文章编号:1001 - 9081(2011)12 - 3275 - 03 doi:10.3724/SP. J. 1087. 2011. 03275

自动文摘中的冗余句消除方法

程传鹏,杨要科
(中原工学院 计算机学院, 郑州 450007)
(33367302@qq. com)

摘 要:针对自动文摘的信息冗余问题,提出了一种冗余语句消除的方法。利用《同义词词林》来定义词语语义距离计算公式,根据词语的相似度,建立主题词和主题句之间的一一对应关系,借用编码理论中海明距离的理论,得到了文摘中主题句的相似度,设置阈值过滤掉相似度较高的主题句,从而实现了主题句的约简。实验结果证明,该方法提高了文摘的精度。

关键词:自动文摘;信息冗余;语义距离;海明距离
中图分类号:TP391.1 **文献标志码**:A

Method for elimination of redundant sentences in automatic abstraction

CHENG Chuan-peng, YANG Yao-ke
(*School of Computer Science, Zhongyuan Institute of Technology, Zhengzhou Henan* 450007, *China*)

Abstract: To solve the problem of information redundancy in automatic abstraction, this paper proposed a method for eliminating redundant sentences in automatic abstraction. Firstly, similarity of words was defined based on *TongYiCi CiLin*. And then, correspondence between topic words and subject sentence was established based on the similarity of words, the similarity of subject sentence was got based on the theory of Hamming distance in encoding theory, and high similarity sentences were reduced by threshold. The experimental results show that the method greatly improves the accuracy of abstraction.

Key words: automatic abstraction; information redundancy; semantic distance; Hamming distance

0 引言

自动文摘应该以尽可能少的文字,最大限度地体现原文所表达的意思。通过自动文摘系统生成的主题句,并不能完全作为文摘提交给用户。因为经过系统初步筛选出的主题句,往往具有较多的冗余信息。

目前,对自动文摘中信息冗余的研究,主要集中在基于词语共现的信息冗余[1-3],有些文献虽然提到了语义信息冗余[4],但并没有给出具体的解决方案。本文针对自动文摘中主题句的冗余现象,比较详尽地描述了自动文摘中消除语义冗余的方法。

1 基于《同义词词林》的词语相似度计算

主题句的相似度主要取决于句中词语语义的相似度。目前基于词语语义相似度的计算,主要采用的是刘群等人[5]提到的方法,该文中词语语义相似性计算公式是基于《知网》[6]的,并将实体概念语义分为 4 个部分,分别计算 4 个部分的相似度,实体的整体概念相似度计算公式为:

$$Sim(S_1,S_2) = \sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2)$$

可以看出在该公式中,有 4 个参数需要设置,人为因素较多,稳定性较差,而且计算量偏大。考虑到稳定性和计算量的问题,本文采用了《同义词词林》来计算词语的相似度,排除了人为选择参数,而且计算量偏小。《同义词词林》是梅家驹等人于 1983 年编纂而成,不仅包括了一个词语的同义词,也包含了一定数量的同类词,即广义的相关词[7]。从《同义词词林》的构造结构来看,很容易想到用树结构来表示,如图 1 示。

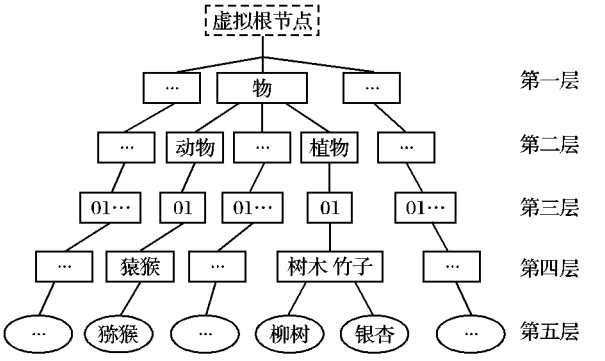

图 1 《同义词词林》语义树形图

第一层是虚拟的根节点,第二层有 12 个节点,表示的是大类别,分别有“人”、“物”、“时间与空间”、“抽象事物”、“特征”、“动作”、“心理活动”、“活动”、“现象与状态”、“关联”、“助语”、“敬语”12 个类别,第三层表示每一大类下面的中类别,共有 94 个中类。第四层节点表示的是中类别下的小类别,共有 1 428 个小类别。第五层共有 3 925 个节点表示小类别下的词群。叶子节点表示的是标题词。为了便于后文的讨论,依据语义树形图给出如下几个定义:

定义 1 绝对高度($Height(P_i)$),指的是节点到根节点

收稿日期:2011 - 06 - 02;**修回日期**:2011 - 08 - 12。 **基金项目**:河南省教育厅自然科学项目(2008B520046)。

作者简介:程传鹏(1977 -),男,河南郑州人,讲师,硕士,主要研究方向:模式识别、自然语言处理; 杨要科(1978 -),男,河南南阳人,讲师,硕士,主要研究方向:软件工程。

的路径长度。比如：$Height$(“动物”) = 3。

定义 2 密度($Density(P_i)$)，指的是节点的兄弟节点数与同一层中所有节点数的比值，文中用 $Density(P_i)$ 表示，计算公式如下：

$$Density(P_i) = \frac{sum(brother_i)}{sum(layer_i)}$$

定义 3 重合度。两个节点第一次到达同一个父节点所经过的最长路径长度，文中用 $Length(P_i,P_j)$ 表示。比如：$Length$(“柳树”,“猴子”) =4。

从图 1 的语义树形图中，可以得出如下结论：

1)对于重合度相同的节点对，处于语义树较高层的，其语义距离较大。比如说：“动物”和“植物”、“柳树”和“银杏”，这两对词语间的重合长度都是 1，但前一对词(“动物”、“植物”)绝对高度为 2，后一对词(“柳树”,“银杏”)绝对高度为 5。

2)对于绝对高度相同的节点对，如果位于语义树中高密度区域，其语义距离应大于低密度区域。这是因为《同义词词林》在分类上粗疏不均所致，有些类别分得比较细，有些类别相对于分得较粗。

Lin 等人[8]认为任何两个事物的相似度取决于它们的共性(commonality)和个性(difference)，并从信息理论的角度给出任意两个事物相似度的通用公式：

$$Sim(x,y) = \frac{p(common(x,y))}{p(description(x,y))}$$

其中：$common(x,y)$ 描述 x,y 共性所需要的信息量的大小，$description(x,y)$ 描述出 x,y 所需信息量大小。

在语义树形图中，节点共性主要体现在两个节点的父节点的高度，个性主要体现在同一层次节点所在分支的密度和节点之间重合度。综合考虑节点的共性信息和个性信息，本文给出如下的词语语义距离计算公式：

$$Dist(W_i,W_j) = \frac{Length(W_i,W_j) + Density(W_i) + Density(W_j)}{Length(W_i,W_j) + Height(pnode)}$$

其中 $Height(pnode)$ 表示 W_i,W_j 共同父节点的绝对高度。

2 主题句消冗的关键技术

为了消除掉自动文摘中的冗余的主题句，需要计算所有主题句之间的相似度，并对相似度过大的主题句进行删减。其中需要涉及到的关键技术包含词语相似度的计算、语义距离表的构建、主题句相似度计算等几个方面，下面一一进行介绍。

2.1 语义距离表的构建

依据词语之间的语义相似度，本文构造了一个词义距离表，结构如表 1 所示。

表 1 词义距离表

词	W_1	W_2	…	W_j	…	W_n
W_1	0	$Dist(w_2,w_1)$	…	$Dist(w_j,w_1)$	…	$Dist(w_n,w_1)$
W_2	$Dist(w_1,w_2)$	0	…	$Dist(w_j,w_2)$	…	$Dist(w_n,w_2)$
⋮	⋮			⋮		⋮
W_i	$Dist(w_1,w_i)$	$Dist(w_2,w_i)$		$Dist(w_j,w_i)$	…	$Dist(w_n,w_i)$
⋮	⋮			⋮		⋮
W_n	$Dist(w_1,w_n)$	$Dist(w_2,w_n)$	…	$Dist(w_j,w_n)$	…	0

表 1 由词 $W_1,W_2,\cdots,W_i,\cdots,W_n$ 构成表的二维的坐标元素。W_i 表示文档经过分词后所得到的所有词条，其中不包括停止词。表的第 i 行 j 列元素 $Dist(w_i,w_j)$ 表示 w_i 与 w_j 的词义距离。$0 \leqslant Dist(w_i,w_j) \leqslant 1$。如果 $Dist(w_i,w_j) = 1$，说明这两个词语意完全相反；如果 $Dist(w_i,w_j) = 0$，说明这两个词语意完全一致，词语和其本身的语义距离也为 0。

