改进的潜在语义分析中文摘录方法*

肖 升[1,2], 何炎祥[1]
(1. 武汉大学 计算机学院, 武汉 430072; 2. 湖南第一师范学院 信息科学与工程系, 长沙 410205)

摘 要: 中文摘录是一种实现中文自动文摘的便捷方法,它根据摘录规则选取若干个原文句子直接组成摘要。通过优化输入矩阵和关键句子选取算法,提出了一种改进的潜在语义分析中文摘录方法。该方法首先基于向量空间模型构建多值输入矩阵;然后对输入矩阵进行潜在语义分析,并由此得出句子与潜在概念(主题信息的抽象表达)的语义相关度;最后借助改进的优选算法完成关键句子选取。实验结果显示,该方法准确率、召回率和 *F* 度量值的平均值分别为 75.9%、71.8% 和 73.8%,与已有同类方法相比,改进后的方法实现了全程无监督且在整体效率上有较大提升,更具应用潜质。

关键词: 自动文摘; 自动摘录; 潜在语义分析; 奇异值分解; 潜在概念
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)12-4507-05
doi:10.3969/j.issn.1001-3695.2012.12.027

Improved extractive summarization of Chinese texts using latent semantic analysis

XIAO Sheng[1,2], HE Yan-xiang[1]
(1. *School of Computer Science, Wuhan University, Wuhan* 430072, *China*; 2. *Dept. of Information Science & Engineering, Hunan First Normal College, Changsha* 410205, *China*)

Abstract: Chinese extractive summarization is a convenient method to realize Chinese text summarization, which extractes sentences and composites summarization corresponding to the extractive rules. This paper proposed an improved Chinese extractive summarization method using latent semantic analysis by optimizing input matrix and the key sentence selection algorithm. First, the method created multi-valued input matrix based on vector space model. Then it abtained the semantic correlation between sentences and latent conceptions(the abstract expression of theme) by latent semantic analysis for input matrix. At last, it extracted the key sentences by improved optimal selection algorithm. The experimental results show that the respective average for precision, recall and *F*-measure are 75.9%, 71.8% and 73.8%, and compared with the existing similar methods, the improved method becomes unsupervised completely and makes dramatical improvement of overall, so it has more potential application value.

Key words: text summarization; extractive summarization; latent semantic analysis(LSA); singular value decomposition(SVD); latent conception

0 引言

网络环境中,面对呈指数级增长的文本,在不阅读全文的前提下判别某文是否与特定主题相关,所依赖的判别对象通常是此文的摘要。作为提取文本关键信息并生成摘要的智能技术,自动文摘(text summarization)对提高文本检索效率的价值不言而喻。自动摘录(extractive summarization)是一种实现自动文摘的便捷方法,它根据摘录规则选取若干个原文句子直接组成摘要,由于不需重新生成语句,此方法在面对大量文本时效率突出。如果句子选取范围仅限于一个文本,则称为单文本摘录,否则称为多文本摘录。本文研究中文单文本摘录(简称中文摘录)。

以摘录规则涉及的语言学特征为标准,目前主流的中文摘录方法可分为两类:a)涉及分布特征,如句子在上下文中的相对位置、词语在相关文本中的出现频度等[1];b)涉及语义特征,如句子与主题的语义相似度、词汇语义链等[2]。相比而言,分布特征对主题的表征能力较弱,但易于提取,因此第一类方法计算简单,但准确率不高;语义特征对主题的表征能力较强,但需要借助外部知识,因此第二类方法的准确率虽有所提高但计算复杂,且难以适应无监督环境。综合考虑,准确率较高、算法复杂度适中且能适应无监督环境的中文摘录方法值得深入研究。

潜在语义分析(LSA)是一种构建潜在语义空间的代数统计方法,主题信息在潜在语义空间中能得到本质性表达,因此 LSA 可以保证中文摘录的准确率。同时,LSA 也是一种矩阵降维方法,其采取的分解、截断及反运算操作可使原有矩阵的计算复杂度降到适中。加之 LSA 在无监督环境中的效率已被信息检索、机器问答、文本分割等多种应用所验证[3,4],因此,运用 LSA 的中文摘录方法显然满足上述要求。

