文章编号：

基于依存句法分析的社会媒体文本挖掘方法

——以微博用户饮食习惯分析为例*

任彬，车万翔，刘挺

（哈尔滨工业大学 社会计算与信息检索研究中心，黑龙江 哈尔滨 150001）

摘要：在进行社会媒体文本挖掘时，传统的基于词表的方法，存在准确率较低、词表难获得等问题。该文提出一种基于依存句法分析的文本挖掘方法，通过规则匹配的方式从社会媒体文本中提取信息。该方法不依赖词表，且实验证明了相比基于词表的方法在准确率上有大幅提高。应用基于依存句法分析的文本挖掘方法，我们在微博文本上进行了饮食习惯分析，实现了性别、地区、时间等维度的饮食习惯分析并可进行交叉分析，最终用词云的方式展示了结果。

关键词：依存句法分析；文本挖掘；社会媒体；饮食习惯分析

中图分类号：TP391　　**文献标识码：**A

Dependency Parsing-Based Social Media Text Mining Method

—— a Case Study in Analysis of Weibo Users' Eating Habits

REN Bin，CHE Wanxiang，LIU Ting

（Research Center for Social Computing and Information Retrieval，
Harbin Institute of Technology， Heilongjiang Harbin 150001）

Abstract: When conducting social media text mining, the traditional lexicon method has the problem of lower accuracy and lexicons difficult to obtain. This paper proposes a dependency parsing-based text mining method, mining information from social media text using matching rules. This method can work without lexicons and the experiment results proved a substantial increase in accuracy compared to the lexicon method. Using the dependency parsing-based method, we conducted an eating habits analysis on the Weibo text and achieved results on gender, region, time, including cross-analysis results, finally we use word clouds to show our results.

Key words: dependency parsing； text mining；social media；eating habits analysis

1 引言

正经历爆发式增长的社会媒体数据对社会科学的影响越来越大[1][2]。通过挖掘社会媒体文本获得信息,相比于社会学研究中传统的问卷调查方式，有着更真实、数据量大、费用低等优点，因而越来越被广泛使用。通过社会媒体文本挖掘，可以预测一个人的性别、年龄、个性等[3]，甚至可以预测股票价格或是电影票房[4]。在本文中，我们则尝试挖掘新浪微博的内容文本，来进行微博用户饮食习惯的分析。

目前在社会媒体文本挖掘中，基于词表的方法使用得最为普遍。其本质是将待分析文本与给定词表中的词相匹配。比如，“鼻子”、“皮肤”、“手”等词语会被放进一个“身体”词表中，通过统计外向的人和内向的人谁的话中这些词出现得更频繁，就可以探究哪种人更常讨论“身体”这个话题[3]。基于词表的文本挖掘方法简单易用，应用广泛。LIWC (Linguistic Inquiry and Word Count) [5][6]，就提供了涉及词性、常见话题等不同方面的英文词表，得以让研究者利用不同词典，开展兴趣、情绪、思维方式、个体差异等方面的研究[7]。

* **收稿日期：**　　**定稿日期：**

基金项目：国家重点基础研究发展计划（973 计划）项目（2014CB340503）；国家自然科学基金面上项目（61370164）；国家自然科学基金重点项目（61133012）

然而，这种基于词表的文本挖掘方法有较明显的缺点。只基于词表，相当于只应用词本身的信息，而不考虑词的多义性和其在句子中有上下文时的特定含义。这样就会使得结果混入较多噪声，准确率较低。比如，“苹果”这个词既有可能指食品苹果，也可能指苹果手机。当利用微博文本研究饮食习惯时，如果简单地应用基于词表的方法，一旦出现词表中的某个词就算作一次饮食行为。那么，如果食品词表包含“苹果”，就会把谈到苹果手机的微博也算作吃苹果出现一次。

另一方面，中文的自然语言处理（NLP）技术实际分为分词、词性标注、句法分析等多个层次。基于词表的文本挖掘方法只应用词本身的信息，相当于只用分词层次的结果，词性以及句法分析信息都没得到有效利用。而 NLP 技术的发展，已经使得词性标注、句法分析等技术相当成熟且容易使用。句法分析就已经广泛用于机器翻译、自动问答、信息抽取等应用。