文档中经过分词后，往往形成成千上万个词语。如果直接进行字符串的匹配非常耗时间。为了方便、快捷地在语义词典查找两个词的语义距离，二维数组的下标可以通过词语首字的 Hash 码来计算：

$$i = (c_1 - 176) \times 94 + (c_2 - 161)$$

其中 c_1 和 c_2 是词首字的区码和位码，对于首字相同的词语，则按顺序存放。

2.2 句子相似度的计算

海明距离是信息论中一个基本概念，能够反映两码字之间的差异，进而提供码字之间的相似程度的客观依据[9]。海明码距离计算公式如下。

令 $x = x_1,x_2,\cdots,x_i,\cdots,x_n; y = y_1,y_2,\cdots,y_i,\cdots,y_n, x_i \in [0,1], y_i \in [0,1]$，它们之间的海明距离(即相异度)可以表示为：

$$Dist(x,y) = \left(\sum_{i=1}^{n} x_i \oplus y_i\right)/n$$

其中⊕表示异或加运算。

假设有一对对码字 X = {0010 1001}，Y = {1001 0011}，它们的距离计算过程如下：

X 异或 Y = {1011 1010}

$X \oplus Y = 5$

$Dist(x_1,y_1) = 0.625$

对于文摘中的主题句，可以将原始文档中的主题词作为码字，然后由上述的方法获得主题词与主题句中，每个词的语义距离。

设文档主题词系列 $\{TS_1,TS_2,TS_3,\cdots,TS_i\cdots,TS_n\}$，文摘中待比较的句子 A，经过分词并去掉停止词后词序列为 $\{A_1,A_2,A_3,\cdots,A_i,\cdots,A_m\}$，文摘中待比较的句子 B，经过分词并去掉停止词后词序列为 $\{B_1,B_2,B_3,\cdots,B_i,\cdots,B_k\}$。由于海明码的取值只能是 1 或者 0，这里设置一个阈值 $\beta(\beta \geqslant 0)$。如果 $\min(Dist(TS_i,A_j)) \leqslant \beta$，那么句子 A 第 i 个码值为 1；反之第 i 个码值取为 0。

下面以一个具体例子，来做进一步的说明。

设文档中的主题词为 TS：

TS = {水果，维生素，丰富，营养，健康，抵抗力}

待比较的句子：

S_1 = “苹果富含大量有益健康的维他命”

S_2 = “梨子含有很多提高抵抗力的维生素”

S_3 = “动物的脂肪里包含有大量的脂肪酸”

经过分词，并去掉无意义的停止词后：

W_{S_1} = {苹果，富有，大量，有益健康，维他命}；

W_{S_2} = {梨子，含有，提高，抵抗力，维生素}

W_{S_3} = {动物，脂肪，包含，大量，脂肪酸}

主题词与 S_1、S_2 中各词语语义距离分别如表 2 ~ 4 所示。

由于海明距离的计算，要求码字的各位取值要么为 0，要么为 1。所以，这里设置阈值 $\beta = 0.1$，语义距离大于 0.1 的码值取为 0；反之取为 1。因此 S_1、S_2、S_3 的码

值分别为:

$S_1 = \{111010\}$

$S_2 = \{111001\}$

$S_3 = \{010000\}$

它们之间的语义距离经过计算分别为:

$Dist(S_1, S_2) = S_1 \oplus S_2 = 0.33$

$Dist(S_1, S_3) = S_1 \oplus S_3 = 0.5$

$Dist(S_2, S_3) = S_2 \oplus S_3 = 0.5$

经过计算发现发现 S_1 和 S_2 语义距离要小于 S_1 和 S_3,计算结果比较符合直观语义。

表 2　S_1 中各词与主题词义距离表

词	水果	丰富	维生素	营养	健康	抵抗力
苹果	0.08	0.00	0.68	0.83	0.83	0.84
富有	0.90	0.05	0.87	0.95	0.86	0.89
大量	0.95	0.03	0.75	0.75	0.80	0.90
有益	0.95	0.91	0.95	0.95	0.95	0.95
健康	0.95	0.91	0.95	0.95	0.00	0.35
维他命	0.95	0.91	0.01	0.95	0.95	0.95

表 3　S_2 中各词与主题词义距离表

词	水果	丰富	维生素	营养	健康	抵抗力
梨子	0.02	0.78	0.56	0.90	0.90	0.90
含有	0.85	0.90	0.89	0.95	0.82	0.86
很多	0.90	0.03	0.90	0.90	0.90	0.90
提高	0.95	0.91	0.95	0.95	0.95	0.95
抵抗力	0.95	0.91	0.95	0.95	0.35	0.00
维生素	0.72	0.85	0.00	0.86	0.90	0.95

