# 运用有向图进行中文分词研究

张培颖
ZHANG Pei-ying

中国石油大学(华东)计算机与通信工程学院,山东 东营 257061
College of Computer & Communication Engineering,University of Petroleum (East China),Dongying,Shandong 257061,China
E-mail:smartfrom1024@yahoo.com.cn

**ZHANG Pei-ying.Method of Chinese word segmentation based on directed graph.Computer Engineering and Applications,2009,45(22):123-125.**

**Abstract**: Chinese word segmentation is the first step for any Chinese information processing and hinders seriously its development.This paper introduces the critical technologies in the segmentation systems.It proposes a refinement of the segmentation algorithm based on the directed graph,this algorithm first constructs the Chinese segmentation directed graph,and calculates the weight of every segmentation path,last evaluates every segmentation path based on the principle of least segmentation,the mutual info of characters and the frequency of words,the highest scores on the path corresponding the correct segmentation results.Open-Test results show that the accuracy rate is more than 90%.

**Key words**: Chinese segmentation;directed graph;Chinese segmentation directed graph;segmentation path;mutual information

**摘 要**:首先说明了分词在中文信息处理中的作用,然后介绍了分词系统中的关键技术。提出了一种基于有向图的中文分词算法,该算法首先构造中文分词有向图,然后计算中文分词有向图中所有可能的切分路径,最后利用了最少分词原则、汉字之间的互信息和词语的频率等信息给中文分词有向图中的每条切分路径打分,分数最高的路径就对应正确的切分结果。开放测试结果表明分词精确率可达90%以上。

**关键词**:中文分词;有向图;中文分词有向图;切分路径;互信息
**DOI**:10.3778/j.issn.1002-8331.2009.22.040 **文章编号**:1002-8331(2009)22-0123-03 **文献标识码**:A **中图分类号**:TP301.6

中文分词是中文信息处理中的重要环节。它在中文搜索引擎、机器翻译、智能检索中有着相当重要的地位,也是智能计算、文献标引、自然语言理解和处理的基础。文中提出了一种基于有向图的中文分词算法,该算法首先构造中文分词有向图,然后计算中文分词有向图中所有可能的切分路径,最后利用了最少分词原则、汉字之间的互信息和词语的频率等信息给中文分词有向图中的每条切分路径打分,分数最高的路径就对应正确的切分结果。

## 1 关键技术

### 1.1 中文分词有向图

设待切分的中文字串 $S=C_1,C_2,\cdots,C_n$,其中 $C_i(i=1,2,\cdots,n)$为单个的字,$n$ 为串的长度,$n\geqslant 1$。建立一个结点数为 $n+1$ 的切分有向无环图 $G$,各结点编号依次为 $V_0,V_1,V_2,\cdots,V_n$。

通过下列两种方法建立 $G$ 的所有可能的词边。

(1)相邻结点 $V_{k-1},V_k$ 之间建立有向边$<V_{k-1},V_k>$,边的长度值为 $L_k$,边对应的词默认为 $C_k(k=1,2,\cdots,n)$;

(2)若 $w=C_iC_{i+1}\cdots C_j$ 是一个词,则结点 $V_{i-1},V_j$ 之间建立有向边$<V_{i-1},V_j>$,边的长度值为 $L_w$,边对应的词为 $w(0<i<j\leqslant n)$。

这样,待切分的中文字串 $S$ 中包含的所有词与切分有向无环图 $G$ 中的边一一对应,把该切分有向无环图 $G$ 称为中文字串的中文分词有向图。例如:“发展中国家”的中文分词有向图如图1所示。

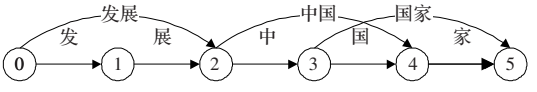


**图1 中文分词有向图示意图**

### 1.2 最少分词原则

对中文字串分词后得到的词数越少越易于对该字串的理解,这称为最少分词原则。

### 1.3 汉字之间的互信息

互信息(mutual information):对有序汉字串 $xy$,汉字 $x,y$ 之间的互信息定义为:

$$I(x;y)=\mathrm{lb}\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

其中 $p(x,y)$是 $x,y$ 的邻接同现概率,$p(x),p(y)$分别代表 $x,y$ 的独立概率。

互信息反映了汉字对之间结合关系的紧密程度:

(1)$I(x;y)>>0$,则 $p(x,y)>>p(x)p(y)$,此时 $x,y$ 之间具有

**基金项目**:中国石油大学(华东)计算机与通信工程学院青年教师创新基金(No.08120907)。
**作者简介**:张培颖(1981-),男,讲师,主要研究方向:自然语言处理、信息检索。
**收稿日期**:2008-04-30 **修回日期**:2008-07-23

