文章编号：1003-0077(2015)06-0193-10

# 基于评论挖掘的药物副作用发现机制

赵明珍，程亮喜，林鸿飞

（大连理工大学，计算机科学与技术学院，辽宁 大连 邮编 116024）

**摘 要**：从医疗社交网站的用户评论中挖掘药物副作用时，由于人们可能采用不同的表述方式来描述副作用，而新药的上市与用药者的差异性也会造成新的副作用出现，因此从评论中识别新的副作用名称并进行标准化十分重要。该文利用条件随机场模型识别评论中的副作用，对识别出的副作用名称进行标准化，最后得到药物的副作用。通过将挖掘出的药物已知的副作用与数据库记录进行对比验证了本文方法的有效性，同时得到一个按评论中的发生频率排序的药物潜在副作用列表。实验结果显示，条件随机场模型可以识别出已知的与新的副作用名称，而标准化技术将副作用名称进行聚合与归并，有利于药物副作用的发现。

**关键词**：药物副作用；用户评论；文本挖掘；实体标准化

**中图分类号**：TP391 **文献标识码**：A

## Detection of Adverse Drug Reactions Based on Comment Mining

ZHAO Mingzhen, CHENG Liangxi, LIN Hongfei

(School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China)

**Abstract**: When mining adverse drug reactions (ADRs) from the user comments on healthcare social networks, it is very important to recognize novel ADR expressions from comments and normalize them, since people probably adopt different expressions to describe adverse reactions and new adverse reactions may emerge with the listing of new drugs as well as the diversity of drug users. This paper utilizes Conditional Random Field (CRF) model to recognize adverse reaction entities, and proposes a normalization method applied to the recognized entities. The effectiveness of this mining method is verified by comparing the mined results of known ADRs with database records, and a list of potential ADRs sorted by occurrence frequency in comments is obtained. Experimental results indicate that CRF model is capable of identifying both known and novel adverse reaction entities, and the standardization aggregates and merges the entities, which benefits the ADR discovery.

**Key words**: adverse drug reaction; user comment; text mining; entity normalization

## 1 引言

随着 Web 2.0 技术的发展，互联网上出现了社区、论坛、博客、微博、Wiki 等各种形式的用户生成内容（User－Generated Content，UGC），它们极大地丰富了网络，并扮演着越来越重要的角色，这其中包括用户对药物的评论。另一方面，药物副作用（Adverse Drug Reaction，ADR）带来的危害越来越大，由它引起的病患占据所有医院病患的 5%，药物副作用已成为导致医院死亡的第五大原因[1]。药物副作用逐渐成为医学界和民众关注的热点，如何判断和预测药物的副作用具有重大的理论和实用价值。近年来互联网上出现的医疗健康类的社交网站与论坛积聚了大量来自用户的用药体验与评论，其中蕴含的最新副作用信息日益受到人们的重视，并逐渐形成从用户评论中挖掘药物副作用的研究方向。

Leaman 等人[2]通过计算滑动窗口中的评论内容与词典中副作用名称之间的相似度进行实体识别，对识别出的副作用名称进行过滤后挖掘药物的副作用，在人工标注的数据集上识别的 F 值为 73.9%。

收稿日期：2015-07-21 定稿日期：2015-09-25

基金项目：国家自然科学基金(661572102，61277370)；辽宁省自然科学基金(201202031，201402003)

Chee 等人[3]利用用户对药物的评论信息评估药物的安全性。由于正负例数量不均衡，他们采用 Bootstrapping 方法增加正例，并融合多个分类器对药物进行分类，预测可能将被监管或召回的药物。Nikfarjam 等人[4]采用关联规则从已标注的评论中挖掘副作用口语化表述的潜在模式，并利用这些语言模式从用户对药物的评论中自动抽取副作用，在测试集上测试得到的 F 值为 67.96%。Yang 等人[1]利用用户健康词汇表(Consumer Health Vocabulary，CHV)[5]构建了一个扩展的副作用名称词典，采用滑动窗口从关于药物的帖子中识别副作用名称，并使用关联规则挖掘药物的副作用，对于五种指定药物的实验得到较好的效果。Wu 等人[6]提出生成式和判别式两种方法来对文本中的药物副作用进行挖掘，实验结果表明网络中关于药物副作用的讨论内容可用于未知副作用的监测，而生成式模型方法在准确率与召回率两方面均比判别式方法更有效。

由于语言表述的自由性与多样性，人们在表达同一个副作用概念时可能会采用不同的措辞方式，而新药的上市以及用药者的差异性又可能会导致新的副作用出现，故而用户评论中会存在数据库未收录的副作用名称，有的甚至是因拼写错误而造成的不同。因此从评论中挖掘药物的副作用时，识别新的副作用名称并将其映射到统一的副作用概念上是十分重要的，否则将无法发现一些潜在的副作用，或挖掘出的副作用发病率与事实存在偏差。以前的工作在对药物副作用进行识别时，或者利用了滑动窗口与词袋模型，或者通过副作用口语化表述的模式来识别，这些方法对新副作用名称的识别效果往往不够理想；另外他们对识别出的新副作用名称的后续处理也很有限，影响药物副作用的发现。针对这些方法的不足，本文采用条件随机场(Conditional Random Field，CRF)模型识别副作用名称，可以有效识别出已知的以及新的副作用名称；对于识别得到的副作用名称，我们将其标准化并映射到已知的副作用概念上。