本文在跟踪研究此方法的基础上,对其中两个关键步骤进行了如下改进:a)在提供输入矩阵时,用多值综合方案代替单值方案;b)在进行关键句子选取时,用句子与潜在概念语义相关度的平均值和潜在概念的平均概念强度来筛选关键句子和概念。与已有成果相比,本文的改进不仅实现了过程的全程无

收稿日期: 2012-04-20; **修回日期**: 2012-05-28 **基金项目**: 国家自然科学基金资助项目(60703008);湖南省教育厅科学研究资助项目(10C0527);湖南省高校科技创新团队支持计划资助项目(湘教通[2010]212 号);湖南省科技厅高新计划资助项目(2010GK3049)

作者简介: 肖升(1980-),男,湖南武冈人,副教授,博士(后),主要研究方向为中文信息处理(xiaosheng@mail.ccnu.edu.cn);何炎祥(1952-),男,湖北应城人,教授,博导,博士,CCF 高级会员,主要研究方向为可信软件、自然语言处理.

监督,而且获得了更全面高效的实验结果,因此也更具应用潜质。

1 潜在语义分析

理解 LSA 的关键是理清降维和语义相关性分析这两大功能的关系,这种关系可描述为“抽象和具体”。降维是一种抽象的数学变化,当这种变化在具体领域发生时,就会使该领域具有较高语义相关度的对象被投射到同一维度所代表的语义空间中,由此便可在该领域展开语义相关性分析。

LSA 完成矩阵降维需要经历分解、截断及反运算三个阶段[5]。目前主流的矩阵分解方法是奇异值分解(SVD),数学上可以严格证明,一个矩阵的 SVD 分解结果是唯一确定且满足最小方差原则的。SVD 的简化定义如式(1)所示:

$$A_{t\times d}=T_{t\times n}S_{n\times n}(D_{d\times n})^{\mathrm{T}}\quad n=\min(t,d) \tag{1}$$

从式(1)可以看出,SVD 将原矩阵 A 分解为三个矩阵乘积的形式,其中:T 和 D 分别是由左奇异值向量和右奇异值向量组成的正交矩阵;S 是以奇异值为主对角线元素的对角矩阵(奇异值沿主对角线降序排列)。分解后对所得矩阵分别作 k 维截断($k<n$),即取 T 的前 k 列,D^{T} 的前 k 行,S 前 k 列中的前 k 行,便可得截断矩阵 $T'_{t\times k}$、$(D^{\mathrm{T}})'_{k\times d}$ 和 $S'_{k\times k}$,再根据需要对截断矩阵做反运算(如 $B_{k\times d}=S'_{k\times k}(D^{\mathrm{T}})'_{k\times d}$)即可得到新的降维矩阵。

中文摘录中,LSA 面对的分析对象是基于向量空间模型(vector space model,VSM)的“词语—句子”矩阵,该矩阵是为适应机器计算而设计的一种文本表达形式,文本主题信息就潜藏于其中。对该矩阵实施降维,可使具有较高语义相关度(通常是包含相同共现词语集)的句子被投射到同一维度,如果将每个维度视为一个支撑句子语义相关性的潜在概念,那么主题信息事实上就能够在潜在概念所组成的潜在语义空间中得到本质性表达,即借助语义相关性可以挖掘潜在主题信息。下面通过一个简单实例来说明 LSA 的工作原理。该实例文本(下文简称例 1)包括如下五个句子:

a)sent1,词汇歧义会涉及到同形异义词和多义词。

b)sent2,同形异义词的词形相同但意义不同。

c)sent3,同形异义词的词汇歧义可通过词形扩充消解。

d)sent4,多义词包含的多个义项是相关的。

e)sent5,根据上下文,多义词的词汇歧义可得到消解。

依次抽取例 1 中的名词及术语,并与句子联立,可以形成一个如表 1 所示的词语—句子矩阵。

表 1 例 1 的词语—句子矩阵

索引号	名词及术语	sent1	sent2	sent3	sent4	sent5
1	词汇歧义	1	0	1	0	1
2	同形异义词	1	1	1	0	0
3	多义词	1	0	0	1	1
4	词形	0	1	1	0	0
5	意义	0	1	0	0	0
6	扩充	0	0	1	0	0
7	消解	0	0	1	0	1
8	义项	0	0	0	1	0
9	上下文	0	0	0	0	1