可信的结合关系,并且 $I(x:y)$值越大,结合程度越强;

(2)$I(x:y)\approx 0$,则 $p(x,y)\approx p(x)p(y)$,此时 $x,y$ 之间的结合关系不明确;

(3)$I(x:y)<<0$,则 $p(x,y)<<p(x)p(y)$,此时 $x,y$ 之间基本没有结合关系,并且 $I(x:y)$值越小,结合程度越弱。

比如:字符串"重点工程"中相邻汉字对之间的互信息如图 2 所示。

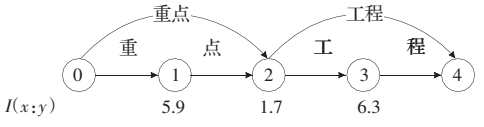


**图 2　"重点工程"相邻汉字对之间的互信息**

可以看到:$I$(重:点)=5.9>>0,说明汉字"重"、"点"之间结合程度较强;$I$(工:程)=6.3>>0,说明汉字"工"、"程"之间结合程度也较强;而 $I$(点:工)=1.7≈0,说明汉字"点"、"工"之间结合关系不明确。显然,可以利用汉字对之间的互信息来比较分词有向图各切分路径的概率大小。具体做法如下:每条切分路径是由许多个结点来连接起来的,以结点为中心,可以计算结点左右两边的两个汉字对之间的互信息,互信息越大,说明这两个汉字对之间结合程度越强,被切分开的概率就越小,从而切分路径的概率就越小。

### 1.4　切分路径的选择

每一条切分路径对应中文字串的一种切分形式。切分路径的选择就是计算每一条切分路径的权重,权重最高的路径就对应中文字串的正确切分结果。那么,路径的权重和哪些因素有关呢?经过分析发现:

(1)对中文字串分词后得到的词数越少越易于对该字串的理解,这称为最少分词原则。

假设切分路径的边条数为 $m$,那么边的条数对应中文字串分词以后的词的个数,即分词后得到的词的个数为 $m$,权重计算公式如下:

$$PowerOfNumber=\frac{1}{m}$$

显然,词的个数越少,对该字串越易于理解,权重越大。

(2)根据中文字串中的汉字与汉字之间的互信息,汉字之间的互信息体现了汉字之间联系的紧密程度,两个汉字之间的互信息越大,这两个字被分开的可能性就越小。

假设:与结点 $i$ 相邻的两个汉字分别为 $LeftC(i)$和 $RightC(i)$,可以计算切分有向图中和每个结点相邻的两个汉字的互信息:

$$I(LeftC(i):RightC(i))=\mathrm{lb}\frac{P(LeftC(i)RightC(i))}{P(LeftC(i))\times P(RightC(i))}$$

定义每条切分路径的互信息为:

$$MutualInfo=\frac{1}{n}\sum_{i=1}^{n}I(LeftC(i):RightC(i))$$

其中 $n$ 为切分路径中结点的个数。

显然,MutualInfo 反映了切分路径的各个断点之间结合的紧密程度。MutualInfo 值越大,说明各个断点之间结合程度越强,该路径成为正确切分路径的概率就越小。所以,定义切分路径的互信息权重如下:

$$PowerOfMutualInfo=\frac{1}{MutualInfo}$$

(3)根据切分之后产生的词语的频度,如果切分之后该词语的使用频度越高,说明该词语被切分成词串的概率就越高。

假设:切分后每个词语的词频为 $freq(W_k)$,那么可以定义切分路径的词频权重如下:

$$PowerOfFreq=\sqrt[l]{\prod_{k=1}^{l}freq(W_k)}$$

$l$ 为切分之后的词语个数。

最后,可以根据上述三种情况对切分路径的影响不同,定义整个切分路径的权重计算公式如下:

$$PowerOfPath=\alpha\times PowerOfNumber+\beta\times PowerOfMutualInfo+\gamma\times PowerOfFreq$$

其中 $\alpha$、$\beta$、$\gamma$ 分别为调整因子,它们反映了每种情况对句子权重的影响程度。并且满足:$\alpha+\beta+\gamma=1$。

## 2　算法描述

提出的基于有向图的中文分词算法,旨在提高中文分词的精度,其算法的流程描述如下:

(1)断句,按标点符号/空格等,生成若干个中文文本字串。

(2)对每一个中文文本字串,做:

①构造中文文本字串的分词有向图;②计算有向图中所有的切分路径;③对每条切分路径,根据 1.4 节中的方法计算每条切分路径的权重;④选择权重最大的切分路径对应的切分形式为正确的分词结果。