## 2 数据与方法

本文从用户评论中挖掘药物副作用的整个流程分为数据准备、副作用实体识别与过滤、副作用实体标准化、药物副作用发现四个部分，如图 1 所示。我们以 DailyStrength 网站[7]上用户对药物的评论作为语料，利用 CRF 模型识别出其中的副作用实体，然后使用本文提出的副作用实体标准化方法对识别出的副作用名称进行标准化，将其映射到统一的副作用概念上，最后统计每种药物评论下副作用概念的发生频率进而挖掘出药物的副作用。

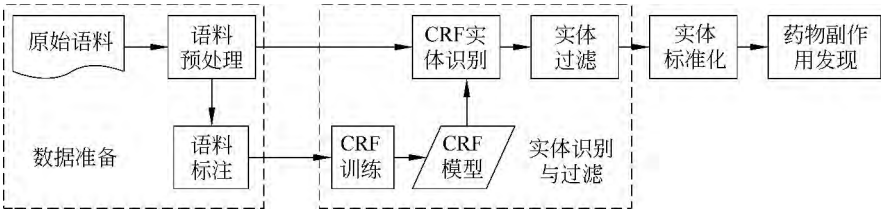


图 1 药物副作用挖掘系统架构图

### 2.1 数据准备

本文利用 SIDER 数据库[8]中的药物副作用数据创建了一个副作用词典，其中共包含 5 719 个副作用概念。每个副作用概念拥有一个统一医学语言系统(Unified Medical Language System，UMLS)的概念编号 CUI (UMLS Concept Id)，并由含义相同的一种或多种副作用名称构成。例如，CUI 为 C0239739 的概念有[sore gums，gum pain，gingival pain，gum tenderness]四种意义相同的副作用名称。

本文从 DailyStrength 网站上抓取了 SIDER 数据库中存在记录的 870 种药物在 2013 年 3 月 24 日之前的用户评论。DailyStrength 上的用户评论是按药物分组的，即每个评论对应的药物是确定的，无需再对评论中副作用名称与药物的关联关系进行判别。用户在撰写评论时具有一定的随意性，导致语料中存在一些不规范的语言现象。为了减少它们对后续处理造成的影响，我们对评论语料进行了一些预处理：在无结束标点符号的评论语句后面加上表示结束的句号；修正一些不规则的写法(如 !!! -> !，isnt -> is not，im -> I am，ive -> I have)等。

预处理之后，对评论内容进行句子划分，共得到 213 466 个不同的句子。从中随机抽取一定数量的句子，采用传统的{B，I，E，S，O}标记方法，标注实体在句子中的起止位置。随机抽取并标注了 1 500

个含有实体的句子，将其作为实体识别的训练集；另外随机抽取出 500 个句子进行标注，将其作为实体识别的测试集（这些句子中有的包含实体，有的不包含，其分布情况与整体语料相同）。从用户评论中挖掘药物副作用这一领域，目前还没有一个权威、公开的标注数据集可以用来测试副作用实体识别方法的性能，因此我们利用自己标注的数据集来对本文副作用实体识别方法的效果进行测试。

### 2.2　副作用实体识别与过滤

副作用实体的识别涉及从用户评论中识别副作用名称，本文将副作用实体识别归结为命名实体识别问题，而序列标注是命名实体识别领域常用的方法。CRF 作为一种无向图模型，常用于序列标注和切分序列化数据等问题。CRF 既能克服 HMM 严格的条件独立性的假设，又克服了 MEMM 的偏置问题，可以更加真实地拟合现实数据，所以在命名实体识别领域得到广泛的应用。唐旭日等人[9]使用 CRF 模型识别中文地名。Settles 等人[10]将 CRF 模型运用于生物命名实体识别，取得很好的效果。刘凯等人[11]使用 CRF 模型识别中医临床病例中的命名实体，相较于 HMM 和 MEMM，取得了最好的结果。

鉴于 CRF 模型在命名实体识别领域的出色性能，本文采用 CRF 模型从用户评论语料中识别副作用实体，具体使用了开源的 CRF++ 工具包①。识别时利用了词语的两类特征：词语特征与词性特征，其中词性特征是使用 Stanford POS Tagger 工具包[12]中的 english-left3words-distsim. tagger 模型对评论语句标注得到的。在利用 CRF 模型进行识别时，对于每个词语，本文考虑的上下文特征包括当前词语与前两个、后两个词语与词性特征。表 1 是评论“very good pain relief but too strong .”中每个单词的特征情况，其中“POS”为词性标记，“标注”是人工标注的结果。

表 1　CRF 特征

| 评论 | POS | 标注 | 当前词与前后两个词特征 | 当前词与前后两个词 POS 特征 |
|---|---|---|---|---|
| very | RB | O | very, good, pain | RB,JJ,NN |
| good | JJ | O | very, good, pain, relief | RB,JJ,NN,NN |
| pain | NN | S | very, good, pain, relief,but | RB,JJ,NN,NN,CC |
| relief | NN | O | good, pain, relief, but, too | JJ,NN,NN,CC,RB |
| but | CC | O | pain, relief, but, too, strong | NN,NN,CC,RB,JJ |
| too | RB | O | relief, but, too, strong,. | NN,CC,RB,JJ,. |
| strong | JJ | O | but, too, strong,. | CC,RB,JJ,. |
| . | . | O | too, strong,. | RB,JJ,. |