表 4　S_3 中各词与主题词义距离表

词	水果	丰富	维生素	营养	健康	抵抗力
动物	0.55	0.65	0.70	0.72	0.90	0.90
脂肪	0.90	0.90	0.68	0.90	0.90	0.90
包含	0.85	0.76	0.96	0.85	0.68	0.78
大量	0.85	0.01	0.68	0.78	0.82	0.90
脂肪酸	0.75	0.90	0.72	0.78	0.80	0.90

2.3　消除冗余主题句的过程描述

在上文论述的基础上,文摘中主题句冗余信息的消除步骤如下:

1)把文中的所有主题词作为码字。

2)对文摘中所有的主题句进行分词,并过滤掉停止词。

3)依据词义距离表,得到主题句中每个词条与码字的语意距离值,形成主题句的码字系列。$w = x_1, x_2, \cdots, x_i, \cdots, x_n$。

4)根据所设置的阈值,来决定码值取 1 还是取 0。

5)根据公式计算相似度,得到主题句之间的相语义距离 $Dist(S_i, S_j)$,计算公式为

$$Dist(x, y) = \sum_{i=1}^{n} x_i \oplus y_i$$

6)设置一个阈值,将相似度小于阈值的主题句进行约减。

7)按照主题句在原文中的顺序进行输出,最终产生较为理想的文摘。

3　实验及评价

对自动文摘冗余信息的评价,目前还没有一种很好的方法。文摘中冗余信息消除的主要工作集中在对句子的相似度的比较上,为了比较全面地评估本文算法,本文中提出冗余率指标来衡量文摘的精度,它的公式定义如下:

$$\text{冗余率} = \frac{\text{文摘中相似句子的总数}}{\text{文摘中句子的总数}} \times 100\% \qquad (11)$$

本文采用通过多个人工专家分别打分,这里假设人工专家冗余率为 0%。本文采集了新浪网上 2010 年的 12 000 多个主题页面,其中包括体育、财经、环保、教育、房产、汽车七个主题,采用机械文摘的方法,形成原始文摘。分别以本文方法与传统的词语共现的方法进行比较。比较结果如表 5 所示。

表 5　不同方法的冗余率实验结果　　%

类别	人工专家	词语共现方法	本文方法
体育	0	25	12
财经	0	15	10
汽车	0	20	10
房产	0	18	8
军事	0	23	8

从表 5 中的实验数据可以看出,本文方法在很大程度上降低了文摘的冗余率,从而提高了文摘的精度,因而本文中的方法具备有一定的实用性。

4　结语

如何以最简练的句子的从文档中提取“主题思想”,已经成为了自动文摘需要迫切解决的一个关键技术。本文利用主题词作为码字,通过构造词的语义距离来计算主题句之间的语义距离,从而得出主题句之间相似度;过滤掉相似度较高的主题句,得到较为精炼的文摘。但是,本文在计算词义距离时,并没有考虑到《同义词词林》中的未登录词,这将在一定程度上影响词语相似度计算的准确性,在下一步的工作中,将对未登录词的语义相似性做进一步的研究。

参考文献:

[1] 张奇,黄萱菁,吴立德.一种新的句子相似度度量及其在文本自动摘要中的应用[J].中文信息学报, 2005, 19(2): 93 - 96.

[2] 张其文,李明.文本主题的自动提取方法研究与实现[J].计算机工程与设计,2006,27(15):2743 - 2766.

[3] 傅间莲, 陈群秀.基于规则和统计的中文自动文摘系统[J].中文信息学报, 2006, 20(5): 10 - 16.

[4] 基于文本聚类的自动文摘系统的研究与实现[J]. 计算机工程, 2006, 32(4): 30 - 33.

[5] 刘群,李素建.基于《知网》的词汇语义相似度的计算[EB/OL].[2011 - 02 - 15]. http://wenku.baidu.com/view/b213af9951e79b8968022660.html.

[6] 董振东, 董强. 知网[DB/OL].[2011 - 02 - 15]. http://www.keenage.com.

[7] 梅家驹, 竺一鸣, 高蕴琦, 等.同义词词林[M]. 上海: 上海辞书出版社, 1993.

[8] LIN DEKANG. An information-theoretic definition of similarity Semantic distance in WordNet[EB/OL].[2011 - 02 - 15]. http://www-rohan.sdsu.edu/~gawron/mt_plus/readings/sim_readings/similarity_lin_98.pdf.

[9] 周荫清.信息理论基础[M].北京:北京航天航空大学出版社, 1993.