因此，我们提出了基于依存句法分析的文本挖掘方法，尝试把词性标注、依存句法分析技术等深层 NLP 技术应用到对社会媒体文本的挖掘上，使得对社会媒体文本的分析更加准确有效。这种方法在对微博文本进行分词处理的基础上，进一步进行词性标注和依存句法分析，然后根据任务需求设定具体的一个或一系列规则，来挖掘文本语料中的信息。还是用刚才关于苹果的例子来说明这种方法能带来的进步和好处。当进行了词性标注和依存句法分析以后，可以用触发词“吃”和动宾搭配的规则过滤出真正吃苹果的行为。因为，很明显，如果你提到的是苹果手机，你肯定不会说“我吃了苹果”。基于依存句法分析的文本挖掘方法，就是用这样的方式，利用更多的上下文信息，减少对文本内容的误读，提高数据利用的准确性。我们还设计并进行了实验，证明了在社会媒体文本挖掘上，基于依存句法分析的方法，的确比基于词表的方法准确率更高。

利用这种基于依存句法分析的文本挖掘方法，我们进行了微博用户饮食习惯分析。做法是对微博文本的依存句法分析结果，通过设定特定的识别规则，从中分析出每条微博是否描反映了真实的饮食行为，如果确实反映饮食行为，相应的食品是什么。再把微博相应饮食行为的食品与微博本身的属性，如发微博时间、发微博人的性别、地区等对应起来进行分析，就能得到关于不同性别、不同地区、不同时间段的饮食习惯。

本文的贡献主要在于：（1） 提出了一种基于依存句法分析的方法，能更准确地进行社会媒体文本挖掘（第 3 节）。（2）将这种基于依存句法分析的文本挖掘方法与基于词表的文本挖掘方法进行了实验对比，证明了前者在准确率上有显著提高（第 4 节）。（3）用基于依存句法分析的文本挖掘方法，对社会媒体新浪微博上的文本，进行饮食习惯分析，获得了不同性别、不同地区、不同时间段的饮食习惯。这是用社会媒体文本进行社会信息挖掘的一种新的尝试（第 5 节）。

2 背景

2.1 基于词表的文本挖掘方法

在用基于词表的方法进行文本分析时，使用最广泛的就是 Linguistic Inquiry and Word Count （LIWC）[5][6]。2007 年版本的 LIWC，包含了将近 4500 个词，这些词被 64 个不同的类别组织起来，即提供了 64 个词表，如其中包括涉及情感倾向性分析的积极情绪（positive emotion）词表和消极情绪（negative emotion）词表。

当一个研究者，想要了解一段文本是有积极情绪的倾向还是消极情绪的倾向时，只需把待分析文本输入给 LIWC 工具，它就能统计出这段文本中词语分属于两个词表的比例，进而确定这段文本的情感倾向性。如果属于积极情绪词表的词语比例高，则文本倾向积极；反之则倾向消极。

所以，基于词表的文本挖掘方法本质上就是通过将待分析文本与给定词表进行匹配，进而获得信息。

2.2 依存句法分析

由于我们提出的是基于依存句法分析的文本挖掘方法，因而有必要阐述依存句法分析的基本概念。

比如对句子“我刚才吃了一个苹果。”进行依存句法分析的结果如图 1 所示：

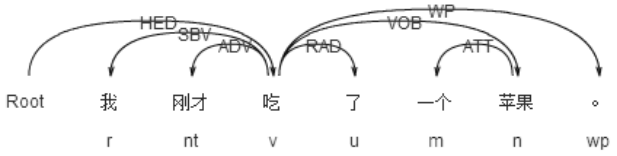

图 1 依存句法分析结果示例

依存分析的结构中，词与词之间直接发生依存关系，构成一个依存对。一个依存对的两个词中，其中一个是核心词，也称为支配词；另一个是修饰词，也称为从属词。依存关系用一个有向弧表示，称为依存弧。在本文中，规定依存弧的方向为由从属词指向支配词。在上图中，每个依存弧上有一个标记，叫做关系类型，表示该依存对中的两个词之间存在什么样的依存关系[8]。本文研究中在依存句法分析过程中使用了哈工大 LTP 开源工具[9]提供的依存句法分析功能，因而依存关系的规定与划分也就默认遵循哈工大 LTP 所使用的依存规则。

