

基于多因素的中文文本主题自动抽取方法

刘金岭, 谈 芸, 李健普, 袁 娜

(淮阴工学院 计算机工程学院, 江苏 淮安 223003)

摘 要:提出了一种基于多因素的文本主题的提取方法,并着重讨论了相应的权值体系。根据概念间的相互关系,对同义概念进行语义归并和上下位概念进行语义聚焦。对于给定的文本,先进行特征词抽取进而生成代表主题概念的重要词汇。综合语句所在位置、语句中的标题、语句中所含重要词汇等多因素构造语句权值表达式,在此基础上,采用主题句选择算法保证每一个主题句被选中,同时解决了主题句的去重问题。实验结果表明,该方法具有较高的抽准率。

关键词:主题句;主题抽取;文本主题;重要词汇;语句权值

中图分类号:TP181

文献标识码:A

文章编号:1673-629X(2010)07-0072-04

Automatic Extraction Method of Chinese Text Theme Based on Multi-Factor

LIU Jin-ling, TAN Yun, LI Jian-pu, YUAN Na

(Dept. of Computer Eng., Huaiyin Institute of Technology, Huaian 223003, China)

Abstract: A multi-factor based on the theme of the text extraction methods is presented, and particularly described the algorithm and the corresponding weight system. Analysis of the interrelation between the concept were the synonymy merging and the superior concept and sub-concept were semantic focusing. For a given text, its features were extracted firstly and the representative of the theme of generating an important words. Consolidated statement location of the statement in the title, key words and other statements contained in the statement that the right to construct the value of multi-factor expression. On this basis, the use of topic sentence selection algorithm to ensure that every sentence of a theme is selected. Meanwhile, remove the redundant sentences from candidates of topic sentences. The experimental results indicate that the method has higher precision.

Key words: subject sentence; subject extraction; text subject; important words; sentence weight

0 引言

文本主题抽取对快速浏览和查询文本资料有着非常重要的意义。一般地,主题抽取的方法大都是利用各种加权算法,并计算关键词对文本主题的重要程度,选定那些重要程度大的关键词^[1]。目前国内外的相关研究中,有些方法是利用词汇频率来提取文本主题^[1];而文献[2]是从语言理解的角度进行了主题抽取;文献[3,4]则是利用关键词匹配和关键词统计的方法抽取主题,所有这些方法都没有考虑表达主题的不同用词之间的语义关联。在国外研究中也涉及到了这方面的内容,文献[5]是利用使用 TF * PDF 算法从日文新闻中提取主题;而文献[6]则利用相关度对词的共现进行

分析,建立词之间的语义关联,进一步生成代表主题概念的种子词类。研究表明,解决同一概念的不同语言表达形式语义的关联问题,在目前可能达到的目标也许只能通过机器学习,对原始语料中概念之间语义关联进行挖掘^[7]。

基于以上原因,文中首先从文本中不同词汇之间的语义关联出发,处理文本词汇的同义关系、上下位关系及文本语句间的相似关系。进而在分析文本语句时根据“兼顾各个方面因素,同时又有所侧重”的原则,综合多方面的因素提出了基于语句权值体系的计算方法。为提高文本主题句选择算法的准确率打下了坚实的基础。

1 语句的权值

一般来说,对文本的主题句抽取主要完成如下两个步骤:一是抽取文本中表达主题的重要词汇;二是正确地评估文本中各语句表达该文本内容的重要性,

收稿日期:2009-11-25;修回日期:2010-02-16

基金项目:淮安市科技项目(HAG09061);江苏省大学生实践创新训练项目(312509001)

作者简介:刘金岭(1958-),男,教授,研究方向为数据仓库及文本数据挖掘。

从中挑选出那些表达文本内容最强的句子作为主题候选句。评估句子重要性的方法目前通常采用计算句子中词的权值、句子间的相互关系以及借助文本的结构形式来评估句子的重要性^[8,9]。