在训练集上对 CRF 模型训练完成后，我们在测试集以及所有的评论语料上进行实体识别。我们利用 CRF 识别出句子中的所有实体，但是，这些实体既包含药物的副作用也包含药物的适应症，如表 1 所示的评论，说明相应的药物缓解了疼痛，即疼痛不是其副作用。因此还需要过滤掉药物的适应症。对于药物适应症的鉴定，参考 Leaman 等人[2]的做法并略作改进，我们根据实体所在子句（Clause）是否含有某些特定词语来确定其是否是药物的适应症。适应症实体所在子句中通常含有 ease, work for, help with, relief 等表示治疗、缓解等意义的词汇，为此我们收集了这样一个指示词表，并根据实体所在子句内是否含有这些词汇确定其为副作用或适应症。此外，为了提高实体识别的准确性，我们还检测了子句中的否定词，并据此进一步滤除非药物副作用的实体。

### 2.3　副作用实体标准化

在副作用实体的标准化中，药物的每种副作用被视为一个副作用概念，它对应着一种或多种表述形式即副作用实体。实体标准化就是通过一定的手段将实体映射到对应的标准概念上，一般可分为精确匹配（Exact Matching）和近似匹配（Approximate Matching）两种方式。在本文中，对于从评论里识别出来的副作用名称，若词典中存在该名称，则直接通过精确匹配得到对应的标准化概念；否则进行近似

① http://crfpp.sourceforge.net/

匹配，即利用本文所述的近似匹配方法将其映射到标准化概念上，或者该实体在词典中无法找到对应的概念，而可能属于一种新的副作用概念。

1. 方法流程

对于一个待标准化实体，如果精确匹配成功，则直接得到标准化概念；否则我们通过近似匹配从副作用词典中寻找与之最相关的副作用名称，并将该名称对应的概念作为标准化概念。本文的近似匹配部分由三个模块组成，这三个模块分别基于常规检索、扩展语义检索以及编辑距离进行标准化。本文提出的药物副作用标准化方法的流程如图2所示。

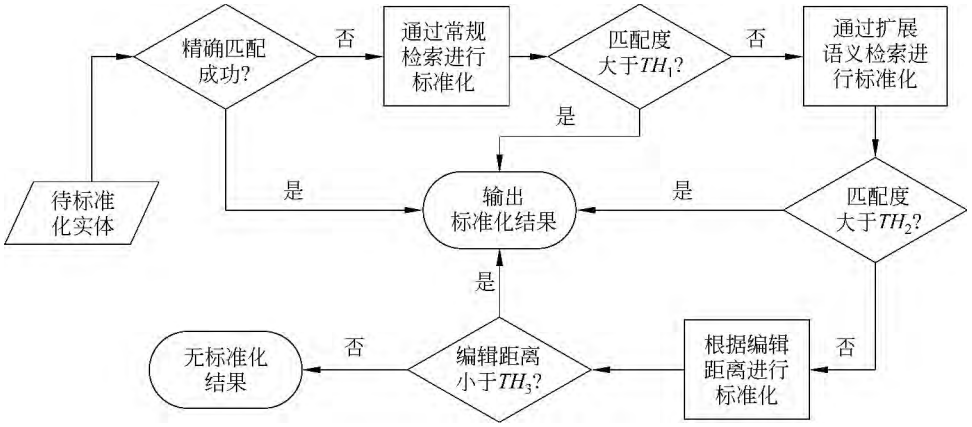


图2 药物副作用名称标准化方法的流程图

本文标准化方法的近似匹配部分首先通过常规检索进行标准化，若得到的匹配度大于设定的阈值 $TH_1$，则将相应的概念作为标准化结果；否则通过扩展语义检索进行标准化，若得到的匹配度大于设定的阈值 $TH_2$，则将相应的概念作为标准化结果；否则根据编辑距离进行标准化，若得到的最短编辑距离小于设定的阈值 $TH_3$，则将相应的概念作为标准化结果。否则，词典中没有任何概念与当前待标准化实体匹配，该实体可能属于一种新的副作用概念。

2. 近似匹配模块

在本文近似匹配的三个模块中，前两个模块利用信息检索中的TF-IDF思想初步限定候选概念范围，然后通过计算副作用名称之间的匹配度进一步确定标准化概念；第三个模块则根据最短编辑距离寻找最佳的标准化概念。

匹配度函数

模块1和2利用匹配度函数 $MD(Ent, Ent_i)$ 计算待标准化实体 $Ent$ 与词典中某一实体名称 $Ent_i$ 之间的匹配程度(Match Degree)，其具体的计算过程如下：