举"重点工程"这个句子切分为例,句子的中文分词有向图如图 3 所示。

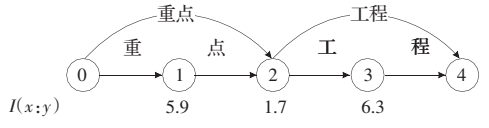


**图 3　"重点工程"的中文分词有向图**

假设互信息和词频同等重要,较最少分词原则重要一些,那么上面的式子中的 $\alpha$、$\beta$、$\gamma$ 可分为取 0.2、0.4、0.4。根据上面的中文分词有向图可以计算出所有的切分路径。

路径 1:重/点/工/程

路径 2:重点/工/程

路径 3:重点/工程

路径 4:重/点/工程

根据北大中文系詹卫东教授所提供的汉语词库,该词库共有 108 783 条词项,主要是根据 1981 年人民日报语料库统计所得,较有实用性。假设语料库的词条总数为 CORPUS_N,通过查询词库,可知路径中各个词的词频分别为:

freq(重)=952/CORPUS_N

freq(重点)=2683/CORPUS_N

freq(点)=1789/CORPUS_N

freq(工)=121/CORPUS_N

freq(工程)=4146/CORPUS_N

freq(程)=15/CORPUS_N

取 $\alpha$=0.2、$\beta$=0.4、$\gamma$=0.4,切分路径的权重计算如下:

路径 1:

$PowerOfPath$=0.05+0.053 666 698 4+0.003 702 626 2=
0.107 369 324 693 700 45

路径 2:

$PowerOfPath$=0.066 666 666 7+0.088 432 736 9+
0.004 511 362 8=0.159 610 766 382 648 96

路径 3:

*PowerOfPath*=0.1+0.100 282 810 6+0.185 405 510 1=
0.385 688 320 669 448 7

路径 4:

*PowerOfPath*=0.066 666 666 7+0.088 432 736 9+
0.124 889 773 9=0.279 989 177 473 124 3

经过比较,显然路径 3 的权重明显大于其余 3 条路径的权重,所以选择路径 3。句子的正确切分结果为:重点/工程。

当然,分词系统也可以根据侧重点不同,动态地调整 $\alpha$、$\beta$、$\gamma$ 这三个参数,以适应于不同的应用场合。

## 3 实验结果及分析

提出的基于有向图的中文分词算法,其目的是提高自动分词的精度,实验表明取得了较好的结果。封闭和开放测试结果参见表 1。

**表 1 封闭和开放测试结果**

| 测试类型 | 测试集句子数 | 测试集词数 | 错误个数 | 精度/(%) |
|---|---|---|---|---|
| 封闭测试 1 | 100 | 923 | 47 | 94.90 |
| 封闭测试 2 | 300 | 3 015 | 103 | 96.58 |
| 开放测试 1 | 100 | 958 | 83 | 91.30 |
| 开放测试 2 | 300 | 2 894 | 116 | 96.00 |

为了进一步改善分词结果,将加入规则来解决部分交叉歧义和组合歧义。另外,根据不同的场合选择不同的 $\alpha$、$\beta$、$\gamma$ 这三个参数,以增强分词系统的适用性。

## 4 结语

提出的基于有向图的中文分词算法,在构造中文分词有向图的时候,每个汉字当作一个词来对待,这样就不会漏掉可能的切分结果,系统的召回率为 100%。计算有向图中的所有路径其实采用的是全切分算法,然后根据最少分词原则、词语之间的互信息和词语频率综合考虑,给每条切分路径打分,分数最高的切分路径对应正确的切分结果。实验表明,该算法具有较高的准确度。

**致谢**:研究实现过程中使用了由北京大学计算语言学研究所和富士通研究开发中心有限公司共同制作的经过标注的 1998 年人民日报语料库,特此致谢。

## 参考文献:

[1] 曹娟,周经野.一种计算汉字串之间相关程度的新方法[J].中文信息学报,2004,18(4):55-59.
[2] 刘群,张华平,俞鸿魁.基于层叠隐马模型的汉语词法分析[J].计算机研究与发展,2004,41(8):1421-1429.
[3] 孙晓,黄德根.基于动态规划的最小代价路径汉语自动分词[J].小型微型计算机系统,2006,27(3):516-519.
[4] 李大农,董慧.汉语分词有向图的快速生成算法[J].情报学报,2004,23(1):36-39.
[5] 赵铁军,吕雅娟,于浩,等.提高汉语自动分词精度的多步处理策略[J].中文信息学报,2000,15(1):13-18.
[6] 张培颖,李村合.一种改进的上下文相关的歧义字段切分算法[J].计算机系统应用,2006(5):46-48.

