，取得了较好的分类效果。在地点抽取中，我们借助已有的地理信息和 POI 数据，使用了自监督的学习方法，避免了繁重的手动标注。最后基于实例抽取的结果，我们分析了活动间的序列关系。

算法 3: 序列关系挖掘

Input: 活动实例集合 I , 用户集合 U , 每个用户 u_i 的活动实例集合 A_i , 活动概念集合 C , 阈值 λ , 时间窗口 W

Output: 对每个活动概念 $c_i \in C$, 随后可能发生活动的序列 seq_i

foreach $A_i \in A$ **do**

foreach $a_{ij} \in A_i$ **do**

foreach $a_{ik} \in A_i, t_{ij} < t_{ik} < t_{ij} + W$ **do**

$seq_occur(c_{ij}, c_{ik}) + = 1$;

$past_time(c_{ij}, c_{ik}) + = t_{ik} - t_{ij}$;

foreach $c_i \in C$ **do**

$follow_{c_i} = \{c_j | seq_occur(c_i, c_j) > \lambda\}$;

 sort $follow_{c_i}$ by $past_time(c_{ij}, c_{ik}) / seq_occur(c_{ij}, c_{ik})$;

$seq_i = follow_{c_i}$;

return $\{seq_i\}$

第 5 章 ActivityNet 系统设计

5.1 设计目标

通过之前的工作，本文从社交媒体中抽取了活动概念、地点、情感等属性，并挖掘了活动间的相关性和序列信息。基于这些结果，本文构建了一个可视化系统 ActivityNet。系统希望实现以下目标：

- 高效的活动检索
- 良好的可视化界面
- 提供用户反馈机制，纠正错误分类

5.2 底层架构

为了系统的高效性，我们选择 SAE 作为系统底层架构。SAE 全称为 Social Analytic Engine，即社交网络分析引擎，是清华大学计算机系 KEG 研究组开发的一套社交网络分析平台，架构图见图5.1，提供了以下功能：

1. 存储和快速检索极大规模社交网络的数据
2. 提供最新的社交网络分析算法和机器学习算法，如话题模型（Topic Model），影响力最大化模型（Influence Maximization）等
3. 提供通用的网络分析引擎和机器学习引擎。

在 ActivityNet 中，我们主要用到 SAE 的存储和索引机制。由于 SAE 主要用于处理社交网络数据，因此图是 SAE 主要的存储模型。本文将活动以及之间的关系抽象为图结构进行存储。为了检索的高效，对所有活动名称建立了索引。同时，由于相似活动搜索是基于语义相似度的，本文也对词向量使用 KD-Tree 进行索引。

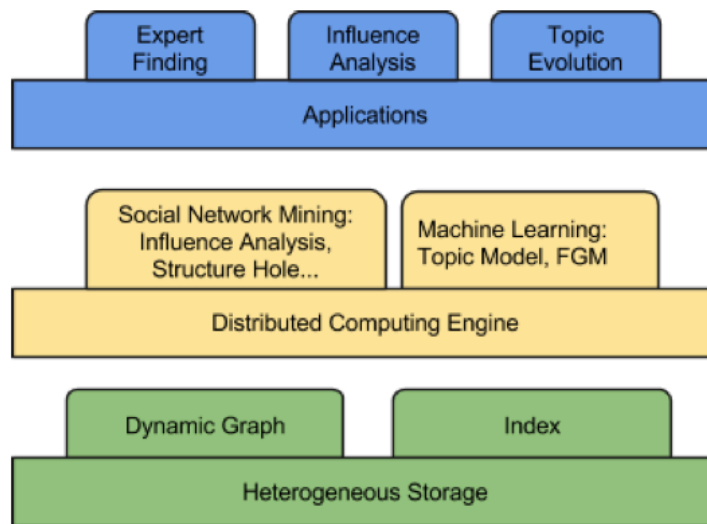


图 5.1 SAE 基本结构

5.3 功能实现

ActivityNet 以 Web 站点 的形式提供服务, 基于成熟的前端框架 *TwitterBootstrap*, 简化了开发工作。下面我们以查询“吃烤鸭”为例, 介绍系统中各个功能。

