中 文 信 息 学 报
第 18 卷 第 6 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 18 No. 6

文章编号:1003 - 0077(2004)06 - 0016 - 07

基于词汇吸引与排斥模型的共现词提取①

郭 锋,李绍滋,周昌乐,林 颖,李胜睿

(厦门大学 计算机与信息工程学院,福建 厦门 361005)

摘要:共现词提取在信息挖掘和自然语言处理中有着十分重要的地位。而传统的共现词提取方法仅仅局限在单一的一种统计量上,其结果十分不精确,需要人工再进行整理。本文提出了一种基于词汇吸引与排斥模型的共现词提取算法,并通过将多种常用统计量进行组合,改进了算法的效果。在开放测试环境下,所提取的共现词其用户感兴趣度为 60.87 %。将该算法应用于基于 Web 的共现词检索系统,在速度和共现词的提取精度上均取得了比较好的效果。

关键词:计算机应用;中文信息处理;共现词;词汇吸引与排斥模型;共现距离

中图分类号:TP391.3 **文献标识码**:A

Co-occurrence Word Retrieval Based on the Lexical Attraction and Repulsion Model

GUO Feng, LI Shao-zi, ZHOU Chang-le, LIN Ying, LI Sheng-rui

(Computer Science of Department, Xiamen University, Xiamen, Fujian 361005, China)

Abstract: Co-occurrence word retrieval is very important in information mining and natural language processing. But traditional co-occurrence word retrieval methods used only a single statistic method, so the result is very imprecise, and needs lots of manual collation. In this paper we present a co-occurrence words extraction algorithm based on the lexical attraction and repulsion model, and combine some common statistical methods with the algorithm to improve its effect. In the open test, our system's Interesting performance is 60.87 %. We show good performance in speed and precision when applied the algorithm on a co-occurrence search system based on web.

Key words: computer application; Chinese information processing; co-occurrence; lexical attraction and repulsion model; co-occurrence distance

1 引言

当前,Internet 作为一个巨大的信息资源库,正在以惊人的速度增长着。为了从这座信息宝库中获得需要的知识,各种信息挖掘技术应运而生,而共现词作为其中的一项重要辅助技术也在不断发展着。

所谓共现词,是指在文档中经常同时出现的词项。共现(Co-occurrence)类似于搭配(Collocation),但范围比搭配要广的多:共现词可以是习惯搭配关系的词对;也可以是属于同一词类的词对,例如 host 与 guest、tea 与 coffee 等;或者是在同一话题中经常出现的词对,例如,英国人吃早饭时习惯喝茶,所以与 tea 同现的词项常会有 morning, breakfast, butter, toast, bacon, eggs, fork

① **收稿日期**:2004 - 05 - 10
基金项目:福建省自然科学基金资助项目(A0310009);福建省重点科技资助项目(2001J005)
作者简介:郭锋,硕士研究生,主要研究方向:信息挖掘、自然语言处理.

等。

共现信息对于自然语言处理有着重要的作用。共现信息能够提高信息抽取的效果[1];共现信息也应用在文本分类系统中[2],例如在"政治新闻"一类的文档中,经常出现的共现词对有"选举—选举权,选举—初选,政权—行政区"等;将共现信息应用在特征词抽取,也能够取得良好的效果[3]。

目前共现词抽取基本只采用单一的统计量进行评估,精度很低,抽取结果还需要人工检查。有的学者采用了一些特殊的方法以提高抽取精度,比如在抽取共现信息之前先抽取文档的关键字,然后针对关键字计算共现信息[4],但这同时也造成了信息的丢失。

本文提出了一种基于词汇吸引与排斥模型的共现词提取算法,并将包括互信息在内的多种统计量进行组合,改进了算法的效果。将该算法应用于基于网络的共现词检索系统,在速度和共现词的提取精度上均取得了比较好的效果。

2 相关知识

2.1 词汇吸引与排斥模型

词汇吸引与排斥模型(Lexical Attraction and Repulsion Model,以下简记LAR Model)[5]是基于共现词能够相互吸引和排斥的语言现象建立起来的。这种语言现象表现为共现的两个词在文本和对话中同时出现的概率随着两者间距离的增加呈指数下降,这就是词汇"吸引"现象;而"排斥"现象则防止共现的两个词过于紧密地出现,当共现的两个词的距离小于一定值时同现的概率将呈指数下降。