(上接 28 页)

是 Vague 值 $x$ 和 $y$ 之间的相似度量的伴随函数,而 $x^{<m>}=(t_x^{<m>}, f_x^{<m>}, \pi_x^{<m>})$ 和 $y^{<m>}=(t_y^{<m>}, f_y^{<m>}, \pi_y^{<m>})$ 分别是 $x$ 和 $y$ 的 $(\alpha,\beta)$ 扩展的第 $m$ 次 Vague 值($m$=0,1,2,…)。

注:当 $m$=0,$k$=$h$=1 时,公式(5)也是文献[6]所提出的 Vague 值 $x$ 和 $y$ 之间的相似度量公式(2)。

**定理** 7 在定理 4 的条件下,则函数

$$M(A,B)=\begin{cases}\text{任意,当 } A=B=\{(0,0,1),(0,0,1),\cdots,(0,0,1)\}\text{时}\\ T(A,B),\text{其他}\end{cases}$$

是 Vague 集 $A$ 和 $B$ 之间的相似度量。其中

$$T(A,B)=\frac{1}{n}\sum_{i=1}^{n}T(x_i,y_i)$$

是 Vague 集 $A$ 和 $B$ 之间的相似度量的伴随函数。

**定理** 8 在定理 5 的条件下,则函数

$$WM(A,B)=\begin{cases}\text{任意,当 } A=B=\{(0,0,1),(0,0,1),\cdots,(0,0,1)\}\text{时}\\ WT(A,B),\text{其他}\end{cases}$$

是 Vague 集 $A$ 和 $B$ 之间的加权相似度量。其中

$$WT(A,B)=\sum_{i=1}^{n}\omega_i * T(x_i,y_i)$$

是 Vague 集 $A$ 和 $B$ 之间的加权相似度量的伴随函数。

## 3 应用实例

**例** 1[8] 设有论域 $U=\{u_1,u_2,u_3\}$ 上的标准模式 $A_1,A_2,A_3$ 和待识别模式 $B$,都有用三维表示的 Vague 集如下:

$A_1=\{(0.1,0.1,0.8),(0.5,0.1,0.4),(0.1,0.9,0.0)\}$
$A_2=\{(0.5,0.5,0.0),(0.7,0.3,0.0),(0.1,0.8,0.1)\}$
$A_3=\{(0.7,0.2,0.1),(0.1,0.8,0.1),(0.4,0.4,0.2)\}$
$B=\{(0.4,0.4,0.2),(0.6,0.2,0.2),(0.0,0.8,0.2)\}$

在公式(3)中取 $k$=$h$=1,$m$=2,计算 Vague 集之间的相似度量,得 $M(A_1,B)$=0.767,$M(A_2,B)$=0.902,$M(A_3,B)$=0.375。于是 $\max\{M(A_1,B),M(A_2,B),M(A_3,B)\}=M(A_2,B)$,则由 Vague 集的模式识别规则[9],知待识别模型 $B$ 应归属于标准模型 $A_2$。

## 4 结束语

在文献[7]找到的高区分能力公式[6]的基础上,提出两类 Vague 集之间的相似度量,自然可称之为高区分能力的相似度量。实例表明这些公式是实用的。

## 参考文献:

[1] Zadeh L A.Fuzzy set[J].Information and Control,1965,8(3):338-353.
[2] Atanasov K.Intuitionistic fuzzy sets[J].Fuzzy Sets and Systems,1986,20:87-89.
[3] Gau W L,Buehrer D J.Vague set[J].IEEE Transaction on Systems,Man,and Cybernetics,1993,23(2):610-614.
[4] 符晓芳,王鸿绪.Vague 集间的相似度量及其在正文检索中的应用[J].计算机工程与应用,2008,44(12):151-153.
[5] 刘华文,王凤英.Vague 值的转化与相似度量[J].计算机工程与应用,2004,40(32):79-81.
[6] 黄国顺,刘云生.Vague 集相似度量及其在模式识别中的应用[J].复旦学报:自然科学版,2004,43(5):869-873.
[7] 朱振国,王国胤.Vague 集相似度量[J].计算机科学,2008,35(9):220-225.
[8] Liang Z Z,Shi P F.Similarity measures on intuitionistic fuzzy sets[J].Pattern Recognition Letters,2003,24:2687-2693.
[9] 刘华文.模糊模式识别的基础——相似度量[J].模式识别与人工智能,2004,17(2):141-145.