5.3.1 首页设计

ActivityNet 的首页如图5.2。

首页提供以下功能

1. 活动搜索。用户可以自行输入自己希望查询的活动
2. 城市热门活动推荐。我们计算出北京、上海、广州的热门活动, 在首页进行推荐。
3. 全网热门活动。基于活动实例抽取结果, 我们选择了 18 个最热门的活动, 在首页底部呈现。活动对应的图片是以活动本身为关键词构建 GET 请求, 在百度图片搜索抓取第一张图。自动抓取的个别活动图片不很理想。我们在自动抓取图片后, 人工进行一些调整。



图 5.2 Activity 首页设计

5.3.2 活动搜索

用户输入一个查询后，可以进入此活动查询结果的二级页面。在此，输入“吃烤鸭”，查询结果如图5.3

在页面右侧，我们列出了和查询相关的前 10 个活动，以相关性排序；其次，利用序列关系抽取的结果，我们列出了此活动经常发生的后续活动。烤鸭是北京的特色菜品，“相关活动”中，我们列出的“炸酱面”，“涮羊肉”，“麻辣烫”也都是北京的特色菜。在序列关系中，ActivityNet 提供了“散步”，“回家”，“游泳”，“探索城市”，“逛街”，“逛超市”等活动，是比较合理的。由于概念的抽取难以做到完全的精准，ActivityNet 的结果中有一些噪声，为此，我们提供用户反馈机制。如

