文摘自动生成中权重计算的对偶性策略

卜东波 白 硕
中国科学院计算技术研究所 北京 2704 信箱 100080
bdb@ncic.ac.cn

摘 要

在指示性文摘自动生成系统中，核心工作是设置语句的权重，以合理地表示出语句在文章中的重要程度。我们发现在权重计算中，存在一种对偶性的现象，并利用迭代的方法来处理和利用这种对偶性，获得了语句的隐含概念。实验结果表明，直接摘取反映重要隐含概念的重要语句能够得到效果良好的文摘。

关键词

文摘自动生成 向量空间模型 特征抽取 对偶性 隐含概念空间

The Duplex Strategy of Weight Computing in Summarization

Bu Dongbo Bai Shuo
Institstute of Computing Technology, CAS Beijing 2704 Box 100080
bdb@ncic.ac.cn

ABSTRACT: The core problem in text summarization is to set the weight of each sentence to represent its importance in the whole text. We find that there is a duplex phenomena in the weight computing and use the iteration method to deal and utilize it. The weight could represent the latent concept of the text and the importance of each sentence.

Keywords: Summarization Duplex Weight Latent Concept Eigen Vector

1. 引言

自从 1958 年 Luhn 提出了第一个文摘系统开始，对文摘自动生成的研究就一直没有中断。大体上说，文摘自动生成可以分两条路线，一条路线是基于理解的自动文摘，试图在完全理解原文的基础上进行综合，产生概述原文的摘要。由于涉及自然语言理解和生成两

本文得到国家自然科学基金"超大规模真实原始数据的浓缩方法"课题资助，课题号 69773008

个不成熟领域，因此目前基于理解的文摘系统还很不成熟，距离实用还有相当长的一段距离。

另一条路线则是基于摘取重要句子的指示性文摘系统，直接摘取能够反映原文主题的语句编辑而成。指示性文摘系统的流程可以分成 4 步，即：单元选取、设置单元的权重、根据对文摘的长度要求选择重要语句形成文摘侯选句、排序和润色等编辑工作。而语句权重的设置则依据提示词、标题关键词、段落位置与其他版式信息、原文统计信息等。比如：一般认为文章的首段、末段以及每段的段首句和文章主题紧密相关，因此赋予较高的权重；含有“综上所述”、“总而言之”等提示词的语句是对全文的概括和总结，也应当是文摘候选句。合理地设置语句的权重，是指示性文摘系统中的核心工作。[8]

衡量一个文摘系统的重要指标就是完全性，即原文中的重要内容是否都已经被覆盖。而完全性则依赖于对原文的理解程度，仅仅依靠关键词和词汇的统计性质，以及简单的版式信息是远不够的，有必要进一步挖掘文章深层次的结构关系和语义信息。本文的目标就是提出一种更加有效的权重计算方法，以求得能够反映出文章的深层概念。

本文仍然使用 G.Salton 的向量空间表示模型 VSM（Vector Space Model），使用一个向量来表示文本中的一条语句，更详细地说，语句的内容被看作主要由一些特征项来表达，这些特征项可以是字、词等语言单位，即语句可以表示为 $Document = D(t_1, t_2 .. t_n)$，其中，$t_i$ 表示各个特征项。换句话说，由这些项张成了一个向量空间，每个项表示一个维度。在一条语句中，每个特征项都被赋予一个权重 W，以表示这个特征项在该语句中的重要程度。权重都是以特征项的频率为基础进行计算，比如采用频数、信息熵等技术。采用 VSM，每条语句都被表示成一个向量，而一篇文本则用一个矩阵来形式化表达。[5]

我们发现，在权重计算中存在一个循环，本文使用迭代的策略来打破这种循环，并证明这种迭代操作是收敛的，特征项的权重最终稳定于一个矩阵的特征向量上。为每个特征项赋予迭代计算出的权重，实际上就得到了语句的隐含概念。和直接采用特征项仅仅反映了语句的表层信息相比，这种隐含概念能够更深刻地反映出语句的深层结构。实验结果表明，和直接在特征项空间中表示语句相比，在概念空间中表示语句能够更好地表达语句的语义信息。