1) 将 $Ent$ 与 $Ent_i$ 进行分词、去停用词、词干化处理，分别得到词袋 $A$ 和 $B$。

2) 对于词袋 $A$ 中的每个单词 $a_m$($m = 1,2,\cdots,p$，$p$ 为 $A$ 中单词数)，遍历词袋 $B$，由如下公式计算出 $a_m$ 与 $B$ 中每个单词 $b_n$ 的字面相似度(Literal Similarity) $LS(a_m, b_n)$：

$$LS(a_m, b_n) = 2 \cdot N_{cc} / (L_a + L_b) \tag{1}$$

其中 $N_{cc}$ 是 $a_m$ 与 $b_n$ 的公共子串长度，$L_a$ 为 $a_m$ 的字符数，$L_b$ 为 $b_n$ 的字符数。若 $N_{cc}$ 小于设定的阈值，则认为不具有表征 $a_m$ 与 $b_n$ 内在相关性的作用，并且可能会作为噪音影响计算结果，将其置为0。

从 $B$ 的所有单词与 $a_m$ 的字面相似度中找出最大值作为单词 $a_m$ 最终的字面相似度 $LS_m$，并将与之匹配的单词从 $B$ 中删除，使之不重复作为 $A$ 中其他单词的最佳匹配。这样，实体 $\mathrm{Ent}_1$ 与 $\mathrm{Ent}_2$ 之间的字面相似度 $LS$ 为

$$LS(Ent, Ent_i) = \sum_{m=1}^{p} LS_m / (L_a + |L_a - L_b|) \tag{2}$$

3) 计算 $Ent_i$ 对应的概念 $Con_i$ 涵盖 $Ent$ 中单词的程度 $WC$ (Word Coverage)。$Con_i$ 的全词袋 $C = \bigcup B_i$ ($B_i$ 为 Con 的第 $i$ 个实体对应的词袋，$i = 1,2,\cdots,q$，$q$ 为 Con 中实体数)，词袋 $A$ 与 $C$ 之间的相同单词集合为 $|A \cap C|$，则 $Con_i$ 涵盖 $Ent$ 中单词的程度 $WC(Ent, Ent_i) = |A \cap C| / |A|$。

(4) Ent 与 $\mathrm{Ent_i}$ 之间的匹配度 $MD(Ent, Ent_i) = LS(Ent, Ent_i) + r \cdot WC(Ent, Ent_i)$，($r \in [0,1]$ 数)。

**模块 1　通过常规检索进行标准化**

将每个副作用概念 $Con_i$ 视为一篇文档，$Con_i$ 对应的所有副作用名称作为文档内容，对该文档进行分词、去停用词以及词干化。将待标准化实体 $Ent$ 也看为文档，并进行相同的处理。根据 TF-IDF 技术，将每个副作用概念 $Con_i$ 与待标准化实体 $Ent$ 分别表示为向量 $v_i$ 和 $v$，然后利用开源搜索引擎框架 Indri[①] 计算 $Ent$ 与 $Con_i$ 之间的相关度，其中 Indri 内部使用的是语言模型和贝叶斯推理网络相结合的检索模型，可以用来有效地计算文档之间的相关度。从模型的检索结果中我们选取 TOP N1 个概念作为候选概念，然后利用上述匹配度函数 MD 计算待标准化实体与候选概念中每种副作用名称之间的匹配度，并将匹配度最大的名称对应的概念作为该实体的标准化概念。若检索结果为空，则该模块的标准化结果为空。

**模块 2　通过扩展语义检索进行标准化**

若待标准化实体 $Ent$ 所含词语在其他副作用名称中很少出现，这时直接将其作为查询进行检索的效果可能不好。因此本文利用 WordNet[13] 数据对 $Ent$ 中的词语进行语义扩展，使每个词语扩展得到最多 5 个同义词（有些词语在 WordNet 中的同义词数量可能少于 5 个），然后将扩展后的词语集合作为 $Ent$ 的新文档，并经过与模块 1 相同的处理，将其表示为新的向量 $\bar{v}$，然后利用模块 1 中的方法选取 TOP $N_2$ 个最相关概念作为候选概念。最后，同样利用匹配度函数计算它与各个候选概念的副作用名称的匹配度，并将匹配度最大的名称对应的概念作为它的标准化概念。若检索结果为空，则该模块的标准化结果为空。

**模块 3　根据编辑距离进行标准化**

副作用名称中存在以下现象：① 两个词语的意义相同但拼写却存在一定差别（如“病毒血症”的两种拼写 viremia 与 viraemia）；② 某一词语为另一短语的缩写形式（如概念 C0079773 的副作用名称有 CTCL 和 cutaneous T cell lymphoma，前者为后者的首字母缩写）。待标准化实体中若含有这些词语，则检索结果可能不理想，无法命中正确的概念。这种情况下我们根据字符串之间的编辑距离寻找与之最匹配的名称。

编辑距离（Edit Distance）用来衡量两个字符串字面上的相异性。字符串 $str_1$ 和 $str_2$ 之间的编辑距离 $ED(str_1, str_2)$ 是指从 $str_1$ 转换成 $str_2$ 所需要的插入、删除和替换的最少次数。对于待标准化实体 $Ent$ 和词典中的实体名称 $Ent_i$ 分别进行分词、去停用词得到词袋 A 和 B，$Ent$ 和 $Ent_i$ 之间的实体编辑距离定义为：

$$Ent\_ED(Ent, Ent_i) = \sum_{w_i \in A} \min_{w_j \in B} ED(w_i, w_j) \quad (3)$$