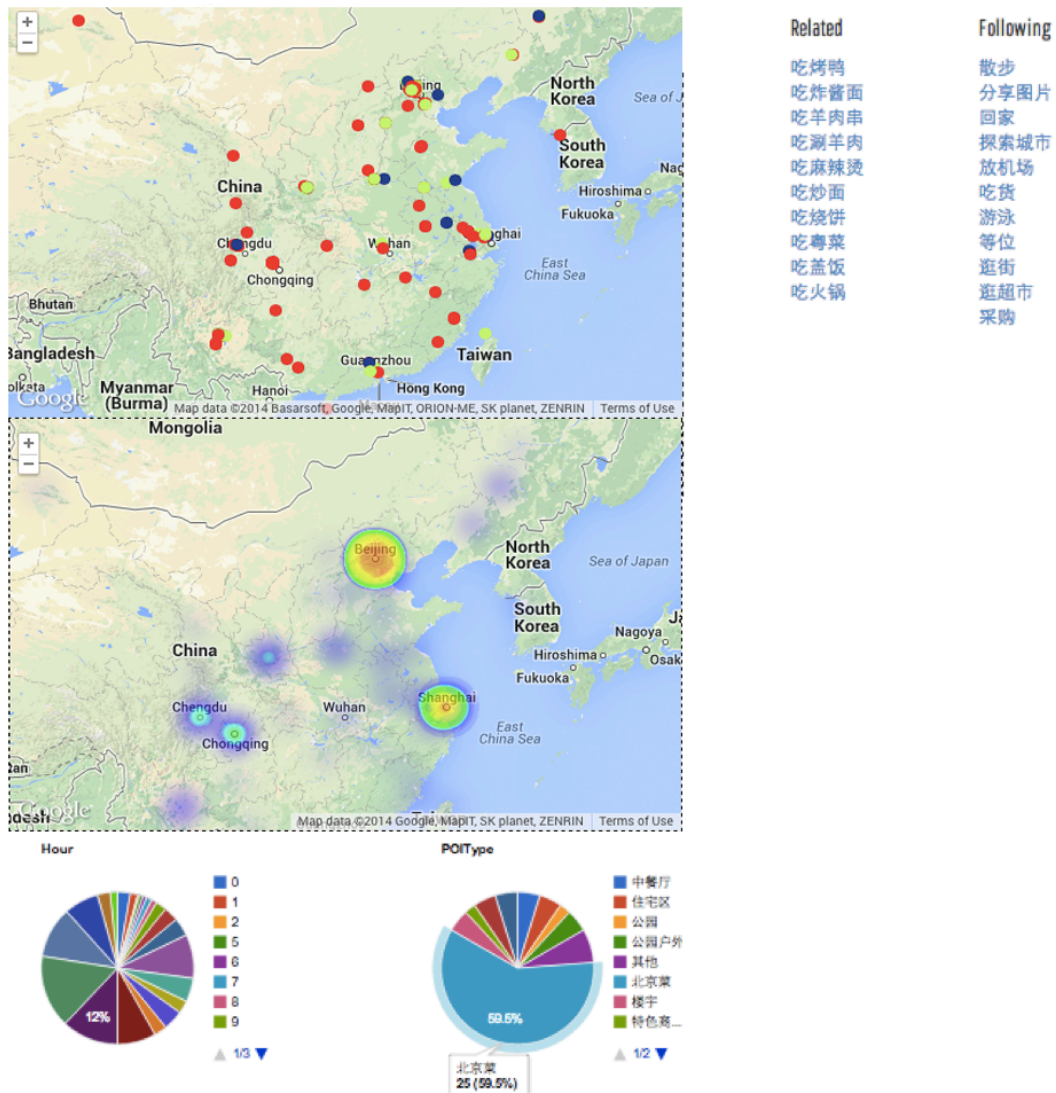


图 5.3 查询结果

果用户发现返回的结果不是活动，鼠标移上之后会呈现“×”按钮，点击后会向系统报告错误。

5.3.3 地点、情感、时间分布

基于地点抽取和情感分类的结果，本文利用 Google Map API 将地理分布和情感分布可视化。上图是用户在参加活动是情感状态的分布，一个圆点表示一个活动实例，红色为正面情感，蓝色为负面情感，绿色为中性。下图是这项活

动的在不同地域的分布情况，以不同颜色的弥散圆表示，红色的强度表示和这项活动相关的活动相关的微博数目。可以看到，“吃烤鸭”这项活动，在北京最为流行；在上海、成都、重庆等地也有着较多的分布。

进一步分析了这项活动发生的 POI 的类型。从 POIType 的饼状图中可以看到，“吃烤鸭”最常发生在“北京菜”这类 POI 中，其次还有为“中餐厅”。而在时间分布中，“吃烤鸭”最常发生的时间为 6 点。

5.4 本章小结

本章从底层架构到界面设计介绍了基于活动信息挖掘的可视化系统 ActivityNet。ActivityNet 基于 SAE(Social Analytic Engined) 实现了高效的存储和检索功能，并利用 gchart 和 Google Map API 实现信息可视化，实现了较好的用户体验。由于方法的局限，ActivityNet 呈现的结果有一些噪声，系统中提供了用户反馈机制，使得不正确的结果可以及时更正。

第 6 章 结论

6.1 工作总结

随着社交网络和移动互联网的蓬勃发展，人们的线下生活越来越多地反映到社交网络发布的信息中。通过对社交数据中日常活动信息的挖掘，我们不仅可以发现关于活动自身的许多知识，诸如地域偏好，时间分布，活动之间的关系，还可以了解到用户的个人喜好，行为模式，进一步对用户进行个性化推荐。活动挖掘是一个有着诸多潜在应用的研究方向。

社交数据以其内容的丰富性和及时性，为日常活动的信息挖掘提供了便利，也带来许多挑战。本文确定了三个研究内容：（1）如果精确地从社交媒体中抽取出日常活动的概念表示？（2）针对用户发布的信息，如何抽取出相关活动及相关属性，如地点、时间、情感？（3）如何构建活动之间的关系？对于第一个问题，本文将概念抽取问题化归为分类问题，通过词的语义向量表示作为特征，通过解优化问题选取了训练数据，进行分类，取得了较好的分类精度。第二个问题中，我借助句法分析、情感分析、信息抽取等方法，实现了活动实例的抽取。第三个问题中，本文基于实例抽取的结果，着重研究了活动之间的序列关系。

此外，基于以上三点研究，本文构建了 ActivityNet，实现了高效的活动检索，将本文的研究成果，进行直观的展示。

6.2 进一步工作

在社交数据中挖掘日常活动相关知识是一个新颖的问题，可以参考的工作不多，本文在这一课题上进行初步研究，取得了一定成果，但也存在一些局限性。在将来，本文的工作可以在以下方面进行改进：

1. 考虑用户在社交网络中的社交关系。本文的工作中，用户是孤立的，没有考虑到用户之间的互动。而利用用户的好友关系和网络结构可以提供可以帮助我们建模更复杂的活动关系，例如发现用户和其好友共同的活动，以及他们在活动中的互动等。这将是一个有趣的研究问题。

2. 对活动关系的建模比较粗糙。活动之间的关系是复杂多样的，本文主要关注活动的序列关系，而根据分类标准的不同，活动之间的关系是多样了。
3. 本文现阶段的工作，仅在社交数据上进行。如果借助外部知识库，如地点、人物、书籍信息，将对活动抽取和关系建模有很大的帮助。