2.3 利用社会媒体文本挖掘饮食习惯

据我们所知，目前尚未有基于社会媒体文本对饮食习惯进行分析的研究。但有一些基于社会媒体数据挖掘社会信息的研究，与我们的研究贴近。Golder 利用 Twitter 数据探究工作、睡眠、昼长对个体情绪的影响，发现了人们在周末更开心[10]。Dodds 利用社会媒体文本分析研究社会层面的幸福感[11]。Hannak 利用大量 Twitter 上的文本数据，研究了天气和时间对群体情绪的影响[12]。这些工作也都是利用社会媒体文本进行社会学统计信息的挖掘。

3 基于依存句法分析的文本挖掘方法

3.1 识别规则

首先，对这些微博文本进行分词、词性标注和依存句法分析。接下来我们需要利用依存句法分析结果判断微博是否反映了真实的饮食行为。我们使用了利用规则来匹配的方法。我们都知道，当谈到饮食行为时，“我吃/喝了某种食品”是最常见的句式。

所以，当给定一条微博，要判断是否反映了真实的饮食行为时，我们可以对微博内容的句法分析结果应用这样一条简单规则：

含词语“吃”/“喝” 且 *以“吃”/“喝”为支配词的句法关系为动宾关系（VOB）* 且 *“吃”/“喝”的宾语为名词（n）*

以此规则来过滤。如果微博内容符合这个有三个条件的规则，则判定其反映了真实的饮食行为，且提取出来的“吃”/“喝”的宾语就是饮食行为所对应的食品。

比如对“我刚才吃了一个苹果”这句话，句法分析结果如图 1 所示。其符合：含“吃”；以“吃”为支配词的句法关系为 VOB 关系；“吃”的宾语“苹果”词性为名词。我们就可以说这条微博反映了真实的饮食行为，且其对应的食品是“苹果”。

而对于“新一代苹果手机即将亮相。”这句话，由于其中不含“吃”或“喝”，更无以“吃”或“喝”为支配词的 VOB 关系，明显不符合我们设定的规则，因而不会被算作饮食行为。所以用我们设定的规则进行匹配，可以从语义上过滤掉那些“苹果”并非以食品的义项出现的微博。

再比如“我喝了一瓶苹果味汽水。”这个句子。在基于词表匹配的文本挖掘方法中，只要食品词表中包含“苹果”和“汽水”，就会导致“苹果”和“汽水”都被当做饮食行为涉及的食品各计算一次。但实际上这里谈到的是喝“汽水”，并非吃“苹果”。通过句法分析分

析，可以得到结果如图 2 所示：

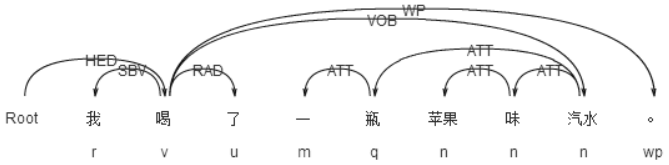

图 2 与吃苹果无关微博的句法分析结果

用我们设定的规则可以很容易地得到这句话描述的饮食行为所对应的真实食品“汽水”，而不会把“苹果”也误算一次。这也是利用上下文信息来理解词语在句子中的实际含义。

3.2 特色饮食

首先，我们先要界定好“饮食习惯”的概念。在我们的分析中，“饮食习惯”主要指既有一定规模、又要有特色的食品。举例来说，想获得北京地区的饮食习惯，那么即使北京地区最常吃的是“饭”、最常喝的是“水”，这也不能算作北京地区的饮食习惯，因为可能整个中国都在吃“饭”喝“水”，“饭”和“水”并不能体现北京地区饮食习惯的特色。

为了满足我们对“饮食习惯”的限定，我们引入互信息值（PMI）进行评价：

$$\mathrm{PMI}(word, category) = \ln \frac{p(word, category)}{p(word) * p(category)} \qquad (1)$$

