第10卷 第25期 2010年9月
1671-1815(2010)25-6300-05

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 10 No. 25 Sep. 2010

从 Web 网页上获取一价事件常识的方法

杨 帆 孙 强

(江苏科技大学,镇江 212003;91206 部队,青岛 266108)

摘 要 一价事件是以一价动词为核心构成的事件。为了提高查询的智能性和准确性,尝试用一价事件设计描述了从《实习词表》中挑选一价动词,根据《一价动词表》设计 Google 查询项,根据 Google 查询项从 Web 网页上获取大规模的语料。用特征提取方法从 Web 语料中抽取事件上下文中的相关词,根据相关词评价 Google 查询项的优劣,并进行调整。得到与一价动词相关的因果逻辑,丰富查询项,从而提高查询精度。

关键词 特征提取 文本 信息检索 分词算法

中图法分类号 TP391.3; **文献标志码** A

特征抽取又称特征提取,是指模式识别中,对某一模式的一组测量值进行变换以突出该模式具有代表性特征的方法。是一种提取有效信息的方法。特征,通常指传感器某一通道反射率测量值。与某一模式有关的特征数目称为其"维数"。特征抽取的目的就是从噪音中分离出有用的信息以及减少数据的维数,以简化分类器中所进行的计算。特征提取(Feature Selection)通常根据某个特征评估函数计算各个特征的评分值,然后按评分值对这些特征进行排序,选取若干个评分值最高的作为特征词。特征提取的对象是海量、异构、分布的文档(Web)[1];文档内容是人类所使用的自然语言,缺乏计算机可理解的语义。目前有关文本表示的研究主要集中于文本表示模型的选择和特征词选择算法的选取上。随着网络知识组织、人工智能等学科的发展,文本特征提取将向着数字化、智能化、语义化的方向深入发展,在社会知识管理方面发挥更大的作用。

1 特征提取方法

1.1 特征项的特性

基于表示文本的基本单位通常称为文本的特征或特征项,而特征项必须具备一定的特性:

(1) 特征项要能够确实标识文本内容;

(2) 特征项具有将目标文本与其他文本相区分的能力;

(3) 特征项的个数不能太多;

(4) 特征项分离要比较容易实现。

1.2 特征选取的方式

(1) 用映射或变换的方法把原始特征变换为较少的新特征;

(2) 从原始特征中挑选出一些最具代表性的特征;

(3) 根据专家的知识挑选最有影响的特征;

(4) 用数学的方法进行选取,找出最具分类信息的特征,这种方法是一种比较精确的方法,人为因素的干扰较少,尤其适合于文本自动分类挖掘系统的应用。

随着网络知识组织、人工智能等学科的发展,文本特征提取将向着数字化、智能化、语义化的方向深入发展,在社会知识管理方面发挥更大的作用。

1.3 三种特征提取方法

1.3.1 基于统计的特征提取方法

这类型算法通过构造评估函数,对特征集合中的每个特征进行评估,并对每个特征打分,这样每个词语都获得一个评估值,又称为权值。然后将所有特征按权值大小排序,提取预定数目的最优特征

2010年5月27日收到

第一作者简介:杨 帆(1985—),女,江苏连云港人,硕士研究生,研究方向:智能信息处理。E-mail:daffodilyang@gmal.com。

作为提取结果的特征子集。显然,对于这种类型算法,决定文本特征提取效果的主要因素是评估函数的质量。

基于评估函数的特征提取方法是建立在特征独立的假设基础上,但在实际中这个假设是很难成立的,因此需要考虑特征相关条件下的文本特征提取方法。

1.3.2　基于体论的特征提取方法[2]

应用本体论(Ontology)模型可以有效地解决特定领域知识的描述问题。具体针对数字图像领域的文本特征提取,通过构建文本结构树,给出特征权值的计算公式。算法充分考虑特征词的位置以及相互之间关系的分析,利用特征词统领长度的概念和计算方法,能够更准确地进行特征词权值的计算和文本特征的提取。

特征权值的计算公式:设特征词 t_i 的特征权为 w_i,则有:

$$w_i = \alpha\delta_i K_d L(t_i) \tag{1}$$

$$L(t_i) = \sum \lg l(t), t \in S \tag{2}$$

式中,α 表示特征词的位置加权系数;δ_i 是表征能力加权;K 是常数(实验中 $K=2$);参数 d 是特征词在本体树中的层次数(根节点 $d=0$);L 表示统领长度,集合 S 是在文本结构树中以特征词 t_i 为根节点的所有叶子节点的集合;t 是集合 S 中的元素,$l(t)$ 表示 t 的统领长度。特征词选取结果集合为 R,公式如下:

$$R = \{w_i \, w_i \geqslant Z\} \tag{3}$$

$$Z = 1\mathrm{n} \sum n_i = 1 w_i \tag{4}$$

其中 Z 为阈值。

1.3.3　基于词性的特征提取方法[3]