注:索引号是根据名词及术语在文本中出现的顺序制定的。

表 1 中,矩阵因子 a_{ij} 的值代表名词或术语 i 在 sent$j$ 中的出现次数,对该矩阵进行 SVD,得到的奇异值矩阵 S 和右奇异值矩阵的转置矩阵 D^{T} 分别如表 2、3 所示,运算结果四舍五入保留两位小数。

表 2 例 1 的词语—句子矩阵经 SVD 后得到的奇异值矩阵

2.98	0.00	0.00	0.00	0.00
0.00	2.00	0.00	0.00	0.00
0.00	0.00	1.46	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	1.00

表 3 例 1 的词语—句子矩阵经 SVD 后得到的右奇异值矩阵的转置矩阵

潜在概念	sent1	sent2	sent3	sent4	sent5
LC_1	0.47	0.30	0.65	0.14	0.49
LC_2	-0.19	0.58	0.38	-0.38	-0.58
LC_3	-0.39	-0.49	0.40	-0.59	0.30
LC_4	-0.75	0.15	0.22	0.59	0.15
LC_5	0.19	-0.56	0.46	0.36	-0.56

分别对表 2、3 中的矩阵作 2 维截断(虚线框中是截断矩阵),再计算截断矩阵 $S'_{2\times 2}$ 和 $(D^{\mathrm{T}})'2\times 5$ 的乘积(做乘法反运算),所得的新降维矩阵如表 4 所示。

表 4 例 1 的词语—句子矩阵降到 2 维的结果

潜在概念	sent1	sent2	sent3	sent4	sent5
LC_1	1.40	0.89	1.94	0.42	1.46
LC_2	-0.38	1.16	0.76	-0.76	-1.16

表 4 的效果如图 1 所示。

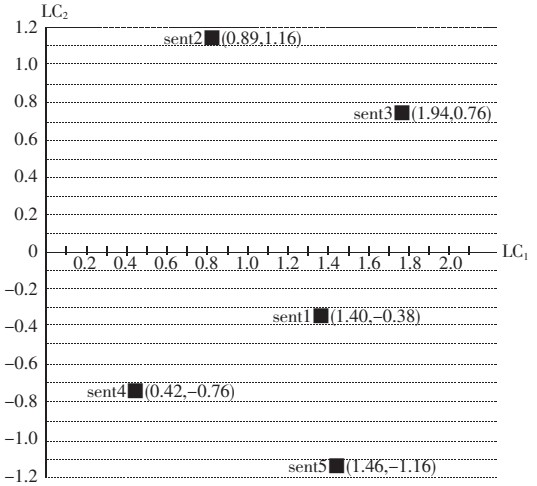

图 1 词语—句子矩阵降到 2 维的效果

对图 1 稍加分析,便可得出如下三个清晰的结论:

a)sent2 和 sent3 以及 sent4 和 sent5 之间的夹角很小,这说明这两组句子内部的语义相关度很高。事实上,sent2、sent3 涉及的是同形异义词,sent4、sent5 涉及的是多义词,分析与事实吻合。

b)sent2 和 sent3 以及 sent4 和 sent5 这两组句子以 sent1 为轴基本形成对称,这说明它们都与 sent1 相关且相关度基本相同。事实上,sent1 是总分结构中的总起句,而 sent2、sent3 及 sent4、sent5 从两个不同角度对 sent1 的主题信息进行论述,因此分析与事实吻合。

c)与 sent2 和 sent4 相比,sent3 和 sent5 与 sent1 之间的夹角更小,因此,它们更贴近 sent1 包含的主题信息。事实上,sent1 包含的主题信息除了有同形异义词和多义词外,还有词汇歧义,而 sent2 和 sent4 都不包含此信息,因此分析与事实吻合。从摘录角度看,sent1 和 sent3、sent5 组成摘要更为合理。