其中，*word* 指食品词汇；*category* 指类别，类别可以是地区、性别、时间，如北京市或是男性等，也可以是交叉条件，如北京市男性等；$p(word, category) = \frac{count(word, category)}{total\_count}$ 指食品 *word* 与类别 *category* 共现的概率；$p(word) = \frac{count(word)}{total\_count}$ 指食品 *word* 出现的概率；$p(category) = \frac{count(category)}{total\_count}$ 指类别 *category* 下微博的出现概率；*count*(*word*)指提取出来的与食品 *word* 有关的饮食行为微博的数量；*count*(*category*)指提取出来的类别 *category* 下的含饮食行为微博的数量；*total_count* 指提取出来的含饮食行为微博数量的总数。

$\mathrm{PMI}(word, category)$ 就代表食品 *word* 在类别 *category* 下的特色程度。比如：PMI（烤鸭，北京）就代表烤鸭在北京地区的特色程度；PMI（烤鸭，男）就代表男性饮食习惯中烤鸭的特色程度；而 PMI（烤鸭，晚上）代表晚上饮食习惯中烤鸭的特色程度。要表示某类别的饮食习惯，只需取与该类别 PMI 值最高的数个食品词语即可。

PMI 同样也可以表示交叉条件下的饮食习惯。比如，PMI（烤鸭，男 and 北京）表示烤鸭在北京男性饮食习惯中的特色程度。而在实际操作中只需令 *category* 满足性别男且地区是北京市，即北京市男性所发饮食微博即可。

4 实验

为了对比基于词表的方法和基于依存句法分析的方法在社会媒体文本挖掘上的效果差别，我们设计了实验，将两者在微博用户饮食习惯分析任务上的表现相对比。在本文所进行的实验中，用到的分词工具均是面向微博语料的分词工具[14]，以期获得更好的分词效果。

4.1 基于词表的饮食习惯分析

将基于词表的文本挖掘方法应用在本文所提出的饮食习惯分析任务上的具体做法，就是

利用待分析文本与已有的食品词表相匹配，当文本中出现词表中的某个词时，就认为发生了一次关于这个词的饮食行为。

可以看出，影响这种方法效果的一个重要因素就是词表的质量。为了使得对比实验真正有意义和有说服力，我们建立了一个质量较高的词表。首先我们获得了百度百科截止 2012 年的全部词条数据 500W 条，并以此为基础提取词表。百度百科是最大的在线中文百科，但有个需要我们考虑的特点，就是其词条标签是开放性的，也就是说所有编辑者都可以为某个词条添加某个标签。因此，通过“食品”单个标签来过滤词条获得词表就会效果较差。因为有时某个食品的词条恰好就会没有“食品”这一标签，而是有“饮食”等其他标签。所以，我们先通过人工筛选的方式，获得 500 个食品词汇，再通过提炼这 500 个食品词汇的所有标签，构成一个与饮食相关的标签候选集，对这个标签候选集再进行人工筛选后，获得与饮食相关的标签集。如果一个词条只含一个标签集中标签，就把词条对应的词算作食品，则会引入较多如“哈尔滨工业大学食品学院”之类的噪声。因而只有当某个词条含有 2 个以上在饮食标签集中的标签时，我们才认为该词条是食品。用这样的方式，对百度百科 500W 词条数据进行筛选，我们获得了一个有 76,754 词大小的食品词表。这个食品词表整体质量较高，但也掺杂有少量与饮食相关、但非食品的词语。

在获得食品词表以后，对每条经过分词处理后的微博文本，我们用饮食词表进行匹配，一旦词表中的某个食品词在微博中出现，我们就认为这条微博对应了一次该食品的饮食行为。

4. 2 对比实验

要评价两种方法效果的差别主要有两方面的困难。一方面是没有标准测试集，需要人工标注数据；另一方面与饮食行为相关的微博在所有微博中所占比例很低，对所有微博进行标注则标注工作量过大。受限于此，Schwartz 在对基于词表的文本挖掘方法进行评价时，就只考虑了准确率的指标，而没考虑召回率[15]。