希望本文的研究可以继续发展，更加深入地揭示社交数据与日常行为的关系，并投入实际应用。

插图索引

图 1.1	社交网络的崛起	1
图 2.1	序列标注模型	8
图 2.2	MaxEnt-LDA	12
图 2.3	Nguyen 的活动抽取框架.....	13
图 3.1	概念抽取框架	14
图 3.2	神经网络语言模型	20
图 3.3	问题归约	24
图 3.4	概念抽取初步结果	28
图 4.1	LTP 句法分析示例	30
图 4.2	分析速度	31
图 4.3	特征选择	32
图 4.4	模型选择	33
图 5.1	SAE 基本结构	39
图 5.2	Activity 首页设计.....	40
图 5.3	查询结果	41

表格索引

表 2.1	现有知识库概况	10
表 3.1	微博数据概况	15
表 3.2	ICTCLAS 分词结果示例	16
表 3.3	χ^2 检验	17
表 3.4	语义相似度样例	21
表 3.5	概念抽取实验结果	28
表 4.1	活动类别抽取实验结果	30
表 4.2	POI 样例	34
表 4.3	地点抽取实验结果	35
表 4.4	地点抽取结果样例	35

参考文献

- [1] Leskovec J, Backstrom L, Kumar R, et al. Microscopic evolution of social networks. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008. 462–470
- [2] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American society for information science and technology, 2007, 58(7):1019–1031
- [3] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009. 807–816
- [4] Tang W, Zhuang H, Tang J. Learning to infer social ties in large networks. Proceedings of Machine Learning and Knowledge Discovery in Databases. Springer, 2011: 381–397
- [5] Weber M. The Nature of Social Action. Cambridge University Press, 1991
- [6] Ciravegna D, et al. Adaptive information extraction from text by rule induction and generalisation. 2001. 1251–1256
- [7] McCallum A, Freitag D, Pereira F C. Maximum Entropy Markov Models for Information Extraction and Segmentation. Proceedings of ICML, 2000. 591–598
- [8] Lafferty J, McCallum A, Pereira F C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
- [9] Sarawagi S, Cohen W W. Semi-Markov Conditional Random Fields for Information Extraction. Proceedings of NIPS, volume 17, 2004. 1185–1192
- [10] Gauch S, Chaffee J, Pretschner A. Ontology-based personalized search and browsing. Web Intelligence and Agent Systems, 2003, 1(3):219–234
- [11] Snow R, Jurafsky D, Ng A Y. Learning Syntactic Patterns for Automatic Hypernym Discovery. Proceedings of NIPS, volume 17, 2004. 1297–1304
- [12] Navigli R, Velardi P, Faralli S. A graph-based algorithm for inducing lexical taxonomies from scratch. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three. AAAI Press, 2011. 1872–1877
- [13] Liu X, Song Y, Liu S, et al. Automatic taxonomy construction from keywords. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012. 1433–1441
- [14] Miller G A. WordNet: a lexical database for English. Communications of the ACM, 1995, 38(11):39–41
- [15] Ponzetto S P, Strube M. Deriving a large scale taxonomy from Wikipedia. Proceedings of AAAI, volume 7, 2007. 1440–1445

- [16] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002. 79–86
- [17] Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2008, 2(1-2):1–135
- [18] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002. 417–424
- [19] Zhao W X, Jiang J, Yan H, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010. 56–65
- [20] Titov I, McDonald R. Modeling online reviews with multi-grain topic models. Proceedings of the 17th international conference on World Wide Web. ACM, 2008. 111–120
- [21] The N M, Kawamura T, Nakagawa H, et al. Automatic Mining of Human Activity Attributes from Weblogs. Proceedings of Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on. IEEE, 2010. 633–638
- [22] Kawamura T, Nakagawa H, Nakayama K, et al. Human activity mining using conditional random fields and self-supervised learning. Proceedings of Intelligent Information and Database Systems. Springer, 2010: 140–149
- [23] Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008, 1997.
- [24] Bollegala D, Matsuo Y, Ishizuka M. Measuring semantic similarity between words using web search engines. www, 2007, 7:757–766
- [25] Liu Z, Li P, Zheng Y, et al. Clustering to find exemplar terms for keyphrase extraction. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009. 257–266
- [26] Hinton G E. Learning distributed representations of concepts. Proceedings of the eighth annual conference of the cognitive science society. Amherst, MA, 1986. 1–12
- [27] Bengio Y, Schwenk H, Senecal J S, et al. Neural probabilistic language models. Proceedings of Innovations in Machine Learning. Springer, 2006: 137–186
- [28] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
- [29] Hornik K. Approximation capabilities of multilayer feedforward networks. Neural networks, 1991, 4(2):251–257
- [30] Che W, Li Z, Liu T. Ltp: A chinese language technology platform. Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2010. 13–16
- [31] Ekman P. An argument for basic emotions. Cognition & Emotion, 1992, 6(3-4):169–200