通过实体编辑距离得到与待标准化实体编辑距离最短的副作用名称，并将其对应的概念作为该实体的标准化概念。

在计算两个副作用名称之间编辑距离时，我们考虑了其中某个词语为缩写词的情况。一般来说，缩写词通常为某个短语的单词首字母的缩写，或单词中前缀的首字母加上剩余部分首字母的缩写（为此我们收集了一个英文前缀表）。因此在计算两个短语之间的编辑距离时，若某词语为另一短语的缩写词，则该词语与短语的编辑距离为 0。

### 2.4　药物副作用发现

根据从评论语料中发现的所有副作用名称，并参照标准化结果，我们得到每种药物的评论中出现的副作用概念及包含此概念的评论所占的比例即发生频率。对于发现的药物已知的副作用，我们将其与已有的数据进行对比，验证本文挖掘方法的有效性；对于数据库中未记录的药物副作用，我们按其在评论中的发生频率由高到低排序，得到一个药物潜在副作用列表。

## 3　实验结果与分析

### 3.1　实体识别效果测试

#### 3.1.1　实体识别效果测试

我们在标注好的 1 500 个评论语句上训练 CRF 模型，然后利用该模型对测试集中的 500 个评论语句进行副作用实体识别，将识别出的实体进行过滤后，得到实体识别的准确率为 87.5%，召回率为 58.7%，F 值为 70.3%。同样以 DailyStrength 上的用户评论作为实验语料，Leaman 等人[2] 在其使用的人工标注的数据集上识别的准确率为 78.3%，召回率为 69.9%，F 值为 73.9%；Nikfarjam 等人[4] 在其人工标注的数据集上识别的准确率为 70.01%，召回率为 66.32%，F 值为 67.96%。文中所用实验数据与 Nikfarjam 等人所用的数据基本相同，都来自 DailyStrength 网站的评论模块，属于同源数据，

① http://www.lemurproject.org/indri/

具有相同的数据分布，因此数据具有可比性。上述结果说明本文所述的方法可以有效的识别用户评论中的副作用实体。

对错误识别的样例进行分析，我们发现CRF模型最主要的错误是不能识别由分散的词语构成的实体，如无法从"… major swelling in my ankles and …"识别出副作用ankles swelling；另外有一部分错误是由于识别出的实体与标准答案不完全相同造成的，例如general feeling of illness与illness、frequent headaches与headaches等（前者为标准答案，后者为识别结果）。

在社交网络中，随意性和表达多样性是用户评论的重要特性。对于同一副作用，用户可以使用多种表达方式来描述，这些表达方式差异性很大。对于某种表达方式，如果在训练数据集中存在其足够的信息和特征，CRF模型就可以对这种表达方式做出正确的标记。如果训练数据集中某种表达方式的信息较少或者出现次数较少，CRF模型倾向于将其标记为普通文本，而不是将其标记为错误的副作用实体。如文中提到的评论："… major swelling in my ankles and …"，其包含的副作用名称为"ankles swelling"，CRF模型并没有将其标记为"swelling"，而是将其标记为普通文本。因此，CRF模型在本文实验中准确率较高。

从识别结果可以看出，本文方法识别实体的准确率较高，而召回率相对来说较低，说明由本文方法识别出的副作用实体大部分是正确的，后续挖掘出的药物副作用关系是可靠的。在未来工作中，可以考虑向CRF模型中引入更多有效的特征来提高副作用实体识别的性能。

#### 3.1.2 从评论中识别副作用实体

我们从870种药物的408 318条评论中识别实体并过滤后，得到了729个词典中存在的副作用名称与3 143个新的副作用名称，表2显示了本文挖掘出的词典中已有名称与新名称的统计情况（括号中为对应数值占总体的百分比），表3显示了识别出的出现频率最高的前10个新的副作用名称。从结果可以看到，利用CRF模型不但识别出了已知的副作用名称，而且能够识别出潜在的新副作用名称。由表可知，新名称出现的总次数占总体的18.0%，而平均出现次数相对于已知名称却少得多，说明用户在评论中使用新的、不同的副作用表述方式是很普遍的，因此进行标准化是很有必要的。

### 3.2 药物副作用标准化

为了验证提出的标准化方法的有效性，本文首先对近似匹配模块标准化的准确率进行了测试。在测试时，我们从副作用词典中随机抽取满足要求（即该副作用名称在词典中须有属于同一概念的其他副作用名称）的副作用名称作为待标准化实体，同时将该名称从词典中删除，并对删除该名称后的词典建立索引。利用上述的标准化方法得到该实体的标准概念，并与正确的标准概念对比，从而得到标准化的准确率。在测试该标准化方法时，我们对其中的三个阈值$TH_1$、$TH_2$、$TH_3$调优，并将最优的阈值用于从评论中识别出的药物副作用的标准化中。

表2 药物副作用实体识别结果统计

| | 数量 | 出现总次数 | 每个名称平均出现次数 |
|---|---|---|---|
| 已知的副作用名称 | 729 (18.8%) | 58 810 (82.0%) | 80.7 |
| 新的副作用名称 | 3 143 (81.2%) | 12 932 (18.0%) | 4.1 |
| 总计 | 3 872 | 71 742 | —— |