对此，我们采取的办法是，随机抽取 10 万条微博，用这两种方法分别识别反映饮食行为的微博，将两种方法识别的结果合并作为候选集。再由三个人对候选集进行人工标注，判断结果是否正确，进而获得标准结果集。在本实验中两种方法识别出的记录总数即候选集大小为 3,371 条，因而只需要对这些记录进行人工标注，而不是对原始的 10 万条微博，这就使得标注和评价变得可行。并且，在这样的评价方法下，我们也能够计算召回率。

人工标注的具体任务为每次给定包含一个词语和对应原始微博的词语-微博对，标注人员要判断这条微博是否反映了真实的饮食行为和这个词语是否是饮食行为对应的食品。只有正确识别出饮食行为和对应食品，才算识别正确。需要多人标注的原因是，有些微博很难判断究竟是否发生了饮食行为，比如“我买了一个苹果回家吃”这种句子，需要进行简单的推断，判断发生饮食行为和没发生均有一定道理，因而需要多人标注。

一共有三名标注人员分别对候选集进行人工标注。为了评价标注结果的一致性，我们计算了用于统计多类多标注人员标注一致性的 Fleiss Kappa 指标[13]，最终三人标注一致性为 75.53%。获得人工标注数据后，对三人标注有差异的数据用投票的方法确定结果。

4. 3 实验结果及分析

用人工标注结果获得标准结果集后，即可评价基于词表的文本挖掘方法和基于依存句法分析的文本方法在候选集上的表现，如表 1 所示：

表 1 两种方法的准确率、召回率、F 值

	准确率（%）	召回率（%）	F 值（%）
基于词表的方法	21.45	63.92	32.12
基于依存句法分析的方法	41. 01	55.08	47.00

可见，基于依存句法分析的文本挖掘方法，在准确率上要比基于词表的文本挖掘方法显著提高，但召回率略低，F 值也有大幅提升。

而基于词表的文本挖掘方法召回率虽然高于基于依存句法分析的方法，但也并不是很高的原因，并非食品词表质量不好，而是无论构建多大的食品词表，考虑到日常生活中的食品种类和说法之多，我们都很难穷尽食品词语。比如，百度百科拥有数百万词条，却尚未收录“甜筒”。日新月异的食品种类和新的称呼，也使得食品词表即使构建得很大，也很难达到很高的覆盖率。而基于句法分析的方法，则不受限于词表的限制，可以识别出关于新食品或食品新表达的饮食行为。

并且，在微博用户饮食习惯分析中，准确率其实比召回率更重要。因为我们通常可以获得大量的微博文本数据，这时只要有较高的准确率，即使召回率较低，通过足够的数据量，也能正确地挖掘出饮食习惯。而如果是召回率较高、但准确率较低的方法，就相当于在饮食习惯统计中掺入了较多错误结果带来的噪声，虽然符合条件的数据多了一些，但结果却没有说服力。

所以，通过实验评价和分析，我们可以说，使用基于依存句法分析的文本挖掘方法，相比于基于词表的方法，能够更准确地挖掘文本的真实含义，在本文的饮食习惯分析中则体现为能更准确地识别一条微博是否反映了真实的饮食行为。

5 饮食习惯分析结果

应用上文介绍的基于依存句法分析的社会媒体文本挖掘方法，我们对大规模微博文本数据进行处理，以获得饮食习惯分析结果。

5.1 数据集

我们随机爬取了新浪微博 5 千万条，时间跨度为 2009 年至 2011 年。使用这部分数据的原因是 2009 年至 2011 年新浪微博刚刚兴起，虚假用户较少。而现在微博上虚假用户及其产生的微博数量则大大增加。如何识别真实用户本身就是一个研究问题，但并不是我们要研究的重点。我们选用这段时间内的新浪微博数据，以便较大限度地剔除虚假用户对我们研究结果的影响。每条微博，我们除了微博本身的内容，还获得了微博发布的时间，以及发微博用户的性别、地区信息。

对这些微博文本应用第 3 节中的规则匹配，最终获得了与饮食相关的记录 45 万余条。再将识别出的对应食品与原微博的性别、地区和时间属性结合起来，即可计算出每个类别下的饮食习惯。