1.1 词的权值计算及调整

一般来讲,具有较高权值的词应该含在文本主题句中。对词的权值的计算,目前大都是利用 TFIDF 公式。以往计算词权值都是假定在所讨论的文本中的关键词对于文本主题重要性的程度是独立于其他词的。这样做的目的主要是为了简化问题的处理工作,其实不然,这样做却忽略了词与词间的相互关系。其实在真实文本中,词汇或字串之间一般联系性是很强的,彼此之间都相互依赖。因此在大量文献研究中关于文本词汇之间不具有相关性的假定使得对于文本处理应用产生了不准确性,这样就会增大文本自动处理应用的出错率。

在同一文本中的词汇、字串之间一般都存在着很强的依赖关系,如上下位关系、同义关系等,因此,对词汇、字串之间的这些关系进行分析将有助于提高文本分析的准确性。例如,“电瓶车”、“电动车”和“二轮电动车”等表达了相同的概念,而它们又可以用上位词“自行车”进一步概化。文中基于《知网》概念库,通过概念调整、概念同义归并等重新调整词的权值计算。

文中根据概念间的同义、上下位等相互关系进行同义词汇和层次词语等之间的归并。在很多情况下,文本中可能会用多个不同的同义词去表示同一个概念。同义词的归并是一种最常见的概念归并。如果在文本中出现多个同义词,可以用代表该同义词簇的标准字串统一替换,并且在计算主题字串的权重时,对它们统一进行度量。例如词语 C 出现在文本 ST 中, C 的同义词 C_1, C_2, \dots, C_n 也在 ST 中出现,则可对这 $n+1$ 个概念进行归并,并调整相应的权值度量:

$$\overline{TF(C, ST)} = TF(C, ST) + \sum_{i=1}^n TF(C_i, ST) \quad (1)$$

$$\overline{TF(C_i, ST)} = 0 \quad i = 1, 2, \dots, n \quad (2)$$

其中 $TF(C, ST)$ 和 $TF(C_i, ST)$ 为概念合并前 C 和 C_i 在文本 ST 中出现的频率; $\overline{TF(C, ST)}$ 和 $\overline{TF(C_i, ST)}$ 为概念归并后 C 和 C_i 在文本 ST 中的频率度量,其中 $\overline{TF(C_i, ST)} = 0$ 是保证词汇 C 的同义词权值不再重复计算。

如果在同一文中出现了上下位概念也可以对它们进行概念归并,同时进行权值的调整。例如在某文本中出现了多个概念的共同上位概念,可以利用该上位概念来表示这些概念的主题。文中对于文本中出现的上

下位概念,用公式(3)、(4)来调整它们的权值:

$$\overline{W_{sup}} = W_{sup} + \frac{1}{k} \sum_{i=0}^{k-1} W_{sub}(i) \quad (3)$$

$$\overline{W_{bott}(i)} = \frac{k-1}{k} W_{bott}(i) \quad (4)$$

在公式(3)、(4)中,用 $W_{sup}, W_{sup}(i)$ 来表示式(1)、(2)计算后的上位、下位字串的权值, $i = 0, 1, \dots, k-1$, 而 $k(k > 1)$ 为下位字串的数目;用 $\overline{W_{sup}}$ 和 $\overline{W_{bott}}$ 分别表示上位词语和下位词语归并后的权值。

一般的,因为上、下位的关系都是一对多的关系,即一个上位概念对应于多个下位概念,是一种 $1:n$ 关系。用式(3)、(4)是来调整上、下位概念权值的意义是:一方面考虑到在文本中下位概念对上位概念的有增强的作用,另一方面也考虑到了这样调整上、下位关系及其权值不会影响整个文本主题的内容。

在文本主题的抽取中,文本所含语句的数量应和所抽取主题句的数量关联在一起。由文献[10],自动抽取的主题中语句数量往往是原始文本句子总数的 20% ~ 30%。作者对若干篇中文文本做了相关实验和参考了大量文献,文本主题的语句数量设定为原文本语句数量的 15% 即可。文中利用中文文本的特点给出一种确定文本主题重点词的计算方法。

算法 1:

1) 对训练短信文本集进行分词:分词方法采用中国科学院计算所的 ICTCLAS 分词系统,去除停用词,转化为文本向量集,得到文本的一个词序列:

$$ST = \{W_1, W_2, \dots, W_m\}$$

2) 利用式(3)、(4)进行同义概念的归并和上下位权值调整;

3) 利用香农信息论^[11]对由步骤1)所转化的文本向量集进行特征抽取;

4) 假设 ST 符合离散的概率分布 $p(W)$, 其中随机变量 W 在词汇集中取值,用 $H(W_i)$ 表示词 W_i 在文本中所含的信息量,计算公式如下:

$$H(W_i) = -TF(W_i, ST) \times \log[p(W_i)] \quad (5)$$

其中 $TF(W_i, ST)$ 表示词 W_i 在文本 ST 中出现的频率, $p(W_i)$ 为词 W_i 的概率分布。

5) 根据参考文献[12],重点词汇的概率不应低于 0.175。可以依次选取文本 ST 的重点词汇。

1.2 语句标题的权值

文本中的标题往往反映文章的主题。标题词在句子中出现的次数也是句子重要性一个因素。语句 L_i 所含的标题词权值按如下公式计算:

$$W_{Title}(L_i) = \frac{TWA}{TTA} \quad (6)$$

其中, $W_{Title}(L_i)$ 表示标题语句的权值; TWA 表示文本

句子中包含标题词的数量;TTA 表示文本中包含的词的数量。

1.3 语句位置权值

文本中的句子在文献文本中的位置对于确定句子的重要性常常是有用的,为了确定这种根据句子在文本中的位置来给语句定义权值,作者对文本数据集作了抽样研究,研究表明,一个文本的前面句子常常能够提供关于文本内容的重要信息,而且句子的位置越靠前,其表达文本主题的语义越强,为了减少其权值份额,文中对文本语句的位置权值给出下列公式:

$$W_{\text{Position}}(L_i) = \frac{1}{\text{TSA} + \text{CSP}} \quad (7)$$

其中, $W_{\text{Position}}(L_i)$ 表示语句的位置权值;TSA 表示文献中所包含的语句数量;CSP 表示当前语句在文献中的位置。

1.4 语句含有重要词权值

从语句所包含的重要词也可以度量该语句的重要性。

$$W_{\text{Importance}}(L_i) = \frac{\text{STKA}}{\text{STWA}} \quad (8)$$

其中, $W_{\text{Importance}}(L_i)$ 表示语句 L_i 含重要词权值;STKA 表示文本语句 L_i 中包含重要词的个数;STWA 表示文本语句 L_i 中包含的所有有效词的数目。

1.5 语句权值

利用式(6)、(7)、(8),每个语句的权值可由几种语句分值经过线形插值相加得到:

$$W(L_i) = \lambda_1 W_{\text{Tide}}(L_i) + \lambda_2 W_{\text{Position}}(L_i) + \lambda_3 W_{\text{Importance}}(L_i) \quad (9)$$

其中, $\lambda_1 + \lambda_2 + \lambda_3 = 1$ 。

2 文本主题的抽取

输入:中文文本 ST,语句相似度阈值 θ 。

/* 当两个语句或两篇文本相似度大于 θ 时,认为是很相似的 */

输出:主题句集合 L' ,即文本 ST 的主题。

算法 2:

1) $L = \{L_i \mid L_i \text{ 表示 ST 的第 } i \text{ 条语句}\}$;

2) $L' = \emptyset, n = 1$; /* n 表示主题句数 */

3) 由算法 1 确定重点词;

4) do while ($n < |L| * 0.15$ and $L \neq \emptyset$) /* $|L|$ 表示 L 中语句个数 */

5) 利用(9)式,存在 k ,使 $W(L_k) = \text{Max}\{W(L_j) \mid L_j \in L\}$;

6) $L' = \{L_k\} \cup L'$; /* 权值最高的语句作为主题句 */

7) for $i = 1$ to $|L|$

8) if $\text{Sim}(L_k, L_i)^{[13]} < \theta$ then

9) if $i \neq k$ then $L = L - L_i$; /* 从 L 中删除 L_i

语句,即去重复语句 */

10) end if

11) end for

12) if $\text{sim}(\text{文本}, L')^{[13]} < \theta$ then

13) Exit /* 如果主题句集合可以代替文本,则退出循环 */

14) end if

15) $n = n + 1$;