表3 识别出的频率最高的前10个新的副作用名称

| 排名 | 副作用名称 | 出现次数 | 排名 | 副作用名称 | 出现次数 |
|---|---|---|---|---|---|
| 1 | gained weight | 814 | 6 | stomach pains | 94 |
| 2 | heart race | 153 | 7 | shakes | 86 |
| 3 | gaining weight | 131 | 8 | stomach hurt | 78 |
| 4 | shaking | 125 | 9 | suicidal | 74 |
| 5 | painful | 123 | 10 | stomach aches | 71 |

#### 3.2.1 检索返回候选概念的数量

为了合理设置检索返回的候选副作用概念的数量，我们对500个待标准化实体进行常规检索并统计返回的前$n$个候选概念中包含正确概念的比例，结果如图3所示。可以看出随着返回候选概念数量的增加，结果中包含正确概念的比例逐渐变大，当返回候选概念数量$n$为20时该比例已达82.0%；但增速却逐渐变缓，当$n$为30时该比例为83.2%，仅增加了1.2%，最终很难达到理想的100%。造成这种现象的一个可能原因是有些待标准化实体为某些生僻词或缩写词，索引中几乎没有与其拼写相同词语，从而无法通过常规检索返回正确的概念。这也是需要利用扩展语义检索与编辑距离进行标准化的原因。

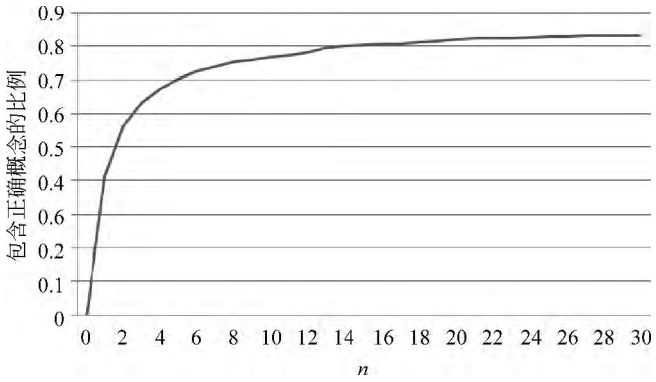


图 3　常规检索返回的前 $n$ 个候选概念中包含正确概念的统计概率

综合考虑检索结果包含正确概念的比例以及检索的效率，我们在实验中将常规检索返回的候选概念数量 $N_1$ 设置为 25；而扩展语义后查询词语得到扩充，与之相关的候选概念数量也会相应地增多，因而我们将扩展语义检索返回的候选概念数量 $N_2$ 设置为 40。

3.2.2　近似匹配模块标准化测试与分析

为了测试各个模块对标准化准确率的提升作用，我们分别采用“模块 1”、“模块 2”、“模块 3”、“模块 1＋2”、“模块 1＋2＋3” 五种组合方式对副作用名称标准化。每种组合方式进行十次实验，每次从词典的 10 498 个副作用名称中随机抽取 500 个用于测试，并根据标准化结果计算其准确率。标准化方法测试的结果如表 4 所示。

表 4　本文标准化方法的测试结果

| 组合方式 | 准确率/% | | |
|---|---|---|---|
| | 最大值 | 最小值 | 平均值 |
| 模块 1 | 70.4 | 66.7 | 68.6 |
| 模块 2 | 56.8 | 51.0 | 53.8 |
| 模块 3 | 66.8 | 58.8 | 63.4 |
| 模块 1＋2 | 74.0 | 69.4 | 71.4 |
| 模块 1＋2＋3 | 75.0 | 71.4 | 73.0 |

由实验结果可以看出，近似匹配模块单独使用时，模块 1 的性能最好，模块 3 次之，模块 2 最差。在模块 1 的结果之上加入模块 2 后，标准化的准确率有了提升，说明将待标准化实体进行语义扩展，通过同义词语寻找正确概念的做法在涉及一些低频率、生僻词语时具有益处。在此基础上，继续添加模块 3 后标准化的准确率进一步提升，说明副作用名称中包含一定数量的缩写词以及意义相同、词形相近的词语，此时根据编辑距离进行匹配具有较好的效果，同时也是对前两个模块功能的补充。由此可见，本文的匹配度函数确实在一定程度上反映了副作用名称之间的内在联系，使得大部分待标准化实体映射到了其正确的概念上。

分析标准化结果中错误的实例，我们发现了以下几种导致标准化错误的情况。

1）有些形式十分接近的副作用名称属于不同的概念，在对其中某个名称标准化时会错误映射至另一名称对应的概念。例如，概念 C0018772 下的 impaired hearing 与概念 C1384666 下 hearing impairment 在词干化并忽略词序后完全匹配，但它们却属于不同的概念。

2）利用 WordNet 数据对副作用名称扩展语义时，由于 WordNet 本身的局限性，有时并不能将合适的词语扩充进来。例如，在对概念 C0549448 下的 elevated hemoglobin 标准化时，WordNet 并不能将 elevate 扩展得到同义词 increase，从而无法匹配到同概念的 increased hemoglobin。

3）副作用名称中的专业词汇常常无法得到扩展，而专业词汇与同概念下的其他名称在词形上的关联又很弱，从而导致标准化错误。例如，cholelithiasis 属于概念 C0008350，而此概念下的所有名称为［gall stone，gallstones，cholelithiasis，biliary calculi］。