5.2 可视化

由于我们对饮食习惯的研究涉及多个维度，且每个维度下的结果是由多个食品词语组成，只用列表的方式展现就显得有些不够直观。我们选择用词云（word cloud）的形式展示我们的研究结果。在一般的词云使用中，词云中词语的大小只是由词语的频率决定。而我们则用词语大小来展示 PMI，即这个词语与该类别的相关性大小，用颜色来表示词语的频率。在结果的展示过程中，我们对明显的错误予以了过滤，最终在词云中展示的词语是 PMI 高的食品词语。

5.3 部分结果展示

5.3.1 性别维度下的结果

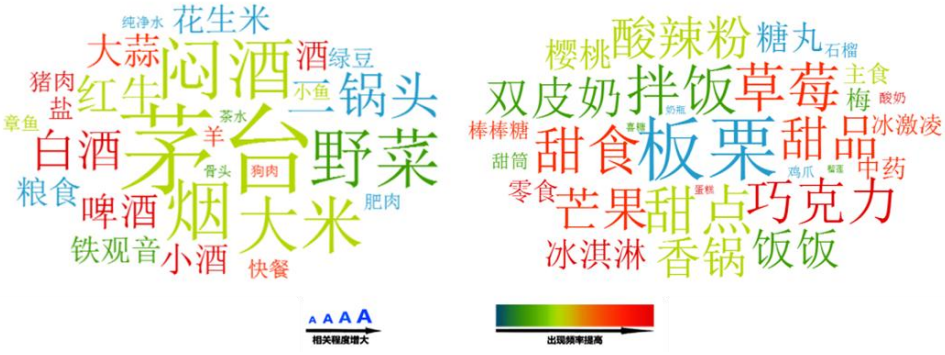

图 3 男性饮食习惯（左）与女性饮食习惯（右）对比

可以看到，不同性别的饮食习惯有很大区别。如男性的特色食品有茅台、啤酒、二锅头等，以酒类为主；女性则偏好巧克力、冰淇淋、甜食、芒果等食品，这比较符合我们的常识认识。

5.3.2 地区维度下的结果

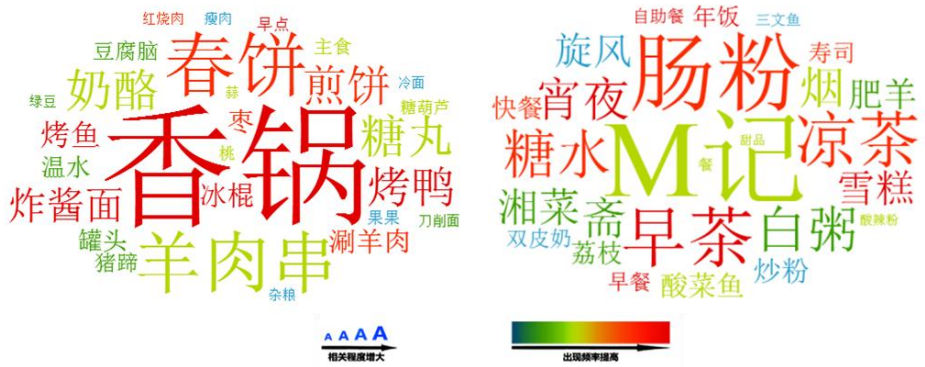

图 4 北京市饮食习惯（左）与广东省饮食习惯（右）对比

可以看到，不同地区的饮食习惯也有很大区别。偏南的广东省的饮食与偏北的北京市距离很远，饮食习惯差别也很大。香锅、烤鸭、春饼等都是北京著名特色食品且在而北京很常见。对于广东省的结果，M 记是对麦当劳的别称，从麦当劳中国官网上，我们也可以看到广东是麦当劳门店数最多的省份。

5.3.3 时间维度下的结果

为了更直观地展现时间维度的结果，我们将一天划分为四个时间段。6:00~10:59 为早上/上午；11:00~13:59 为中午；14:00~17:59 为下午；18:00~次日 5:59 为晚上。