考虑到汉语言中,能标识文本特性的往往是文本中的实词,如名词、动词、形容词等。而文本中的一些虚词,如感叹词、介词、连词等,对于标识文本的类别特性并没有贡献,也就是对确定文本类别没有意义的词。如果把这些对文本分类没有意思的虚词作为文本特征词,将会带来很大噪音,从而直接降低文本分类的效率和准确率。因此,在提取文本特征时,应首先考虑剔除这些对文本分类没有用处的虚词,而在实词中,又以名词和动词对于文本的类别特性的表现力最强,所以可以只提取文本中的名词和动词作为文本的一级特征词。

1.4　影响特征词权值的因素

1.4.1　词频

文本内空中的中频词往往具有代表性,高频词区分能力较小,而低频词或者示出现词也常常可以做为关键特征词。所以词频是特征提取中必须考虑的重要因素,并且在不同方法中有不同的应用公式。

1.4.2　词性

汉语言中,能标识文本特性的往往是文本中的实词,如名词、动词、形容词等。而文本中的一些虚词,如感叹词、介词、连词等,对于标识文本的类别特性并没有贡献,也就是对确定文本类别没有意义的词。

1.4.3　文档频次

出现文档多的特征词,分类区分能力较差,出现文档少的特征词更能代表文本的不同主题。

1.4.4　标题

标题是作者给出的提示文章内容的短语,特别在新闻领域,新闻报道的标题一般都要求要简练、醒目,有不少缩略语,与报道的主要内容有着重要的联系,对摘要内容的影响不可忽视。

1.4.5　位置

美国的 EE. Baxendale 的调查结果显示:段落的论题是段落首句的概率为 85% ,是段落末句的概率为 7% 。而且新闻报道性文章的形式特征决定了第一段一般是揭示文章主要内容的。因此,有必要提高处于特殊位置的句子权重,特别是报道的首句和末句。

1.4.6　句法结构

句式与句子的重要性之间存在着某种联系,比如摘要中的句子大多是陈述句,而疑问句、感叹句等则不具内容代表性。而通常“总之”、“综上所述”等一些概括性语义后的句子,包含了文本的中心内容。

1.4.7　专业词库

通用词库包含了大量不会成为特征项的常用词汇,为了提高系统运行效率,系统根据挖掘目标建立专业的分词表,这样可以在保证特征提取准确

性的前提下,显著提高系统的运行效率。

1.4.8 信息熵

熵(Entropy)在信息论中是一个非常重要的概念,它是不确定性的一种度量。信息熵方法的基本目的是找出某种符号系统的信息量和多余度之间的关系,以便能用最小的成本和消耗来实现最高效率的数据储存、管理和传递。

1.4.9 文档、词语长度

一般情况下,词的长度越短,其语义越广泛。

1.4.10 单词的区分能力

在 TF-IDF 公式的基础上,又扩展了一项单词的类区分能力。新扩展的项用于描述单词与各个类别之间的相关程度。

1.4.11 词语直径(Diameter (t))

词语直径是指词语在文本中首次出现的位置和末次出现的位置之间的距离。词语直径是根据实践提出的一种统计特征。

1.4.12 首次出现位置(FirstLoc(t))

Frank 在 Kea 算法中使用候选词首次出现位置作为 Bayes 概率计算的一个主要特征,他称之为距离(Distance)。简单的统计可以发现,关键词一般在文章中较早出现,因此出现位置靠前的候选词应该加大权重。

1.4.13 词语分布偏差(Deviation(t))

词语分布偏差所考虑的是词语在文章中的统计分布。在整篇文章中分布均匀的词语通常是重要的词汇。

1.5 特征提取的一般步骤

1.5.1 候选词的确定

(1) 分词(词库的扩充)

尽管现在分词软件的准确率已经比较高了,但是,它对专业术语的识别率还不是很好,所以,为了进一步提高关键词抽取的准确率,我们有必要在词库中添加一个专业词库以保证分词的质量。