3.2.3　对识别出的实体进行标准化

对 3 143 个新的副作用名称进行标准化处理，其中 2 337 个新名称映射到了 974 个概念上，平均每个概念约对应 2.4 个新名称；剩余的 806 个新名称无法对应到词典中已有的概念上，可能属于新的副作用概念。图 4 显示了副作用概念 C0043094 在词典中已有的名称以及本文从评论中挖掘得到的新名称，其中实线框里的是词典中已有的名称，虚线框里的是挖掘出的新名称（有些拼写是错误的）。可以

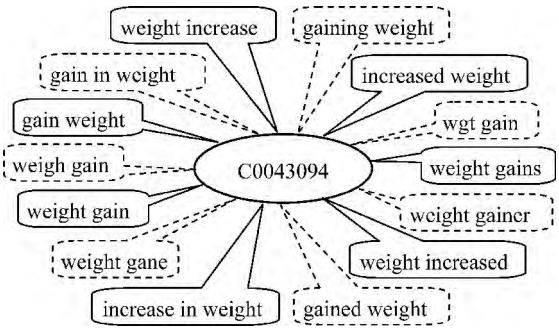


图4 概念C0043094在词典中已有的副作用名称及本文挖掘的新名称

看出，通过对新名称进行标准化，我们可以将用户对同一概念的不同表述形式（包括评论中常见的因拼写错误而产生的不同形式）映射到其真正所指的概念上，实现副作用名称的有效聚合与归并，使副作用概念在评论中的发生比例更接近其在用药者中真正的发生频率，从而有利于药物潜在副作用的发现。

### 3.3 药物副作用发现

通过对识别出的副作用名称进行标准化，我们将不同的表述形式映射到了其所指的概念上，从而可以统计出副作用概念在每种药物评论中的发生频率。为了避免偶然现象而使结果更具统计意义，我们选择评论数量大于50的药物，将挖掘出的副作用概念按照在对应药物评论中的发生频率由高到低排序，得到药物副作用的列表。

对于药物已知的副作用，我们将挖掘出的发生频率与SIDER数据库中记录的发生频率进行了对比。表5显示了挖掘得到的发生频率最高的前十种已知的药物-副作用对与数据库中记录的相应数据，其中“postmarketing”表示副作用在药物上市后得到确认，“potential”表示药物可能的副作用。从表中可以看出，我们从评论中挖掘出的具有较高发生频率的药物副作用，其在数据库中记录的发生率一般也相应地较高，两种来源的药物副作用发生频率具有较大程度的相似性与对应性，说明本文的药物副作用挖掘方法是有效的，挖掘得到的药物副作用结果具有较大的可信度。

对于发现的数据库中未记录的药物副作用，可以认为副作用概念在某种药物的评论中的发生频率越高，则其为该药物潜在副作用的可能性越大。因此，我们按挖掘到的发生频率由高到低排序，得到了一个可能性由大到小排列的药物潜在副作用列表。表6显示了本文挖掘到的前十个最有可能的潜在药物-副作用对。对于挖掘出的具有较高发生频率的药物副作用，可以作为药物潜在的副作用以备参考。

表5 挖掘出的发生频率最高的10种已知的药物-副作用对与数据库记录之间的对比

| 排名 | 药物 | 副作用概念 | 挖掘得到的发生频率/% | 数据库记录的发生频率 |
|---|---|---|---|---|
| 1 | Nitrostat (Nitroglycerin) | C0018681[headache, …] | 18.5 | 18% |
| 2 | Mirena (Provera) | C0030193[unspecified pain, …] | 13.7 | potential |
| 3 | Gleevec (Imatinib) | C0027497[nausea, …] | 13.6 | 9% |
| 4 | Zyprexa (Olanzapine) | C0043094[weight gain, …] | 12.0 | 5% |
| 5 | Oxytrol (Oxybutynin) | C0043352[dry mouth, …] | 11.6 | 8.33% |
| 6 | Indocin (Indomethacin) | C0038354[gastric disorder, …] | 11.5 | potential |
| 7 | Lupron (Leuprolide) | C0600142[hot flushes, …] | 10.7 | postmarketing |
| 8 | Doxorubicin | C0027497[nausea, …] | 10.1 | 18.2% |
| 9 | Danazol | C0085633[mood swings, …] | 10.0 | potential |
| 10 | Parlodel (Bromocriptine) | C0012833[dizziness, …] | 9.3 | postmarketing |

表6 挖掘出的发生频率最高的前10种潜在药物-副作用对

| 排名 | 药物 | 副作用概念 | 挖掘得到的发生频率/% |
|---|---|---|---|
| 1 | Arthrotec | C0038354[gastric disorder, …] | 15.7 |
| 2 | Piroxicam | C0038354[gastric disorder, …] | 11.3 |
| 3 | Robaxin | C0234450[sleepy] | 7.7 |

续有

| 排名 | 药物 | 副作用概念 | 挖掘得到的发生频率/% |
|---|---|---|---|
| 4 | Tussionex | C0030193[unspecified pain, …] | 7.6 |
| 5 | Nitrostat | C0008031[chest pain, …] | 7.4 |
| 6 | Carboplatin | C0015672[fatigue, …] | 7.2 |
| 7 | Seasonique | C0030193[unspecified pain, …] | 7.2 |
| 8 | Fosavance | C0038354[gastric disorder, …] | 6.8 |
| 9 | Mirena | C0026821[cramps, …] | 6.8 |
| 10 | Danazol | C0149931[migraine, …] | 6.7 |