该模型描述了词共现的三个主要特征:

1. 假设(s,t)是在文本中出现的词对,t与s的距离为k个词,那么t是s的共现词的概率将随着k的增加而指数下降。该性质可以使用单参数指数函数表示:$P_{\mu}(x) = \mu e^{-\mu x}$

对应的离散函数[5]是 $P_{\mu}(k) = (1 - e^{-\mu}) e^{-\mu k}$ (1)

2. 当k等于某个阈值,t是s的共现词的概率达到最大值;当k小于该阈值时,t是s的共现词的概率随着k的减小而指数下降;当k大于该阈值时,t是s的共现词的概率随着k的增加而指数下降。该性质可以使用带两个参数的指数函数表示:

$$P_{\mu_1,\mu_2}(x) = \frac{\mu_1\mu_2}{\mu_2 - \mu_1}(e^{-\mu_1 x} - e^{-\mu_2 x}) \qquad \mu_1 \neq \mu_2$$

对应的离散函数[5]是 $P_{\mu_1,\mu_2}(k) = \sum_{i=0}^{k} P_{\mu_1}(i)\, P_{\mu_2}(k-i)$ (2)

3. 当k足够大时,t是s共现词的概率将趋向于一个常量。使用公式可表示为:

$$P_{\mu_1,\mu_2,c}(x) = \gamma^{-1}(P_{\mu_1,\mu_2}(x) + c)$$

对应的离散函数[5]是 $P_{\mu_1,\mu_2,c}(k) = \gamma^{-1}(\sum_{i=0}^{k} P_{\mu_1}(i)\, P_{\mu_2}(k-i) + c)$ (3)

在实际应用中,我们使用公式(2)来对共现词的出现概率进行计算,并使用EM算法对统计的数据进行处理,以得到最合适的参数。假设$\theta \in \Theta$表示LAR Model的参数,$\tilde{P}(k)$是实验统计的当距离为k时共现词对出现的概率。为了使得到的曲线适合统计的数据,我们需要找到使下式达到最大值的θ:$\ell(\theta) = \sum_{k=0} \tilde{P}(k) \log P_{\theta}(k)$,对于公式(2)即找到合适的$(\mu_1,\mu_2)$,使$\ell(\mu_1,\mu_2) = \sum_{k} \tilde{P}(k) \log(\sum_{j=0}^{k} P_{\mu_1}(j)\, P_{\mu_2}(k-j))$有最大值。使用EM算法得到的迭代公式[5]

为:

$$\mu_1' = \log(1 + \frac{1}{\sum_{k=0}^{k} \widetilde{P}(k) \sum_{j=0}^{k} j P_{\mu_1,\mu_2}(j \mid k)}) \quad (4)$$

$$\mu_2' = \log(1 + \frac{1}{\sum_{k=0}^{k} \widetilde{P}(k) \sum_{j=0}^{k} (k - j) P_{\mu_1,\mu_2}(j \mid k)}) \quad (5)$$

其中, $P_{\mu_1,\mu_2}(j \mid k) = \frac{P_{\mu_1,\mu_2}(k, j)}{P_{\mu_1,\mu_2}(k)}$, $P_{\mu_1,\mu_2}(k, j) = P_{\mu_1}(j) P_{\mu_2}(k - j)$。

2.2 四种常用的评估词与词结合紧密度的统计量

表 1 四种常用统计量的公式

方法	记为	公式
互信息	MI	$log_2 \frac{P_{ts}}{P_t P_s}$
Z-Score	ZScore	$\frac{f_{ts} - \xi_{ts}}{\sqrt{\xi_{ts}(1 - \xi_{ts} \mid N)}}$
Student's t-Score	TScore	$\frac{f_{ts} - \xi_{ts}}{\sqrt{f_{ts}(1 - f_{ts} \mid N)}}$
频次	Freq	f_{ts}