2. 权重计算中的对偶性策略

文[3][4]认为，在整个文本聚类操作中，存在一个基本循环（Basic Cycle），即：要想聚类必须首先进行特征抽取和设置权重，聚类结果的好坏直接依赖于特征抽取和设置权重的合理与否；而合理的特征抽取和权重设置应当使得样本类内方差尽量地小，同时类间方差尽量地大，这样就要求必须首先知道聚类的结果。也就是说，聚类和特征抽取及权重设置互为因果，两者构成一个循环。文本聚类中的循环关系提示我们在语句和词之间可能存在某种对偶性，而在文摘自动生成中也存在这种对偶性。这种对偶性在以下的权值计算中表现得更明显。

假设共有 m 条语句，使用 n 个特征项。我们为每个语句和每个特征项都定义一个权值。

语句 f_i 的权值 Wf_i 表示该语句对整篇文本语义的反映程度和概括程度，权值越大的语句越重要，概括程度越强；特征项 t_j 的权重 Wt_j 表示该词对于整篇文本语义的反映程度，权值越大的特征项就越重要，其反映文本主题的能力也就越强。

语句和特征项的重要性之间存在着这样的一种对偶关系：

- 一个重要的语句就是包含许多重要词的语句；
- 一个重要的词就是经常出现在重要语句中的词。

这种对偶关系实际上是对重要性的一个循环定义，无法各自独立地定义语句和词的重要性。如何由这种循环定义来定量地计算出语句和词重要性呢？文对超语句进行链接分析的技巧给我们以很大的启发：对付这种循环，迭代方法是一件利器。

开始时赋予 Wf，Wt 随机值，这里我们使用 Wf 表示语句权重向量，Wt 表示词权重向量，即：

$$Wf = (Wf_1, Wf_2, ...Wf_m)'$$

$$Wt = (Wt_1, Wt_2, ...Wt_n)'$$

然后进行如下的两步迭代过程：

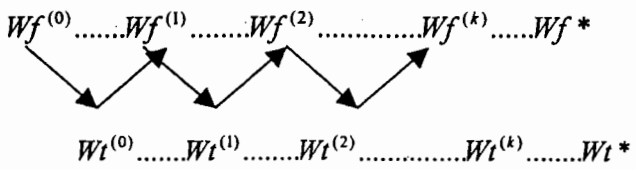
① 使用当前对词权重的估计值 Wt 来改进对语句权重的估计值 Wf，找出当前比较重要的词，包含这些词的语句就是比较重要的语句，因此相应地增加这些语句的权重。具体来说，每个语句更新后的权重 Wf_i 等于它包含的所有词的词频和词权重乘积的总和，也就是词权重向量 Wt 和词频矩阵第 i 行向量的内积。直观地看，包含重要词较多的语句将获得较高的得分。

② 使用当前对语句权重的估计值 Wf 来改进对词权重的估计值 Wt，找出当前比较重要的语句，经常在这些语句中出现的词就是比较重要的词，因此相应地增加这些词的权重。具体地说，每个词更新后的权重 Wt_j 等于所有含有这个词的语句的权重与词频乘积的总和，也就是语句权重向量和词频矩阵第 j 列向量的内积。直观地看，在那些重要语句中经常出现的词将获得较高的得分。

反复进行以上的两步迭代过程，语句权重向量 Wf 和词权重向量 Wt 将稳定在一个不动点上，这个不动点仅仅和 m 行 n 列的词频矩阵 $A_{m\times n}$ 相关。

我们使用 $Wf^{(0)}$ 和 $Wt^{(0)}$ 分别表示向量 Wf 和 Wt 的初始值，使用 $Wf^{(k)}$ 和 $Wt^{(k)}$ 分别表示经过 k 次迭代之后得到的改进值，Wf^* 和 Wt^* 表示最终的稳定值，

下图形象地描述了迭代求解的过程。



使用线性代数可以更清楚地分析迭代过程，每次迭代操作实际上是在做向量和矩阵的乘法运算，即：

$$Wt^{(k+1)} = (A_{m\times n})^T \times Wf^{(k)}$$
$$Wf^{(k+1)} = A_{m\times n} \times Wt^{(k+1)}$$