16) end do

17) 输出 L'

3 文本集主题的抽取

输入:文本集 $U = \{S_1, S_2, \dots, S_p\}$ 。

输出: U 的主题集 U' 。

算法 3:

1) $U' = \emptyset$;

2) for $i = 1$ to p

3) 调用算法 2,计算文本 S_i 的主题 L'_i ;

4) $U' = U' \cup L'_i$;

5) End for

6) 输出 U'

4 试验与分析

文本的主题提取的试验结果可以使用精度和召回率度量来衡量,其中:

$$\text{抽取精度} = \frac{\text{反映主题的主题句数量}}{\text{抽取出的主题句总数}}$$

$$\text{召回率} = \frac{\text{抽取出的主题句数量}}{\text{文本中实际主题句总数}}$$

4.1 自动抽取测试结果

该测试文本是从香港中国资讯行页面上随机下载的 200 篇真实财经新闻。为了试验简单而有效,由有经验的标引人员对主题句进行自动抽取而且将试验结果进行判定,并将判定结果与全文本文献相比较。

对式(9)中参数 $\lambda_1, \lambda_2, \lambda_3$ 的确定,也可以由文献[14]的方法生成。其实根据文本类型的不同可以估计该参数,如在科技类、财经类文献中 λ_2 的参数值比文学类的文献要略高一些。作者经过一些测试,当取参数 $\lambda_1 = 0.39, \lambda_2 = 0.17, \lambda_3 = 0.44$ 时得到最佳评价结果如表 1 所示。

如果将准确反映文本内容和基本反映文本内容主题都视为符合文本主题抽取要求,由表 1 可见,其准

准确率达到了 92.5%。

表 1 文本主题自动抽取测试结果

可接受程度	评价结果	比例(%)
准确反映文本内容	143	71.5
基本反映文本内容	42	21
没有很好反映文本内容	15	7.5

4.2 准确率对比实验

对当前两种常见的文本主题抽取方法和文中给出的基于多因素的文本主题自动抽取方法在三组不同类型的文本数据上进行准确率实验。为了易于标示,将文献[1]中基于词汇频率方法提取文本主题简记为 WFSE,文献[4]的采用匹配和统计方法从文本中抽取主题简记为 WMSE,文中基于多因素的文本主题自动抽取方法简记为 MFSE。

为了实验简单,笔者从一些网站上选取了较典型的三组短文,一组数据从香港中国资讯行页面上随机下载的财经类新闻,共 200 篇,另两组是分别从新华网的时政论坛和军事论坛上各下载的 200 篇短文。由于抽取的这些文本大多没有段落标题,为了测试标题语句对文本主题提取的作用,笔者故意在时政论坛上下载的短文中加上了一些语句标题,从实验结果可以看出,大大提高了抽取的准确率(包括准确和基本准确反映文本内容两类)。

实验结果如表 2 所示。

表 2 文本主题提取准确率比较

	WMSE	WFSE	MFSE
财经新闻	0.824	0.782	0.925
时政论坛	0.842	0.743	0.947
军事论坛	0.802	0.801	0.916

从表 2 可以看出,基于多因素的文本主题提取方法的自动抽取文本主题的准确率是比较高的。这是因为文献[4]中的采用匹配和统计的方法从文本中抽取主题方法只是从主题词、主题概念和主题句三个不同层面上利用词汇匹配进行主题选取,因此对含有一些主题句的时政论坛文本主题抽取时准确率相对高一些,达到了 84.2%,其他两类中准确率低一些;而文献[1]中基于词汇频率提取文本主题方法中,是以字处理为基本单位的,而文中所选取的实验资料中的财政、军事类,专业词汇重现的较多,因此准确率较高一些,但是该算法由于完全避开了分词和抽词过程,因此该方法的执行速度最快;文中给出的基于多因素的文本主题提取方法先是利用概念间的同义关系、上下位关系等进行同义、层次的归并,在计算主题字串的权重时,