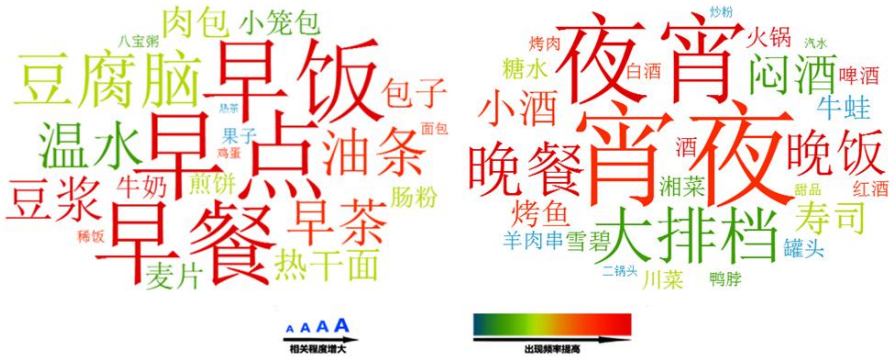

图 5 早上/上午饮食习惯（左）与晚上饮食习惯（右）对比

时间维度下的结果，也可以很好地反映饮食习惯。比如晚上时间段的宵夜、夜宵、烤肉、啤酒等，确实能反映晚上的饮食习惯特色；而在早上/上午，除了早饭、早点、早餐外，豆浆、油条、包子等也主要为早餐食品，与我们的日常认识很接近。

5.3.4 交叉条件下的结果

除了上文提到的三个维度，我们的分析还能得到交叉条件下的饮食习惯结果。比如可以查看北京市男性晚上的饮食习惯，也可以分析北京市女性晚上的饮食习惯。也就是说，我们可以分析出性别、地区、时间这三个维度交叉所可能形成的所有特定群体的饮食习惯。

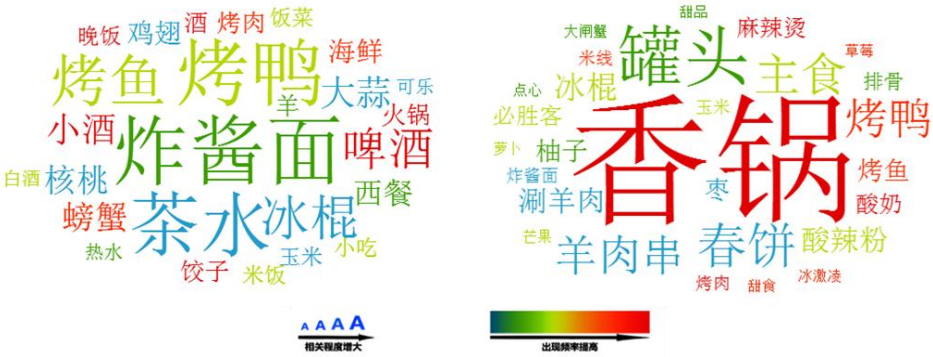

图 6 北京市男性晚上的饮食习惯（左）与北京市女性晚上的饮食习惯（右）

6 结论

我们提出了一种基于依存句法分析的文本挖掘方法，能更准确地挖掘社会媒体文本中的信息。并应用这种方法，从性别、地区、时间三个维度对微博用户的饮食习惯进行分析和交叉分析，用词云的形式可视化地展现了结果。实验也证明了在社会媒体文本挖掘上，基于依存句法分析的方法的确要比基于词表的方法有更高的准确率，因而能获得更有说服力的饮食习惯分析结果。并且，基于依存句法分析的方法，可以不受限于词表内的食品进行饮食行为的识别和食品的提取，甚至可以识别出新食品或是食品的新说法。

同时，用微博语料分析特定群体的饮食习惯，也有着重要意义。用传统的问卷调查等方法，很难获得关于饮食习惯的有效结果，但应用我们的方法，可以得到有一定说服力的结果。并且，我们经过分析获得的关于特定群体的饮食习惯结果，不仅是社会信息的统计结果，还可以进一步应用于为食品企业或餐饮行业的细分市场营销提供信息等方面。