(2) 停用词的过滤

停用词是指那些不能反映主题的功能词。例如:“的”、“地”、“得”之类的助词,以及像“然而”、“因此”等只能反映句子语法结构的词语。

(3) 记录候选词在文献中的位置

为了获取每个词的位置信息,需要确定记录位置信息的方式,以及各个位置的词在反映主题时的相对重要性。

1.5.2 词语权重计算

(1) 词语权值函数的构造。

(2) 关键词抽取。

候选词的权值确定以后,将权值排序,取前 n 个词作为最后的抽取结果。

2 系统分析与实现

2.1 系统设计流程图

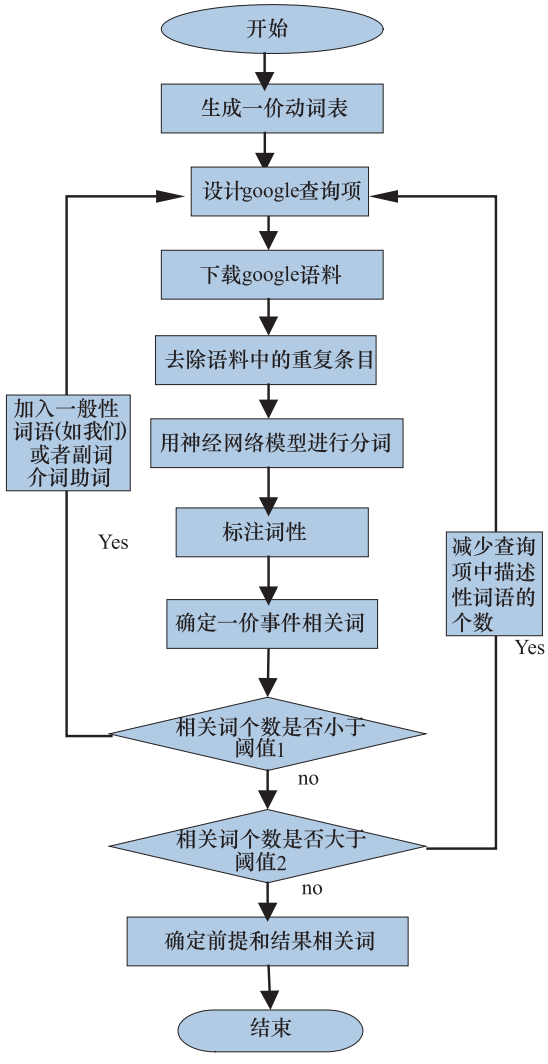

图 1 系统设计流程图

2.2　下载语料模块

下载语料模块共分为四个部分,分别为取词部分、查找部分、句子选择部分和下载语料部分。

2.3　系统实现

(1) 从《实习词表》中挑选一价动词,生成《一价动词表》,如图 2 所示:

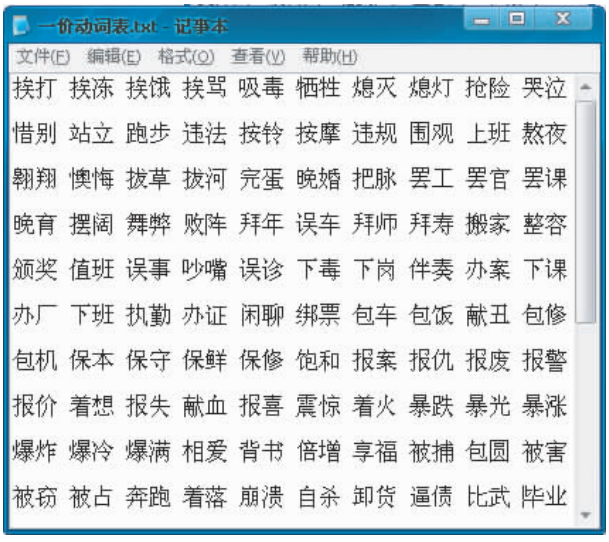

图 2　一价动词表

(2) 根据《一价动词表》设计 Google 查询项,并且根据查询项从 Google 中获得 Web 语料。

①根据《一价动词表》设计 Google 查询项,如表 1 所示。

表 1　查询项

挨打///了	挨冻///了	挨饿//了	挨骂///了
明星///吸毒	牺牲///了	熄灭///了	熄灯///了
抗洪///抢险	在///哭泣	依依///惜别	站立///着
我///跑步	违法///了	在///按铃	为客人///按摩
违规///了	群众///围观	在///上班	在///熬夜
在天空///翱翔	很///懊悔	在///拔草	在///拔河
完蛋///了	是///晚婚	为病人///把脉	工人///罢工
官员///罢官	学生///罢课	决定///晚育	老是///摆阔
考试///舞弊	败阵///下来	给爷爷///拜年	误车///了

②根据查询项从 Google 中获得 Web 语料。

以“被捕”为例获取语料的过程如图 3 所示。

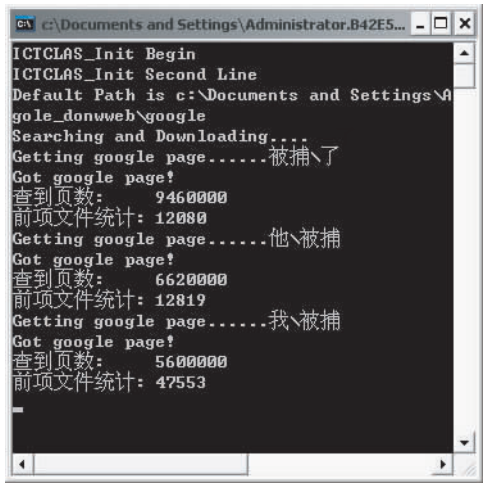

图 3　获取语料过程

2.3.3　下载的语料中有重复的语料

如:劫匪担心被捕专抢学生小钱抢了 20 元还嫌多”就有重复,一条语料是从“中新网”中获取,另一条语料是从“金鹰新闻”中获取的,这就需要我们有去除复重的工作,图 4 为对下载的语料(去重复后)分词并进行标注。