此外,LSA 对包含多义词和同义词的语义相关性分析也有独特功效,而这些并未在例 1 中得到展示。总而言之,LSA 的分析能力足以使其满足中文摘录的需求。

2　中文摘录流程

基于潜在语义分析的中文摘录主要包括如下三个步骤:

a)构建输入矩阵,即创建 LSA 的分析对象;

b)对输入矩阵进行 LSA;

c)应用 LSA 的分析结果选取关键句子。

上述三个步骤中,步骤 b)是核心也是贯穿整个流程的主线,因此其内容被单独列出并用第 1 章整章进行了阐述,下面重点讨论步骤 a)和 c)。

2.1　输入矩阵构建

如第 1 章所述,中文摘录中用于 LSA 的输入矩阵是词语—句子矩阵。之所以选择该矩阵,一方面因为"词语成句、句成文本"是一种合理的文本结构;另一方面词语与句子之间的语义相关度可以推演出词语与词语、句子与句子、句子与文本的语义相关度。因此不论从文本结构还是语义关联的角度看,词语—句子矩阵都适用于中文摘录。

词语—句子矩阵中存储的是词语与句子的语义相关度,表达这种相关度的方式有多种,每种方式都有其独特的语言学视角,选取不同方式,输入矩阵将具有不同的初始值。由于单一初始值会导致后期分析的片面性[6],本文选用多值综合方案,即用不同初始值所得效率的算术平均值作为最终综合效率。方案中涉及的初始值生成方式如下:

a)词语频度,即某个词或短语在某个句子中的出现次数,出现次数越多(频度越高),则语义相关度越高。

b)二值频度,即某个词或短语在某个句子中是否出现的逻辑值,如果出现,则记为 1,表示两者语义相关,否则计为 0,表示两者语义无关。

c)tf * isf 值,即词语频度和反句子频度的乘积,该方式仿照 tf * idf 权重公式族中的公式设计。其中 tf 为

$$\mathrm{tf}(i,j)=f(i,j)/\sum_k f(k,j) \tag{2}$$

其中:$f(i,j)$是词语 i 在句子 j 中的出现次数;$\sum_k f(k,j)$是句子 j 中所有出现的词语的次数之和。tf 值说明,如果词语 i 对句子 j 的组成越重要,则该词语与句子的语义相关度越大,否则越小。isf 为

$$\mathrm{isf}(i)=\log(|D|/s_i) \tag{3}$$

其中:$|D|$是文本包含的句子数;s_i 是包含词语 i 的句子数。isf 值说明,如果包含词语 i 的句子占文本句子总数的比例越小,则该词语与该句子的语义相关度越大,否则越小。综合考虑,tf * isf 值越高,说明该词语对该句子的语义识别度越高。

d)log 熵。log 熵用于描述词语在句子中的信息分布,其计算公式如下:

$$\text{log-entropy}=\text{global}\times\text{local} \tag{4}$$

$$\mathrm{global}(i)=1+(\mathrm{sum}/\log_2(n)) \tag{5}$$

$$\mathrm{sum}=\sum_j p(i,j)\log_2(p(i,j)) \tag{6}$$

$$\mathrm{local}(i,j)=\log_2(1+f(i,j)) \tag{7}$$

式(5)中的 n 是文本中包含的句子数;式(6)中的 $p(i,j)$是词语 i 在句子 j 中的出现概率;式(7)中的 $f(i,j)$是词语 i 在句子 j 中的出现次数。

由于并非每个句子都会包含所有词语,因此,通过上述任何一种表达方式建立的输入矩阵都有可能是一个高维稀疏矩阵,LSA 的降维功能有助于使其变为低维紧密矩阵。

2.2　关键句子选取

对输入矩阵进行 LSA 后,接下来需要根据分析结果选取关键句子。利用分解、截断和反运算,LSA 可以产生多个分析矩阵。考虑到关键句子选取的主体是句子,由句子向量组成的 D^{T}(或 D)无疑最为适合。此矩阵中存储的是句子与潜在概念的语义相关度,而潜在概念本质上是经过抽象的主题信息,因此与潜在概念的语义相关度事实上就是与主题的相关度,这种相关度正是关键句子选取的判断条件。