对于任意给定的初始值，这种迭代过程都是收敛的，并且最后的稳定值恰好是矩阵 $A*A^T$ 和 A^T*A 的一个特征向量。首先得到以下的引理：

引理 1 矩阵 $A*A^T$ 和矩阵 A^T*A 有相同的非零特征值。

证明：设 A*Tr(A)有某个非零特征值 λ，相应的特征向量为 η，即：

A*Tr(A)* η = λ * η

两边同乘以 Tr(A)，得到下式：

Tr(A)* A *Tr(A)* η = λ * Tr(A)* η

因此，Tr(A)* η 就是 Tr(A)* A 的一个特征值为 λ 的特征向量。

同理可证，对 Tr(A)*A 的特征向量 ζ，A * ζ 就是 A*Tr(A)的特征向量。

所以，矩阵 A*Tr(A)和矩阵 Tr(A)*A 有相同的非零特征值，它们的特征向量之间存在着一一对应关系，并且非零特征值的个数就是矩阵 A*Tr(A)的秩。

这个看起来不起眼的引理的作用却很重要。在很多应用场合下，我们需要求出 $A*A^T$ 的特征值或者特征向量，但是有时方阵 $A*A^T$ 的维数特别高，而求特征向量过程的时间复杂度是 $O(n^3)$ 的，非常耗时。如果方阵 A^T*A 的维数较低的话，一个变通的方法就是先求出 A^T*A 的特征值或特征向量，然后再依据引理 1 求出 $A*A^T$ 的特征值或特征向量。这样不仅能够节省大量的时间，更重要的是可以避免大规模运算带来的误差累积，使得结果更准确。

定理 2 对于任意给定的初始向量 $Wf^{(0)}$ 和 $Wt^{(0)}$，迭代过程都是收敛的。Wf 将稳定于矩阵 $A*A^T$ 的某个特征向量上，Wt 将稳定在 A^T*A 的某个特征向量上。[6][7]

证明：设矩阵 $A*A^T$ 的秩为 r，非零特征值对应的特征向量分别是：

$$\eta_1, \eta_2 \cdots\cdots \eta_r$$

矩阵 A^T*A 的诸非零特征值对应的特征向量是：

$$\xi_1, \xi_2 \cdots\cdots \xi_r$$

相应的特征值是 $\lambda_i (i = 1,2..r)$，并且按照降序排列，即：

$$\lambda_1 \geq \lambda_2 \geq \cdots\cdots \geq \lambda_r$$

因为矩阵 $A*A^T$ 是一个对称矩阵，所以这 r 个特征向量是相互正交的。

设初始向量 $Wf^{(0)}$ 在这个特征空间中的坐标为 $(c_1, c_2 ... c_r)$，即：

$$Wf^{(0)} = c_1 * \eta_1 + c_2 * \eta_2 + ...c_r * \eta_r$$

因为经过一次迭代操作后，有：

$$Wt^{(k+1)} = (A_{m\times n})^T \times Wf^{(k)}$$
$$Wf^{(k+1)} = A_{m\times n} \times Wt^{(k+1)}$$

故有下式成立：

$$Wf^{(1)} = A * A^T * Wf^{(0)}$$

继而推出：

$$\begin{aligned} Wf^{(k+1)} &= A_{m\times n} \times (A_{m\times n})^T \times Wt^{(k)} \\ &= (A * A^T) * (c_1 * \chi_1 * \eta_1 + c_2 * \chi_2 * \eta_2 + ...c_m * \chi_m * \eta_m) \\ &= c_1 * \chi_1^{\,2} * \eta_1 + c_2 * \chi_2^{\,2} * \eta_2 + ...c_m * \chi_m^{\,2} * \eta_m \end{aligned}$$

$$Wf^{(2)} = A * A^T * Wf^{(1)}$$

$$\begin{aligned} &= (A * A^T) * (c_1 * \eta_1 + c_2 * \eta_2 + ...c_m * \eta_m) \\ &= c_1 * \chi_1 * \eta_1 + c_2 * \chi_2 * \eta_2 + ...c_m * \chi_m * \eta_m \end{aligned}$$

.....

$$\begin{aligned} Wf^{(k)} &= c_1 * \chi_1^{\,k} * \eta_1 + c_2 * \chi_2^{\,k} * \eta_2 + ...c_m * \chi_m^{\,k} * \eta_m \\ &= \chi_1^{\,k} * (c_1 * \eta_1 + c_2 * \left(\frac{\lambda_2}{\lambda_1}\right)^k * \eta_2 + ...c_m * \left(\frac{\lambda_m}{\lambda_1}\right)^k * \eta_m) \end{aligned}$$