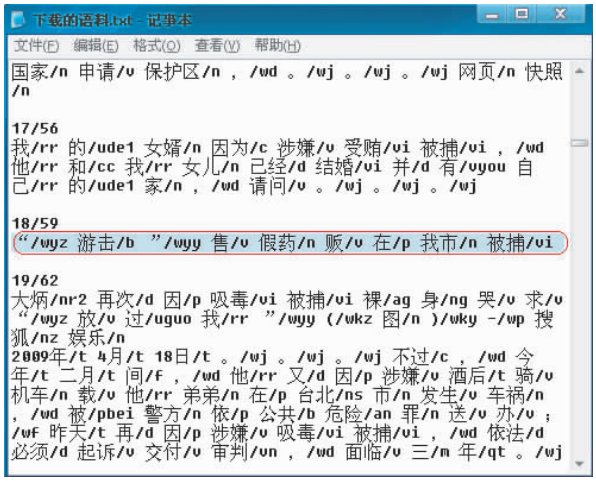

图 4　分词并标注

2.3.4　常识事件的提取

常识事件的提取:我们提取图 4 中“被捕事件”和“销售假药事件”,我们可以得出“售假药事件”是“被捕事件”的前提。

2.4　规律总结

① 编写的查询项不宜过长,查询项过长则返回

语料少。可由两三个词组成的短语作为查询项。例如:【跑步】“他每天都到操场上跑步”或“天天跑步有益身体健康。”这样的查询项太长,就不能很好的从 Google 里返回语料。而“去跑步”,“跑步了”,"跑步时" 等作为查询前项,加上一般性词语(如“我们”,“大家”,“一个”等)作为后项组成最后的查询项或者直接作为查询项,就会很好地从 Google 里返回语料,有利于我们从语料里提取相关事件。

② 注意副词和助词介词的使用

例如:【钻研】“仔细地钻研”,“认真钻研”,“钻研了”,“正钻研”,“正在钻研”,这个词语加上一些副词或介词之后其程度就会略有改变。

③ 有些动词意义不明确,需要与其他名词搭配,获得一组含义同一的查询项

A 组"遵守纪律","遵守了纪律","严格遵守纪律","严格地遵守纪律"

B 组"遵守学校纪律","遵守了学校纪律","严格遵守学校纪律","严格地遵守学校纪律"(暗含了角色在里面,即学生)。

3 结束语

本文主要论述了 3 种特征提取方法,本系统设计采用了基于统计的特征提取方法研究。这类型算法通过构造评估函数,对特征集合中的每个特征进行评估,并对每个特征打分,这样每个词语都获得一个权值。然后将所有特征按权值大小排序,提取预定数目的最优特征作为提取结果的特征子集。显然,对于这类型算法,决定文本特征提取效果的主要因素是评估函数的质量。

基于评估函数的特征提取方法是建立在特征独立的假设基础上,但在实际中这个假设是很难成立的,因此需要考虑特征相关条件下的文本特征提取方法。本文为了克服这一缺点,在对 Google 语料分析时确定了因果事件,加强了文本的特征相关性,从而提高了查询精度。

参 考 文 献

1 庞景安. Web 文本特征提取方法的研究与发展. 信息系统,2006;29(3): 338—367

2 晋耀红,苗传江. 一个基于语境框架的文本特征提取算法. 计算机研究与发展,2004;41(4):582—586

3 胡 燕,吴虎子,钟 珞. 中文文本分类中基于词性的特征提取方法研究,武汉理工大学学报,2007;29(4):132—135

Gainting Method Univalence Thing from Web Page

YANG Fan, SUN Qiang

(Jiangsu Polytechnieal University, Zhenjiang 212003, P. R. China; No. 91206 Group of PLA, Qingdao 266108, P. R. China)

[**Abstract**] Univalence thing is composed of univalence verb. For improving the searching intelligence and accuracy, using univalence thing to describe picking univalence verb form “practical verb table is acceptd to”, designing google searching item according to “univalence verb table” and gaining extensive language material from Web page. Use feature extraction method to extract interrelated verb from context. Evaluating and adjusting the Google searching item according to related word. Causal logic words can be gain which are related to univalence verb. The causal logic words will be used to enrich the searching item and proved the searching precision.

[**Key words**] feature extraction text information retrieval segmentation algorithm