最朴实的选取算法是利用潜在概念在 D^{T} 中是按与主题语义相关度大小呈降序排列的特点,从上至下选取固定个潜在概念,并按 D^{T} 因子值的大小从每个潜在概念中选取一个语义相关度最高的句子作为关键句子[7]。依此算法分析,假如固定值为 2,那么表 3 中分别与 LC_1 和 LC_2 最相关的 sent3(0.65)和 sent2(0.58)会当选为例 1 的关键句子。该算法虽然操作简单,但却存在如下三个问题:

a)由于入选概念与主题的相关度存在差别,因此统一为每个入选概念选取一个关键句子显然不够灵活,重要的多选、次要的少选更为恰当。

b)由于与主题密切相关的句子可能并不与某个入选概念高度相关,因此以单个 D^{T} 因子值作为关键句子的选取标准并不可靠。

c)由于适合某些主题的概念数对其他主题可能过大或过小,因此按固定值选取潜在概念并不合适,过大会使与主题相关但并不重要的概念入选,过小会使重要的概念漏选。

针对上述问题,文献[8~11]提出了一些改进,包括用 $S\times D^{\mathrm{T}}$ 放大入选概念的差别,用句子长度代替单个 D^{T} 因子值来筛选关键句子,用新维度决定关键概念入选数目等。

本文在分析上述改进的基础上提出用句子与潜在概念语义相关度的平均值来筛选关键句子,用潜在概念平均语义强度来筛选关键概念。理论上,平均值方案将有助于获得更为全面和客观的分析结果。此外,句子与潜在概念语义相关度的平均值和潜在概念平均语义强度都无须额外输入,因此本文的改进也将使整个方法过程实现全程无监督。

改进后所得算法的 C 语言伪代码如下:

a) $B_{n\times d}=S_{n\times n}*(D_{d\times n})^{\mathrm{T}}$;//放大潜在概念的差别;

b)for $i=1$ to n {

　　for $j=1$ to d {

　　sum = sum + b_{ij};}

　　avg[i] = sum/d;}/ * 计算每个潜在概念与句子语义相关度的算术平均值 * /;

c)for $i=1$ to n {

　　for $j=1$ to d {

　　if(b_{ij} <= avg[i]) $b_{ij}=0$;}}/ * 将小于等于平均值的因子值置零(不相关),即句子选优 * /;

对例 1 而言,执行步骤 a)~c)后,结果如表 5 所示。

表 5　例 1 的"潜在概念—句子"矩阵经放大和句子优选后的结果

潜在概念	sent1	sent2	sent3	sent4	sent5	avg
LC_1	1.40	~~0.89~~ 0	1.94	~~0.42~~ 0	1.46	1.22
LC_2	~~-0.38~~ 0	1.16	0.76	~~-0.76~~ 0	~~-1.16~~ 0	-0.08
LC_3	~~-0.57~~ 0	~~-0.72~~ 0	0.58	~~-0.86~~ 0	0.44	-0.23
LC_4	~~-0.75~~ 0	0.15	0.22	0.59	0.15	0.07
LC_5	0.19	~~-0.56~~ 0	0.46	0.36	~~-0.56~~ 0	-0.02

注:结果四舍五入保留两位小数。

d)init($C_{n\times n}$);/ * 初始化一个 $n\times n$ 的潜在概念—潜在概念矩阵 * /;

e) for $i=1$ to n {

for $m=1$ to n {

for $j=1$ to d {

if(b_{ij} ! =0 && b_{mj} ! =0) $c_{im}=b_{ij}+b_{mj}$; } } } /* $C_{n\times n}$ 中的因子值 c_{im} 等于潜在概念 c_i 和 c_m 公共句子的语义相关度之和,如 LC_3 和 LC_4 的公共句子是 sent3 和 sent5,因此c_{34} = (0.58 +0.22) + (0.44 +0.15) =1.39 */;

f) for $i=1$ to n {

for $m=1$ to n {

stren[i] = stren[i] + c_{im}; }

sum = sum + stren[i]; }

avg_stren = sum/n; /* 计算每个潜在概念的概念强度和平均概念强度 */;

g) for $i=1$ to n {

if(stren[i] < = avg_stren) stren[i] =0; } /* 将小于等于平均值的概念强度置零,即概念选优 */;

h)入选概念的入选句子依次排列后,成为最终的摘录结果。由表 6 可知,例 1 的入选概念为 LC_1;由表 5 可知,LC_1 的入选句子为 sent1、sent3 和 sent5,因此摘录结果为 sent1、sent3 和 sent5。此结果与图 1 的结论 c)吻合,可见本文算法正确。另外在本算法中,所有数据都来源于输入矩阵,不需要外部知识,因此该算法是一个全程无监督算法。