致 谢

衷心感谢导师唐杰副教授的精心指导。他的言传身教将使我终生受益。

感谢陈麒聪、曹烨、杨洋同学，他们参与了我的毕业设计的工程中，并给了我很多帮助。在美国南加州大学进行暑期研修期间，承蒙 Ugur Demiryurek 教授热心指导与帮助，不胜感激。感谢 IBM 中国研究院的蔡柯柯研究员，和她的交流给了我许多灵感和想法，让我获益良多。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名： 王凝辉 日 期： 2014年6月17日

附录 A 外文资料的调研阅读报告

Online social networks, such as Facebook, Twitter, Weibo and Renren, have achieved great success in the past years. Now Facebook has over 1 billion users and Sina Weibo has about 59 million users. It's reported that people in US are spending their 16% online time on Facebook, even more than Google (10%). With the rapid growth of online social networks, information is produced at an amazing speed. More than 500 million tweets are sent everyday on average. people share their opinions about various events from the Olympics to commercial promotions. Each event consists of several aspects and people's interests and opinions are changing over time. Take the Olympics for example. The opening ceremony is mostly talked about in the first few days. Then with the event going on, different sport matches and athletes are talked about. When an athlete won a medal, people will cheer for him. However, several days later, he made a blunder and lost, people's opinion changed. Faced with the information overwhelming, we need a automatic way to track the event process and people's opinion. Given an event, we want to know, 1. What aspects does people interested in 2. How does people's interest and opinions evolve over time and 3. Who are the most influential users in the discussion of the topic. And finally a flow chart visualization is helpful to demonstrate the evolution process of the opinion.

Several classical research areas in data mining are closely related to the problem, such as opinion mining, topic modeling, event detection and tracking.

A.1 Opinion Mining

Opinion mining aims to find out people's opinion about an *entity* from corpus like product reviews or blogs. An entity can be a product, service, event, topic, anything that can be evaluated. An entity can be represented as a combination of different aspects (or features). For example, for an mobile phone (entity), screen, battery, memory are three aspects. Given a collection of documents D , the objective of opinion definition is, opinion mining aims to extract entities, aspects, associated opinions and analyse their

sentiment orientations. To get a high-level perspective of the whole corpus, an addition summary step is optional.

Aspect extraction is a fundamental step in aspect-based opinion mining and is closely related with our task. The first work in aspect extraction is a two-step unsupervised method^[1]. First, find frequent nouns and noun phrases. Then, find infrequent aspects by using the relationships between aspects and known opinion words. The method is based on the intuition that important aspects are talked about frequently and phrases which co-occur with opinion words often are likely to be aspects. Opinion words can be generated in two ways. The dictionary-based approach defines a seed set of opinion words and search their synonyms and antonyms in a dictionary WordNet. However, it fails to capture the characteristic that the same word can express different sentiment orientation in different domains. The corpus-based approach also uses a seed set. But it finds new words by exploiting syntactic or cooccurrence patterns in a large corpus^[2]. The aspect extraction and opinion words finding can reinforce each other. New-found opinion words provide hint for identifying new aspects, while new aspects result in more opinion words. So many following researches combine them in a unified framework.

More unsupervised and supervised algorithms were proposed since then. Supervised methods treat aspect extraction as a classification problem. CRF is used in^[3]. Jin et al.^[4] used a HMM based sequence tagging method to find entities and opinion words simultaneously. Su et al.^[5] proposed a clustering-based method to find hidden association of opinion words and aspects. Topic models can also be used in aspect extraction. However, as Titov^[6] pointed out, plain LDA is not suitable for aspect extraction because it can't distinguish global topics (like hotels in China, hotels in America) and local topics (aspects of hotels). He proposed a multi-grain LDA (MG-LDA) to solve that problem. MG-LDA models global and local topics at the same time. Another way to solve that problem is to run LDA on sentence level^[7]. But the two methods mess opinion words and aspects together. Zhao et al.^[8] proposed MaxEnt-LDA to model aspect and aspect-specific opinion words jointly.