评估词与词之间结合紧密度的统计量有很多[6,7],在实验中,我们使用到的统计量有四种:互信息、Z-Score、Student's t-Score 以及频次,并将这四个统计量进行组合用于评估一对词结合的紧密度。选择这四种统计量的原因是它们两两之间的相关度比较低[8],具有不同的评估性能,通过组合它们能够进一步提高共现词抽取性能。

给定词对 (s, t),表 1 列出了使用的四种统计量。其中,f_t 和 P_t 分别表示词 t 出现的频次和概率,f_{ts}和 P_{ts}分别表示词对 (s, t) 出现的频次和概率。ξ_{ts}则表示在词 t、s 独立的条件下词对 (s, t) 出现的期望值,于是有:$\xi_{ts} = P_{ts}N = P_t P_s N = f_t f_s \mid N$ (N 为训练语料库规模)。

这里通过加权策略进行组合,设第 i 种统计量对词对 (s, t) 的评估值为 $score_i(ts)$,则四种统计量的综合评估值为:

$$score(ts) = \sum_{i=1}^{4} (wt_i \times score_i(ts)) \quad (6)$$

其中 wt_i 为第 i 种方法的权重,且满足约束条件 $\sum_{}^{4} wt_i = 1$。

采用阈值型分类器来判断是否是共现词:当候选的词对的综合评估值超过给定阈值时,就认为是共现词对。

由于这四种统计量的数量级不同,在这种直接组合方式中,数量级大的统计量会湮没其他统计量的信息,因此有必要先将统计量的评估值进行归一化:

$$score_i(s_j) = score_i^*(s_j) \mid \sqrt{\sum_{\forall s_k \in TS} (score_i^*(s_k))^2} \quad (7)$$

其中 TS 是语料库,s_j 为属于 TS 的一个词对,$score_i^*(s_j)$ 和 $score_i(s_j)$ 分别是归一前和归一后第 i 种统计量对词对 s_j 的评估值。

2.3 LAR Model 与四种统计量的组合

为了更好的提高共现词抽取的性能,我们在使用四种统计量的基础上,应用了 LAR Model。对于词对 (s, t),$score(ts)$ 表示使用四种统计量组合后对 (s, t) 是共现词的评估值,那么组合了 LAR Model 后的新的评估值表示如下:

$$dis\_score(ts) = score(ts) \times P_{\mu_1,\mu_2}(k) \quad (8)$$

在这里,k 表示 s 与 t 的平均距离。

由于 LAR Model 提供的 $P_{\mu_1,\mu_2}(k)$ 值很小,所以新的评估值也减小了很多,为了适应新的评估值,需要修改原有的阈值,即将阈值乘上 $P_{\mu_1,\mu_2}(k)$ 的期望值。

3 实验和结果

一般情况下,一篇文档都只阐述一个主题,因此关于同一主题的共现词在该主题的文档中的出现几率会更高一些,所以在我们的共现词抽取系统中,选择整篇文档作为窗口单元显得更有意义;其次,利用词与词的互信息来统计共现词信息,并未考虑单个词本身的重要程度,容易引入噪音,并且对统计量带来的影响比较大,所以在使用互信息来统计共现信息前,要先对文本进行加工,滤去停用词,然后选择中高频词(可通过设定阈值确定)进行共现信息的统计计算,有助于生成更合理的共现资源。

3.1 测试集

在实验中,我们选择北京大学计算语言学研究所已经分好词并标注过的 1998 年 1 月的《人民日报》作为训练语料库,人工整理在该语料库中出现频率大于 500 次的词序列 $W_1, W_2, \cdots, W_i$,并对每个 W_i 整理出其在语料库中的共现词序列 $C_1, C_2, \cdots, C_j$,其中 $i <= 10$。部分结果如表 2 所示。

表 2 人工整理的共现词列表

C / W	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
社会主义	特色	市场经济	建设	现代化	精神文明	体制	邓小平理论	阶段	文化	制度
香港	回归	特区	一国两制	董建华	内地	繁荣	稳定	香港特别行政区	政府	贸易
研究	理论	成果	问题	分析	科学	技术	深入	调查	开发	专家
……										