当然，我们也注意到了用微博文本进行饮食习惯的挖掘，所获得的结果，会与现实有一

定偏差。这是由于微博数据相对于真实社会的偏置所造成。我们的工作，目前只限于尽可能准确地理解微博文本内容，使分析结果更贴近微博的真实含义。而对于微博数据和真实社会之间的偏置，还有待进一步研究。

接下来，我们进一步的研究工作主要有两方面：

一方面，用基于依存句法分析的方法，其实还可以细化规则，从而更准确地识别饮食行为。按照目前的规则，他人的饮食行为，比如“他吃了一个苹果。”也被算作发微博的人的饮食行为。通过细化规则，可以设定当主语不是“我”时不识别为饮食行为，就能过滤掉这种错误。另外，还有类似“我没吃饭”这样的否定句或者疑问句，也可以用通过细化规则如限制“吃”的修饰语挖掘出真实含义并处理，从而较少错误。

另一方面，我们目前设定的规则只有一条，只是匹配单一的由三个条件组成的规则。本文证明了，即使只用这一个最简单的规则，我们的方法也比基于词表的文本挖掘方法在准确率上有大幅提高。但实际上，还可以设定更多规则从文本中挖掘信息。比如，针对饮食行为的识别，除了“我吃/喝了某种食品”,“某种食品很好吃/好喝”也很常见。扩充规则的方法，可以人工制定，也可以用机器学习的方式进行扩充。通过扩充规则，可以进一步提高依存基于句法分析方法进行社会媒体文本挖掘时的召回率，这也是我们未来的一个研究方向。

参考文献

[1] Miller G. Social scientists wade into the tweet stream[J]. Science, 2011, 333(6051): 1814-1815.

[2] Lazer D, Pentland A S, Adamic L, et al. Life in the network: the coming age of computational social science[J]. Science (New York, NY), 2009, 323(5915): 721.

[3] Schwartz H A, Eichstaedt J C, Kern M L, et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach[J]. PloS one, 2013, 8(9): e73791.

[4] Asur S, Huberman B A. Predicting the future with social media[C]//Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. IEEE, 2010, 1: 492-499.

[5] P Pennebaker J W, Francis M E, Booth R J. Linguistic inquiry and word count: LIWC 2001[J]. Mahway: Lawrence Erlbaum Associates, 2001, 71: 2001.

[6] Pennebaker J W, Chung C K, Ireland M, et al. The development and psychometric properties of LIWC2007[J]. Austin, TX, LIWC. Net, 2007.

[7] Tausczik Y R, Pennebaker J W. The psychological meaning of words: LIWC and computerized text analysis methods[J]. Journal of Language and Social Psychology, 2010, 29(1): 24-54.

[8] 李正华. 依存句法分析统计模型及树库转化研究[D]. 哈尔滨工业大学硕士学位论文，2008.

[9] Che W, Li Z, Liu T. Ltp: A chinese language technology platform[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2010: 13-16.

[10] Golder S A, Macy M W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures[J]. Science, 2011, 333(6051): 1878-1881.

[11] Dodds P S, Harris K D, Kloumann I M, et al. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter[J]. PloS one, 2011, 6(12): e26752.

[12] Hannak A, Anderson E, Barrett L F, et al. Tweetin'in the Rain: Exploring Societal-Scale Effects of Weather on Mood[C]//ICWSM. 2012.

[13] Fleiss J L. Measuring nominal scale agreement among many raters[J]. Psychological bulletin, 1971, 76(5): 378.
[14] Liu Y, Zhang M, Che W, et al. Micro blogs Oriented Word Segmentation System[J]. CLP 2012, 2012: 85.
[15] Schwartz H A, Eichstaedt J, Dziurzynski L, et al. Choosing the Right Words: Characterizing and Reducing Error of the Word Count Approach[J]

作者简介：

任彬（1990—），男，硕士研究生，主要研究领域为自然语言处理。
E-mail：bren@ir.hit.edu.cn

车万翔（1980—），男，副教授、博士生导师，主要研究领域为自然语言处理、计算社会语言学。
E-mail：car@ir.hit.edu.cn

刘挺（1972—），男，教授、博士生导师，主要研究领域为中文信息处理、社会计算、信息检索。
E-mail：tliu@ir.hit.edu.cn