To decide the sentiment orientation of an opinion, machine learning classification methods such as SVM and naive Bayesian can be used. However, lexicon-based

method can capture more subtle semantic elements of an opinion^[9]. For example, opinion shifters like negation words (not, never, none) and sarcasm, and but-clauses.

A.2 Evolutionary Topic Models

LDA, first proposed by Blei et al,^[10] has become the most popular topic modeling tool. In general, LDA is extended in three main directions. 1. Modeling more latent attributes like sentiment, social role, personal preference. 2. Incorporating inter-document relations. For example, relational LDA^[11] on a citation network and author-topic LDA^[12] on an author-document network. 3. Build time-aware topic model to model *topic evolution* in corpus, which we are most interested in.

The first try to model topic evolution with LDA is by Blei et al^[13]. In their model, time is discretized and topic distributions satisfy Markov attribute. They chain prior α and β in LDA with parameters in adjacent time slices using Gaussian distribution. A drawback of discrete-time model is we have to choose a suitable grain of time slices. In continuous dynamic topic model (cDTM)^[14], Brownian motion is used instead of Gaussian distribution. To get rid of the Markov assumption, Wang^[15] associated a Beta distribution of time with each topic to model topic popularity evolution, but topic distributions keep the same over time. Moreover, topic number is fixed over time. Topic birth and death are ignored. For example, at the dawn of artificial intelligence, areas like pattern recognition, natural language processing are rarely talked about explicitly. However, they're now so important that PR and NLP should be seen as separate topics. So a *split* of a topic happens. A non-parametric version of dynamic topic model called Infinite Dynamic Topic Models (iDTM) is given by Ahmed^[16]. iDTM models each document using a hierarchical Dirichlet process.

Visualization is a perfect way to demonstrate the evolution trend of topics. Liu composed several great visualization tools for rendering topic dynamics. TextFlow^[17] renders the whole corpus as a multi-branch flow. Each branch is a topic. Different branches can split or join each other, representing topic birth and death. And width of a flow indicates the popularity. It was further developed into an interactive analysing tool^[18].

A.3 Event Tracking and Burstiness detection

The task of Event tracking is about discovering temporal intensities of events in text streams such as weblogs or newswires. Event detection and tracking is closely related with dynamic topic modeling and share methods in common. However, event detection has its own features. First, topic evolution analysis is usually done off-line. We accumulate data over a time window and run algorithms on whole data set. While event detection is executed at the time new data comes, aka *on-line*. Second, topic model focuses on topics that are heavily talked about. While event detection emphasizes the *burstiness* more. A new event may not be talked about intensively now, but its occurrence rises rapidly in a short time.

Variations of LDA are used in event tracking. AlSumait^[19] proposes an on-line version of LDA to track the event. Ha-Thuc et al. proposed a relevance-based topic model for news event tracking^[20]. However, apart from LDA, many other models can be used. Leskovec et al.^[21] studied a novel problem of meme-tracking. meme is a quoted text which varies during spreading over the web. They adopted a clustering method to group variants of a meme together and analysis global and local intensity.

Term burstiness has been extensively researched as a mechanism to address new event detection. Kleinberg^[22] used a state machine to model the burstiness and inspired most following work. Lappas et al.^[23] further explore how burstiness can enhance document searching.

参考文献

- [1] Hu M, Liu B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004. 168–177
- [2] Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1997. 174–181
- [3] Jakob N, Gurevych I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010. 1035–1045
- [4] Jin W, Ho H H, Srihari R K. OpinionMiner: a novel machine learning system for web opinion mining and extraction. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009. 1195–1204
- [5] Su Q, Xu X, Guo H, et al. Hidden sentiment association in chinese web opinion mining. Proceedings of the 17th international conference on World Wide Web. ACM, 2008. 959–968
- [6] Titov I, McDonald R. Modeling online reviews with multi-grain topic models. Proceedings of the 17th international conference on World Wide Web. ACM, 2008. 111–120
- [7] Brody S, Elhadad N. An unsupervised aspect-sentiment model for online reviews. Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. 804–812
- [8] Zhao W X, Jiang J, Yan H, et al. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010. 56–65
- [9] Ding X, Liu B, Yu P S. A holistic lexicon-based approach to opinion mining. Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008. 231–240
- [10] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. the Journal of machine Learning research, 2003, 3:993–1022
- [11] Chang J, Blei D M. Relational topic models for document networks. Proceedings of International Conference on Artificial Intelligence and Statistics, 2009. 81–88
- [12] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents. Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004. 487–494