表 6 例 1 的"潜在概念—潜在概念"矩阵经概念优选后的结果

潜在概念	LC_1	LC_2	LC_3	LC_4	LC_5	stren
LC_1	9.64	2.70	4.42	3.77	3.99	24.52
LC_2	2.70	3.84	1.34	2.29	1.22	~~11.39~~
LC_3	4.42	1.34	2.04	1.39	1.04	~~10.23~~
LC_4	3.77	2.29	1.39	2.22	1.63	~~11.30~~
LC_5	3.99	1.22	1.04	1.63	2.02	~~9.90~~
					avg_stren	13.47

注:结果四舍五入保留两位小数。

3 实验及其分析

虽然例 1 已证明本文方法可行,但面对成规模的真实语料本文方法是否有效,还依赖于实验及其结果分析。本文从政治学、物理学、计算机科学、心理学和语言学中各选取 10 个概念,并在百度百科中找到解释这些概念的中文文本作为实验语料。之所以选用学术文本作为实验语料,是因为评价学术文本摘要的主要标准是该摘要对原文信息的包含效率(准确率和召回率),而非在言语重新组织中被视为主要因素的连贯性,这与自动摘录的评价标准吻合[12]。此外,上述五个学科依次可基本被认定为文、理、工、偏理边缘、偏文边缘五种学科类型,通过这种设定希望消除因学科差异带来的实验误差。各学科的语料规模及句子平均信息量如表 7 所示。

表 7 各学科的语料规模及句子平均信息量

比较项	政治学	物理学	计算机科学	心理学	语言学
本文所含平均句子数	25.8	41.7	24.6	48.7	17.3
句子所含平均词语数	23.7	18.4	21.2	19.5	26.2
句子平均信息量	91.9	44.1	86.2	40.0	151.4

注:a)句子平均信息量 =(句子所含平均词语数/文本所含平均句子数)×100;b)结果四舍五入保留一位小数。

针对上述语料,本文进行如下实验:

a)预处理,包括以句号、疑问号和惊叹号为句末标志进行分句,用 ICTCLAS 进行分词及词性标注,去停用词(仅保留名词性词语和动词性词语);

b)运用四种不同的初始值生成方式(见 2.1 节)建立词语—句子矩阵,并以此作为输入矩阵;

c)用 MATLAB 中的 svds 函数对步骤 b)所提供的四个输入矩阵进行 SVD;

d)用优选算法(见 2.2 节)从步骤 c)中产生的四个潜在概念—句子矩阵中摘录关键句子,并生成自动摘要;

e)借助人工选取为每个文本建立标准摘要。

经步骤 a)~e)后,可获得如下三组数据:

a)1 篇自动摘要的句子数 k 及 10 篇的总数 $\sum k$。

b)1 篇标准摘要的句子数 s 及 10 篇的总数 $\sum s$。

c)在某组自动摘要和标准摘要中共现的句子数 l 及 10 组的总数 $\sum l$。根据这三组数据可得出如下三个用于评测某学科摘录效率的关键指标:

(a)准确率 $P=\sum l/\sum k$,即在两种摘要中共现的句子占自动摘要的比例越大,该摘要包含原文信息越准确;

(b)召回率 $R=\sum l/\sum s$,即在两种摘要中共现的句子占标准摘要的比例越大,该摘要包含原文信息越全面;

(c)F 度量值,$F=(2\times P\times R)/(P+R)$。由于 P 和 R 存在相互制约,因此要用 F 值综合评测。

依据上面三个关键指标所得的实验结果如表 8 所示。

表 8 中文摘录的实验结果 /%

指标		词语频度	二值频度	tf * isf	log 熵
政治学	P	72.5	82.6	74.6	75.2
	R	68.4	75.3	70.7	70.8
	F	70.4	78.8	72.6	72.9
物理学	P	73.0	83.1	75.7	76.8
	R	70.2	77.8	71.4	71.6
	F	71.6	80.4	73.5	74.1
计算机科学	P	71.5	80.3	73.4	74.2
	R	69.1	78.2	68.4	69.6
	F	70.3	79.2	70.8	71.8
心理学	P	72.9	81.4	75.4	77.2
	R	73.6	80.2	72.1	73.3
	F	73.2	80.8	73.7	75.2
语言学	P	71.7	80.9	73.1	73.4
	R	62.9	72.5	69.8	69.1
	F	67.0	76.5	71.4	71.2