3.2 LAR Model 参数的建立

根据表 2 中的数据,对所有的 (W_i, C_j) 共现词对,统计实验所需的 $\tilde{P}(k)$ 值。实现的算法如下:

```
定义整形数组 Dis_Match_Array      //其中 Dis_Match_Array[k]用于保存距离为 k 的共现词
                                  //对的出现次数,初始化置 0
定义整形数组 Dis_Total_Array      //其中 Dis_Total_Array[k]用于保存距离为 k 的所有词对
                                  //的出现次数,初始化置 0
定义实形数组 P_Array              //其中 P_Array[k]用于保存当距离为 k 时共现词对出现
                                  //的概率,初始化置 0
```

(1) 对所有的 W_i 进行如下步骤

在语料库中找到 W_i 所在的每个位置依次执行下列操作

a) 将整数 k 置为 0

b) 向前对每个词进行下列操作

i. 将 k 加一

ii. 如果该词是噪声(包括标点符号,单纯的数字,助词,不常见词以及常见词等),则停止

往下执行,返回 b 继续检查下个词

iii. 将 Dis_Total_Array[k]加一

iv. 如果该词不同于 W_i,且是 W_i 的共现词,则将 Dis_Match_Array[k]加一;如果该词与 W_i 相等,则由人工判断是否共现词。

c) 按照 a、b 的步骤类似的完成向后的搜索

(2) 计算 P_Array[k] = Dis_Match_Array[k]/ Dis_Total_Array[k]

实验所需要的参数 $\tilde{P}(k)$ 值即保存在数组 P_Array 中。具体结果曲线如图 1 所示。

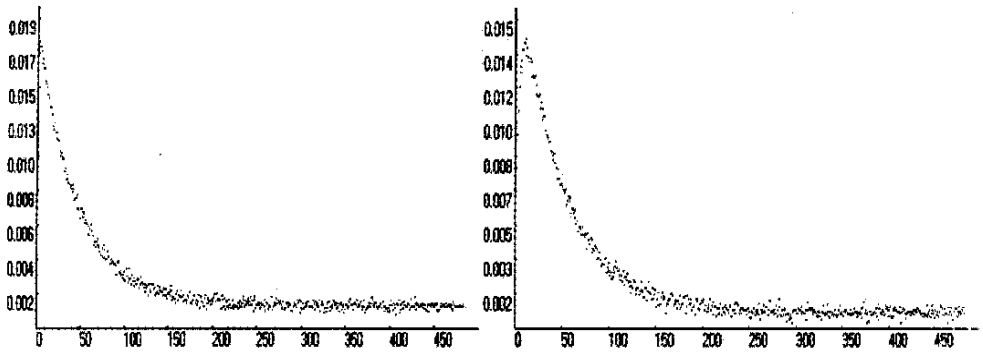

图 1 横轴表示 W_i 与 C_j 的距离 k,纵轴表示对应的 $\tilde{P}(k)$ 值,其中左图是当 $W_i \neq C_j$ 时,右图为 $W_i = C_j$ 时。

在实验中,我们发现对于共现词对 (W_i, C_j),当 $W_i \neq C_j$ 时,排斥效果不明显;而当 $W_i = C_j$ 时就有比较明显的排斥效果。这说明对于不同的两个词,距离越近则组成共现词的概率越高;而对于一篇文档中相同的两个词,排斥效果则阻止相隔太近的两个相同的词组成共现词。

将得到的 $\tilde{P}(k)$ 值代入到公式(4)、(5)中,通过迭代,最后收敛得到的参数值为

$\mu_1 = 7, \mu_2 = 0.02$ 当 $W_i \neq C_j$ 时

$\mu_1 = 0.2, \mu_2 = 0.02$ 当 $W_i = C_j$ 时

3.3 四种统计量各自的性能以及组合后的性能

在实验中,我们使用了查全率(Recall)和查准率(Precision)来表示系统的性能。公式表示为:查准率 = $\frac{\text{机器正确判定}}{\text{机器实际判定}}$,查全率 = $\frac{\text{机器正确判定}}{\text{应有共现词数}}$。在实际中,我们根据对训练语料库中频次大于 500 的所有词 W_i 分别进行共现词检索,得到共现词对序列 (W_i, C_j),对于每一个共现词对如果在人工整理的共现词对列表中存在,则视为正确判定;而机器所检索的所有共现词对的总和即机器实际判定;表 2 中所列的所有共现词对即为应有共现词数。