- [13] Blei D M, Lafferty J D. Dynamic topic models. Proceedings of the 23rd international conference on Machine learning. ACM, 2006. 113–120
- [14] Wang C, Blei D, Heckerman D. Continuous time dynamic topic models. arXiv preprint arXiv:1206.3298, 2012.
- [15] Wang X, McCallum A. Topics over time: a non-Markov continuous-time model of topical trends. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006. 424–433
- [16] Ahmed A, Xing E P. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. Proceedings of Uncertainty in Artificial Intelligence3. AUAI, 2010
- [17] Cui W, Liu S, Tan L, et al. Textflow: Towards better understanding of evolving topics in text. Visualization and Computer Graphics, IEEE Transactions on, 2011, 17(12):2412–2421
- [18] Liu S, Zhou M X, Pan S, et al. Interactive, topic-based visual text summarization and analysis. Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009. 543–552
- [19] AlSumait L, Barbará D, Domeniconi C. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. Proceedings of Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008. 3–12
- [20] Ha-Thuc V, Mejova Y, Harris C, et al. A relevance-based topic model for news event tracking. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009. 764–765
- [21] Leskovec J, Backstrom L, Kleinberg J. Meme-tracking and the dynamics of the news cycle. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009. 497–506
- [22] Kleinberg J. Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery, 2003, 7(4):373–397
- [23] Lappas T, Arai B, Platakis M, et al. On burstiness-aware search for document sequences. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009. 477–486

在学期间参加课题的研究成果

个人简历

1992 年 1 月 23 日出生于山东省枣庄市。

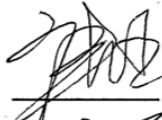
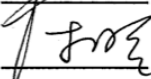
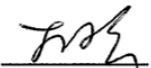
2010 年 9 月进入清华大学计算机科学与技术系计算机科学与技术专业攻读工学学士学位至今。2010 年获新生二等奖学金，2011 年获国家奖学金，2012 年获清华之友-董氏东方奖学金，2013 年获钟士模奖学金。另外，于 2013 年获得中国计算机学会颁发的 CCF 优秀大学生奖，以及 ASC13 亚洲大学生超级计算机竞赛冠军，ISC13 世界大学生超级计算机竞赛第二名。

参与的科研项目

[1] 国家自然科学基金项目：知识发现与知识工程（合同号：61222212）

[2] 华为公司资助研究项目：大规模社会网络分析引擎技术合作项目

综合论文训练记录表

学生姓名	王凝桦	学号	2010011281	班级	计 03
论文题目	社交网络中观点制导的事件追踪				
主要内容以及进度安排	<p>1. 基于事件的相关微博发现 3月22日前</p> <p>2. 观点 aspect-opinion 抽取 尝试序列标注的 CRF 模型 4月12日前</p> <p>3. 动态的观点和事件追踪 5月3日前</p> <p>4. 可视化完善修改 5月10日前</p> <p>5. 论文写作与提交 6月1日前</p> <p>指导教师签字: </p> <p>考核组组长签字: </p> <p>2014年 5月12日</p>				
中期考核意见	<p>开题报告. 进行中</p> <p>考核组组长签字: </p> <p>2014年 4月1日</p>				

指导教师评语	<p>论文主要研究社交网络中基于双气的事件分析, 包括事件抽象, 事件属性事件关系抽象, 以及基于事件抽象场的抽取方法. 并基于该方法实现了实验系统 activity Net. 论文内容饱满, 过程顺利! 祝顺利!</p> <p>指导教师签字: <u>李锐</u></p> <p>2014 年 6 月 日</p>
评阅教师评语	<p>论文主要研究社交网络中基于双气的事件分析, 包括事件抽象, 事件属性事件关系抽象, 以及基于事件抽象场的抽取方法. 并基于该方法实现了实验系统 activity Net. 论文内容饱满, 过程顺利! 祝顺利!</p> <p>评阅教师签字: <u>李锐</u></p> <p>2014 年 6 月 18 日</p>
答辩小组评语	<p>陈述清晰, 回答问题正确, 建议授予 288 分.</p> <p>答辩小组组长签字: <u>李锐</u></p> <p>2014 年 6 月 13 日</p>

总成绩: 96

教学负责人签字: 李锐

2014 年 6 月 18 日