注:结果四舍五入保留一位小数。

依据表 8 可以计算出 P、R、F 的算术平均值 Pa、Ra、Fa 分别为 75.9%、71.8% 和 73.8%。其中,$Pa=\sum_{ij}P_{ij}/(i\times j)$,$Ra=\sum_{ij}R_{ij}/(i\times j)$,$Fa=\sum_{ij}F_{ij}/(i\times j)$,$i$ 是初始值种类(本例为 4),j 是分组数(本例为学科种类 5)。值得说明的是,用算术平均值来实现多角度综合并非唯一选择,针对不同语料为不同初值产生的摘录结果分配不同权值再进行加权综合似乎更为合理,但鉴于权值分配暂无统一标准,本实验只得放弃。当然,作为全程无监督方法,表 8 的结果足以说明本方法已具备应用潜质,值得深入研究。

除测定综合效率外,还可对表 8 中的数据进行对比分析。

a)由表 8 可知,无论是考察准确率、召回率还是 F 度量值,对同一个学科而言,二值频度作为初始值的摘录效率都高于其他方式。当然这并非巧合,稍作分析便知,虽然二值频度仅记录简单的出现信息,但正是这种信息最直接、最突出地表达了词语的共现,而共现恰恰就是 SVD 赖以生效的基础,因此二值频度带来的高效完全在情理之中[13]。

b)表 8 中除政治学和计算机科学相反外,其余各学科的

摘录效率(以 F 为标准)与该学科语料的句子平均信息量(表7)基本成反比,即语料的句子平均信息量越大,摘录效率越低。如果注意到计算机科学的语料中存在影响语料规模统计的因素(如公式等),这个规律就更加可靠了。究其原因,如果每个句子包含的词语越多且文本包含的句子数越少,即句子平均信息量越大,则优选过程中删除一个句子所损失的关键信息就越多,摘录效率自然也就越低。

除上述用于测定方法自身效率的实验外,本课题组还基于相同语料将本文方法与文献[8~11]中的方法进行了对比性实验(具体步骤请参看相关文献)。对比性实验中所有的方法选用了相同的输入矩阵,并进行了相同的 LSA,因此实验结果所反映的事实上就是不同关键句子选择算法对方法整体效率带来的影响。该实验用准确率、召回率和 F 度量值的平均值作为比较参数,实验结果如表9 所示。

表9　本文方法与其他方法的对比分析　/%

对比项	本文方法	文献[8]	文献[9]	文献[10]	文献[11]
P 的平均值	75.9	65.4	69.1	71.7	66.3
R 的平均值	71.8	70.2	68.5	67.9	71.2
F 的平均值	73.8	67.7	68.8	69.7	68.7

表9 的结果显示,与文献[8~11]的同类方法相比,本文方法在各个比较参数的对比中都具有一定优势。可见改进后的方法不仅实现了全程无监督,而且在整体效率上有较大提升,更具应用潜质。

4　结束语

通过引入 LSA,本文提出了一种无监督的中文摘录方法,该方法借助 SVD 将主题信息抽象为潜在概念,并利用句子与潜在概念的语义相关度来判断句子与主题的语义相关度,由此完成关键句子的选取。实例及实验证明,本文方法可行且有效。

作为本文方法的核心和基础,LSA 在发挥其功效的同时也不可避免地遇到了如下两个问题:

a)LSA 所处理的输入矩阵其构建基础是向量空间模型,受限于独立性假设(单向量模型的分量或多向量模型的分向量表征的特征相互独立),此模型在支撑文本到输入矩阵转换时丢失了所有关联信息[14],而这些关联信息(如词序、句法语义约束等)很可能成为提升摘录效率的辅助条件。

b)当主题信息集中于个别句子时,输入矩阵可能出现数据不均的状况,此时,LSA 的分析效率会呈现一定程度的下降,这对摘录结果无疑会产生负面影响。

为解决上述问题,本研究下一步将重点思考如何用富含关联信息的模型(如图模型)来构建输入矩阵。此外,如何将 LSA 的应用范围从单文本摘录升级到多文本摘要也将是本研究努力的方向。

参考文献:

[1] 刘兴林,郑启伦,马千里. 一种基于主题词集的自动文摘方法[J]. 计算机应用研究, 2011,28(4):1322-1324.