## 4　结论与展望

从社交网络的用户评论中提取药物副作用信息是一种快捷、有效的渠道，而评论中含有大量数据库未收录的副作用名称，识别这些新名称并标准化对药物副作用的挖掘十分重要。针对前人工作中对新副作用名称的识别效果不佳以及对识别出的新名称后续处理不足的问题，本文利用 CRF 模型识别评论中的副作用，可以识别出已知的与新的名称。将副作用名称标准化可以对其进行有效的聚合与归并，有利于药物副作用的发现。我们通过将挖掘出的药物已知的副作用与数据库记录进行对比验证本文方法的有效性，对挖掘出的数据库中未记录的药物副作用按其在评论中的发生频率排序，得到了一个可能性由大到小排列的药物潜在副作用列表。

在未来工作中，1）考虑在副作用实体识别的 CRF 模型中加入更多有效的特征，如药物的分子式特征、药物适应症特征、副作用词典特征以及单词的分布式矢量特征等，以便提高实体识别的效果；2）鉴于 WordNet 数据存在的局限性，在标准化时可以考虑引入生物医学领域的专业词典，或是借助语义相似度数据来衡量词语之间的关联程度，提高标准化方法的准确率；3）对于挖掘出的新的副作用名称，如果无法映射到现有的副作用概念上，则考虑通过它们之间的关联度将其进行聚类，从而更好地发现药物潜在的副作用以及新的副作用概念；4）相较于发现潜在药物不良反应，发现产生不良反应的原因和条件则具有更加深远的意义，未来工作中会着重挖掘不良反应发生的原因。

## 参考文献


[1] Yang C C, Jiang L, Yang H, et al. Detecting signals of adverse drug reactions from health consumer contributed content in social media[C]//Proceedings of ACM SIGKDD Workshop on Health Informatics. 2012.

[2] Leaman R, Wojtulewicz L, Sullivan R, et al. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks[C]//Proceedings of the 2010 workshop on biomedical natural language processing. Association for Computational Linguistics, 2010: 117-125.

[3] Chee B W, Berlin R, Schatz B. Predicting adverse drug events from personal health messages[C]//Proceedings of the AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2011: 217.

[4] Nikfarjam A, Gonzalez G H. Pattern mining for extraction of mentions of adverse drug reactions from user comments[C]//Proceedings of the AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2011: 1019.

[5] Zeng Q T, Tse T. Exploring and developing consumer health vocabularies[J]. Journal of the American Medical Informatics Association. 2006, 13(1): 24-29.

[6] Wu H, Fang H, Stanhope S J. Exploiting online discussions to discover unrecognized drug side effects[J]. Nervenheilkunde. 2007, 26(11): 969-980.

[7] Online Support Groups and Forums at DailyStrength. Available[DB]. www. dailystrength. org. Accessed March 28, 2014.

[8] Kuhn M, Campillos M, Letunic I, et al. A side effect resource to capture phenotypic effects of drugs[J]. Molecular systems biology. 2010, 6(1):343-348.

[9] 唐旭日，陈小荷，许超，等. 基于篇章的中文地名识别研究[J]. 中文信息学报，2010，24(02)：24-32.

[10] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets[C]// Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Association for Computational Linguistics, 2004: 104-107.

[11] 刘凯,周雪忠,于剑,等. 基于条件随机场的中医临床病历命名实体抽取[J]. 计算机工程,2014(9): 312-316.

[12] Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 173-180.

[13] Miller G A. WordNet: a lexical database for English [J]. Communications of the ACM. 1995, 38(11): 39-41.

**赵明珍**(1989—),硕士研究生,主要研究领域为文本挖掘和自然语言处理。
E-mail: zmz@mail.dlut.edu.cn

**林鸿飞**(1962—),博士,教授,博士生导师,主要研究领域为搜索引擎、文本挖掘、情感计算和自然语言处理。
E-mail: hflin@dlut.edu.cn

**程亮喜**(1986—),硕士研究生,主要研究领域为生物医学文本挖掘和自然语言处理。
E-mail: liangxicheng@mail.dlut.edu.cn

(上接第171页)

[6] 常晓龙,张晖. 融合语素特征的中文褒贬词典构建[J]. 计算机应用, 2012, 32(7): 2033-2037.

[7] 徐琳宏,林鸿飞,潘宇,等. 情感词汇本体的构造[J]. 情报学报,2008,27(2): 180-185.

[8] 桂守才. 基础心理学[M]. 北京: 人民教育出版社, 2007.

[9] 林传鼎. 社会主义心理学中的情绪问题[J]. 社会心理学科,2006,21(83):37-62.

**蒋盛益**(1963—),博士,教授,主要研究领域为数据挖掘与自然语言处理。
E-mail: jiangshengyi@163.com

**蔡茂丽**(1992—),本科生,主要研究领域为 Web 数据挖掘。
E-mail: 996992867@qq.com

**黄卫坚**(1993—),本科生,主要研究领域为 Web 数据挖掘。
E-mail: 634101364@qq.com