对它们统一进行度量,又根据语句标题、语句在文本中的位置及重要词汇权值等多个因素得到重要语句,因而对文本主题抽取的准确率最高,但是在这三种文本主题抽取方法中花费的时间也最长。

5 结束语

在目前对现有文本主题抽取方法研究的基础上,提出了一种主题抽取的具体实现算法。该算法是基于文本的特点,从词的同义归并来调整词的频率和上下位概念的聚焦来调整文本词汇概念的权值,文中一方面考虑到了下位词汇对上位词汇的进一步说明作用,另一方面又考虑到这种调整不影响整个文本主题的分布。

为了精确地抽取出文本的主题采用多因素权值综合度量方式,评估句子反映主题的价值。在此基础上,采用主题句选择算法将文本的主题数量与所抽取的主题句的数量相联系,以进行进一步的分析,并在保证每一个主题句被选中后,对后面主题句的选取考虑其去掉重复的问题,从而进一步提高所抽取主题的覆盖性和概括性。

参考文献:

- [1] 马颖华,王永成,苏贵洋,等.一种基于字同现频率的汉语文本主题抽取方法[J].计算机研究与发展,2003,40(6):874-878.
- [2] 麻志毅,姚天顺.基于情境的文本主题求解[J].计算机研究与发展,1998,35(4):344-348.
- [3] YN Z H, WANG Y C, CAI Wei, et al. Extracting Subject from Internet News by String - Match[J]. Journal of software, 2002, 13(2): 159-167.
- [4] 韩客松,王永成,沈洲,等.三个层面的文本主题自动提取研究[J].中文信息学报,2001,15(4):20-27.
- [5] BUN K K, ISH IZUKA M. Topic extraction from news archives using TF * PDF Algrithm[C]//The Third International Conference on Web Information Systems Engineering. Singapore:[s. n.], 2002: 73-82.
- [6] 陈炯,张永奎.一种基于词聚类的中文文本主题抽取方法[J].计算机应用,2005,25(4):755-756.
- [7] 何请,史忠植.机器学习与概念语义空间生成[J].中文信息学报,2002,16(3):814-819.
- [8] Sunayama W, Yachida M. Panoramic view system for extracting key sentences based on viewpoints and application to a search engine[J]. Journal of Network and Computer Applications, 2005, 28(2): 115-127.
- [9] 王继成,武港山,周源远,等.一种篇章结构指导的中文 Web 文本自动摘要方法[J].计算机研究与发展, 2003, 40

(下转第 79 页)

平台:奔腾 4 系列 2.80GHz, 1G RAM 的 3 台 PC 机。

测试方法:针对改进前和改进后的混合通信模型, 分别进行如下测试:订阅方发送某种类型消息的订阅请求后, 发布方发布大量不同长度的消息, 订阅方收到自己所订消息以后, 回复给发布方确认信息。经过大量严格的测试, 最后得到的结果如图 6 所示。