[2] 陈燕,龙建勋. 基于明确语义分析的自动文摘算法[J]. 计算机工程, 2011,37(3):183-185.

[3] 余正涛,樊孝忠,郭剑毅,等. 基于潜在语义分析的汉语问答系统答案提取[J]. 计算机学报, 2006,29(10):1889-1893.

[4] 石晶,戴国忠. 基于 PLSA 模型的文本分割[J]. 计算机研究与发展, 2007,44(2):242-248.

[5] 盖杰,王怡,武港山. 潜在语义分析理论及其应用[J]. 计算机应用研究, 2004,21(3):9-13.

[6] 徐永东,徐志明,王晓龙. 基于信息融合的多文档自动文摘技术[J]. 计算机学报, 2007,30(11):2048-2054.

[7] GONG Yi-hong,LIU Xin. Generic text summarization using relevance measure and latent semantic analysis[C]//Proc of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM Press,2001:19-25.

[8] STEINBERGER J,JEZEK K. Using latent semantic analysis in text summarization and summary evaluation[C]//Proc of ISIM. 2004:93-100.

[9] MURRAY G,RENALS S,CARLETTA J. Extractive summarization of meeting recordings[C]//Proc of the 9th European Conference on Speech Communication and Technology. 2005:112-115.

[10] OZSOY M G,ALPASLAN F N. Text summarization using latent semantic analysis[J]. Journal of Information Science,2011,37(4):405-417.

[11] 马亮,何婷婷,李芳,等. 以关键词抽取为核心的文摘句选择策略[J]. 中文信息学报, 2008,22(6):50-54.

[12] 张瑾,王小磊,许洪波. 自动文摘评价方法综述[J]. 中文信息学报, 2008,22(3):81-88.

[13] KONTOSTATHIS A,POTTENGER W M. A mathematical view of latent se mantic indexing:tracing term cooccurrences,LU-CSE-02-006[R]. Bethlehem:Lehigh University,2002.

[14] 张晓艳. 新闻话题和关联追踪技术研究[D]. 长沙:国防科学技术大学,2010.

(上接第 4506 页)进一步的理论探讨,改进完善 Tableau 算法及其规则,探讨算法的复杂性以及实际应用的可行性等。

参考文献:

[1] 陆建江,张亚非,苗壮,等. 语义网原理与技术[M]. 北京:科学出版社,2006.

[2] 王金环,李宝敏. 基于本体 DL 的语义推理研究[J]. 计算机技术与发展,2009,19(11):94-100.

[3] WACHE H,VOEGELE T,VISSER U,*et al*. Ontology-based integration of information:a survey of existing approaches[C]//Proc of Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence. 2001:108-117.

[4] BAADER F,MCGUINNESS L D,NARDI D,*et al*. The description logic handbook:theory,implementation and application[M]. Cambridge:Cambridge University Press,2003.

[5] 石莲,孙吉贵. 描述逻辑综述[J]. 计算机科学,2006,33(1):194-197.

[6] HORROCKS I. Description logics in ontology applications[C]//Proc of the 9th International Conference on Automated Reasoning with Analytic Tableaux and Related Methods. Berlin:Springer,2005:2-13.

[7] 梅靖,林作铨. 从 ALC 到 SHOQ(D):描述逻辑及其 Tableau 算法[J]. 计算机科学,2005,32(3):1-11.

[8] HORROCKS I,SATTLER U. A tableau decision procedure for SHOIQ[J]. Automated Reasoning,2007,39(3):249-276.

[9] GLIMM B,HORROCKS I,LUTZ C,*et al*. Conjunctive query answering for the description logic SHIQ[C]//Proc of the 20th International Joint Conference on Artificial Intelligence. 2007:399-404.