同时,我们还使用了兴趣度(Interesting)来表示系统的性能。公式表示为:兴趣度 = $\frac{\text{用户认为感兴趣的共现词}}{\text{机器实际判定}}$。其中"用户认为感兴趣的共现词"表示机器抽取的共现词列表中与共现词所在的文章密切相关,并对阅读该文章有所帮助的共现词数目。之所以采用这个度量方法是因为共现词的出现与特定文档相关,而通过人工阅读共现词所在的文章并对该共现词作出的评估会更有价值些。

四种统计量单独运行的性能如表 3 所列,其中以 TScore 的性能最好、Freq 的性能最差。由结果可以看出每种统计量存在着各自的问题:对于MI,容易提取出出现次数很少,但与检索词

共同出现次数比较多的词,比如在训练集中检索“中国”,则人名“华新”会被提出出来作为共现词,这是因为,“华新”共出现了6次(大于低频词的阈值),且所在的文章中均包含有“中国”,因此虽然“华新”与“中国”并无共现关系但评估值却很高;而Freq的问题在于提取出的共现词基本都是出现次数比较高的词,比如“记者”,几乎每篇文章都会出现,而又不属于高频词的范畴,因此自然会作为“中国”的共现词被抽取出来。同时表3的数据也表明部分共现词的出现并不局限于特定的主题,在各种类型的文档中都有可能出现。

表3 四种统计量单独运行的性能

方法	共现词数目	精确率	召回率	兴趣度
MI	2601	72.3%	70.1%	41.2%
ZScore	2756	76.5%	76.7%	42.7%
TScore	2671	80.6%	81.9%	44.6%
Freq	2413	65.1%	61.0%	35.1%

表4 四种统计量在组合时的权重

方法	MI	ZScore	TScore	Freq
权重	0.261	0.272	0.355	0.112

为了提高抽取共现词的能力,我们对四种统计量进行组合。在实验中,我们首先通过公式(7)对四种统计量得到的评估值进行归一化,根据表2的数据和执行的结果使用遗传算法对公式(6)中的权重进行自动调整,最后得到的权重值如表4所示:

将组合后的新的提取算法对训练集中词频大于500的词语进行共现词提取,共提取出了2613个共现词。经过对结果进行分析,发现MI和Freq方法经常出现的低频和高频词被抽取出来的问题得到了很好的解决。实验得到的精确率是88.32%、召回率是89.53%、兴趣度是55.77%,相比单一的统计量有了明显的提高。

3.4 LAR Model与四种统计量的组合

为了将LAR Model与四种统计量进行组合,我们使用了公式(8)所表示的评估值计算公式,将组合后的新评估值应用于实验,共提取出了2273个共现词,比起没有组合LAR Model的情况共现词数目少了300多,这是由于阈值的下降和距离较远的词对的评估值很低两方面的原因造成的。因此一些经常出现在文章开头和结尾的常用词如“我国”、“新华社”、“电”等就会被排除。同时实验结果表明召回率并不会因为提取的共现词的数目的减少而下降,最后的结果为精确率95.75%、召回率96.29%、兴趣度63.66%,相对于没有使用LAR Model的情况各项指标均得到了提高。

3.5 其他工作

目前,我们的系统运行在基于Web的共现词检索中。在实际使用中,我们利用对语料库建立索引提高了检索的速度;由于需要检索的语料库基本是汉语未分词的生语料,所以我们使用了厦门大学自然语言处理实验室史晓东教授开发的分词系统,然后再利用本文介绍的共现词检索程序加以检索。

在开放环境下,语料库的规模要大大于训练集,进行检索的词以及提取出来的共现词大部分在训练集中无法找到,因此无法使用查准率和查全率进行评价,而只使用兴趣度(Interesting)作为性能评价的指标。