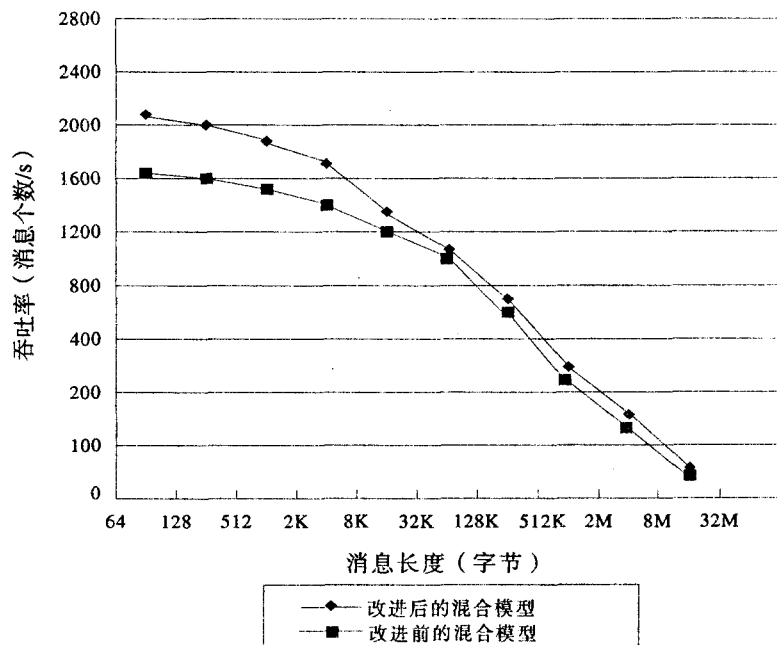


图 6 通信模型改进前后吞吐量对比图

4 结束语

在现有各种通信模型的基础上, 通过分析 PTP 通信模式、Pub/Sub 通信模式和混合通信模式在通信过程中存在的缺点——扩展性和实时性差, 提出了改进的混合通信模型。该模型采用主题树结构和基于向量 Vector 的匹配机制, 大大提高了系统匹配能力。该模型针对数据消息的不同规模, 既避免了在面对大规模数据消息时目录服务器中转传输的瓶颈问题, 也在一

定程度上避免了在面对小量数据消息时, 先经 Pub/Sub 处理再经过 PTP 处理而造成的消息延时问题。

参考文献:

- [1] Eugster P T, Felber P A, Guerraoui R, et al. The many faces of publish/subscribe[J]. ACM Computing Surveys, 2003, 35 (2): 114-131.
- [2] Nilsson D R, Mauget L E. J2EE 应用与 IBM WebSphere[M]. 马竹青, 鞠文飞, 译. 北京: 电子工业出版社, 2004.
- [3] 潘慧芳, 周兴社, 杨刚. 基于混合通信模型的消息中间件的设计与实现[J]. 计算机工程, 2006, 32(3): 116-118.
- [4] 石扬, 张燕平. 基于 Struts + Spring + Hibernate 的 Web - MIS 开发研究[J]. 计算机技术与发展, 2007, 17(1): 46-48.
- [5] 王伟卿, 孙莉. 基于 Java 消息服务的消息中间件的应用研究[J]. 计算机技术与发展, 2009, 19(7): 220-222.
- [6] Object Management Group. The Common Object Request Broker: Architecture and Specification, Version 3.0, Formal/02-06-01 [EB/OL]. 2002-06. <http://www.omg.org/technology/documents>.
- [7] 詹先银. 基于发布/订阅的消息中间件技术及其应用[D]. 西安: 西安电子科技大学, 2005.
- [8] 朱方娥, 曹宝香. 基于 JMS 的消息队列中间件的研究与实现[J]. 计算机技术与发展, 2008, 18(5): 172-175.
- [9] Schmidt D C, Huston S D. C++ Network Programming: Systematic Reuse with ACE and Frameworks[M]. [s. l.]: Addison-Wesley Longman, 2003.
- [10] Tran P, Greenfield P. Behavior and Performance of Message-oriented Middleware Systems[C]//Proceedings of the 22nd International Conference on Distributed Computing Systems Workshops. Vienna: [s. n.], 2002: 645-650.
- [11] LIH, YAMAN ISH I K. Topic analysis using a finite mixture model[J]. Information processing and management, 2003, 39 (3): 521-541.
- [12] Tombros A. Reflecting user information needs through query biased summaries[R]. Glasgow: Department of Computing Science, University of Glasgow, 1997.
- [13] 刘金岭. 一种基于语义的中文短信文本高质量聚类算法[J]. 计算机工程, 2009, 35(10): 201-205.
- [14] Zhang Y, Callan J, Minka T. Novelty and redundancy detection in adaptive filtering[C]//In: Proc. of the ACM SIGIR 2002. New York: ACM, 2002: 81-88.

(上接第 75 页)

(3): 398-405.

- [10] Morris A, Kasper G, Adams D. The effects and limitations of automated text condensing on reading comprehension performance[J]. Information Systems Research, 1992, 3(1): 17-35.
- [11] LIH, YAMAN ISH I K. Topic analysis using a finite mixture model[J]. Information processing and management, 2003, 39 (3): 521-541.