为了比较各种统计方法在开放测试条件下的结果,表5列出了本文使用过的六种统计方法以及对应的用户的兴趣度。从数据看出,Freq方法所提取的共现词的数目比起其他的方法少了许多,这是因为在开放测试条件下,用户要求检索的可能是低频词,因此对应的共现词的出现次数也比较低,无法大于阈值,造成了提取的共现词数目的减少,同时其效果是最差的,对用户阅读文章的帮助最小。最后通过四种统计方法的组合,并组合了LAR Model之后,检索的结果就比较理想,达到了60.87%,这与训练集中的词检索的效果几乎一样好。这说明,该共

现词检索程序已经能够很好的运行在开放的环境中。同时,相比 Yuen-Hsien Tseng[4]的 53 %的兴趣度也有了很大的提高。

表 5　开放测试条件下各种统计方法的比较

方法 / 数据	MI	ZScore	TScore	Freq	四种统计量的组合	与 LAR Model 的组合
检索词的数目	100	100	100	100	150	200
提取的共现词数目	1295	1136	1172	871	1076	947
兴趣度	34.21 %	29.18 %	30.52 %	17.50 %	51.77 %	60.87 %

4　结束语

实验结果表明:在共现词检索算法中,选择整篇文档作为窗口单元,并统计中高频词的共现信息,可以改进检索的效果;将四种常用统计量进行组合能够弥补单一统计量局限在某一词类的缺点;通过使用词汇吸引与排斥模型能够更进一步提高共现词抽取的性能。

在实验中发现,分词的效果对于共现词抽取统计实验具有重要影响,如果能够在分词中更好的处理数字、地名、人名、机构名等未登录词,并结合禁用词表,那么将对共现信息的生成产生良好的影响。

由于在实验中,我们只针对训练集中词频大于 500 的词项进行统计,并且在计算共现信息前滤去了低频词,所以对稀疏数据的共现词的抽取效果比较差,将来可以根据词项相似度原理将低频词的共现词的抽取问题转换成具有相似词义的高频词的共现词抽取问题上来[9];同时,由于共现词的出现具有很强的时效性,某一个时期的共现词在另一个时期的文档中可能并不是共现词,所以将来要针对不同时期通过对不同的训练集进行训练以完善共现词检索系统。

今后,可以在文本过滤、文本分类中结合特征词的共现信息,研究其对文本过滤或文本分类的影响;也可以在非汉语共现词检索中结合词汇吸引与排斥模型,研究对其他语言的作用。

参　考　文　献:

[1] Ying Ding, IR and AI. Using Co-occurrence Theory to Generate Lightweight Ontologies[A]. Proceedings of 12th International Workshop on Database and Expert Systems Applications[C], Pages:961-965, Sept., 2001.

[2] 吴光远,何丕廉,等. 基于向量空间模型的词共现研究及其在文本分类中的应用[J]. 计算机应用,2003, 23(6): 138-145.

[3] El-Sayed Atlam, A New Method for Construction Field Association Terms Using Coroccurrence Words and Declinable Words Information[A]. Proceedings of 2002 IEEE International Conference on Systems, Man and Cybernetics[C], Volume 4, Pages:5, Oct. 2002.

[4] Yuen-Hsien Tseng, Fast Co-occurrence Thesaurus Construction for Chinese News[A]. Proceedings of 2001 IEEE International Conference on Systems, Man, and Cybernetics[C], Volume 2, Pages:853-858, Oct. 2001.

[5] Doug Beeferman, Adam Berger, John Lafferty. A Model of Lexical Attraction and Repulsion[A]. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. [C], Pages:373-380, 1997.

[6] 王丽坤,王宏,等. 文本挖掘及其关键技术与方法[J]. 计算机科学, 2002, 29(12): 12-19.

[7] 许伟,黄昌宁,等. 基于语料库的语言建模[J]. 清华大学学报, 1997, 37(3): 71-75.

[8] 罗盛芬,孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究[J]. 中文信息学报, 2003, 17(3): 9-14.

[9] Ido Dagan, Shaul Marcus. Contextual word similarity and estimation from sparse data[J]. *Computer Speech and Language*, Vol. 9, Pages:123-152, 1995. 9.