因为：$\lambda_1 > \lambda_2 > ..... > \lambda_r$

所以：

$$\left(\frac{\lambda_2}{\lambda_1}\right) < 1, \left(\frac{\lambda_3}{\lambda_1}\right) < 1 ..... \left(\frac{\lambda_m}{\lambda_1}\right) < 1$$

因此，当 k 足够大时，有：$Wf^{(k)} \approx c_1 * \chi_1^{\,k} * \eta_1^{\,k}$。即，$Wf$ 的稳定值 Wf^* 是 $A*A^T$ 的一个特征向量，Wf^* 依赖于 Wf 的初始值。

同理可证，Wt 将稳定在 $A^T * A$ 的某个特征向量上。

熟悉线性代数的人马上就可以看出，上述过程就是幂法求矩阵特征值和特征向量的过程。

推论 Wf 和 Wt 的稳定值 Wf^* 和 Wt^* 满足下面的关系式：

$$Wt^* = (A_{m\times n})^T \times Wf^*$$
$$Wf^* = A_{m\times n} \times Wt^*$$

上述推论实际上说明了这么一种关系：语句集合在词向量空间中表现成一群点，每条语句在这个空间中的坐标构成矩阵 A 的一个行向量，而 Wt^* 表示词向量空间中的一个方向，Wf^* 的则是这些语句在这个方向上的投影。

Wf^* 是语句权重向量，它的各个分量表示相应的语句对文本合语义的概括程度，权重越大的语句越重要，然而，这种重要性只是从某个侧面看的结果，因为 Wf^* 是这些文件在 Wt^* 方向上的投影，它仅仅反映了从 Wt^* 方向上对各个文件重要性的衡量。Dumais 提出的 LSI[1][2]技术中将这种方向称为“隐含的概念”，这种概念不是仅由某一个词就能完整表达的，而是由一类词共同拥有的语义或者经常共同出现来表达的。因此，Wt^* 只是反映了文本中的某一个“隐含的概念”，或者说某一个主题，Wf^* 则表示了各个语句对这个主题的贡献的大小，从这个主题来看各个语句的重要与不重要。

在 Clever 系统中使用类似的技巧来进行超语句的链接分析。但是和我们在这里的应用不同的是在 Clever 系统中为每个页面赋予两个权重，分别表示页面的内容权威程度和引用程度，它处理的矩阵仅仅是 m 个节点之间的关联矩阵，是一个 m 阶方阵；而且矩阵的每一个元素都是 0/1 二值，以表示两个节点之间是否有链接关系。[6][7]

3. 概念空间

很少会出现仅仅只有一个主题的文本，通常的文本都会有多个主题或曰“隐含的概念”。比如，随着 Wt 选取不同的初始值，Wt^* 会得到不同的稳定值 $\xi_1, \xi_2 \ldots \xi_r$。$\xi_1$ 反映了语句集合中的一个概念，ξ_2 则反映了 ξ_1 所不能表达的另一个概念，而 ξ_3 则反映了 ξ_1 和 ξ_2 都不能表达的某个概念…… 每个 ξ_i 都反映了语句集各不相同的主题。任意两个稳定值 ξ_i 和 ξ_j 都是两两

正交的，直观地说，某个ξ_i所对语句集合主题的反映作用，是不能被其他的ξ_j完全代替的。

针对某个特定的主题ξ_i，可以定义各条语句对这个主题的反映程度，也就是语句的重要程度。对于一个语句，我们使用其在ξ_i方向上的投影来定量地刻画该语句对主题ξ_i的反映程度，投影为正数的文件可以看作这个主题的赞同，投影为负数的语句可以视为对这个主题的否定，而投影的绝对值大的那些语句对于反映这个主题的作用也比较大，绝对值小的语句的反映力也较小。所有语句的权重合起来恰好就是和ξ_i对应的向量η_i。

诸个“隐含概念”ξ_i有着不同的重要性，概念之间也有主次之分。这种重要性可以使用和ξ_i相应的特征值λ_i来定量刻画，对应特征值越大的概念就越重要，特征值相对较小的概念就比较次要。从信息的角度看，特征值λ_i的大小定量地表达了概念ξ_i蕴涵的信息量的多少，严格的数学证明表明，特征值λ_i恰好是所有语句在概念方向ξ_i上投影的方差，它表示投影的散布情况，散布越大则蕴涵的信息量就越大，散布越小蕴涵的信息量就越小。

以上对重要性的定义和迭代过程中观察到的现象是吻合的。从任意的初始值$Wt^{(0)}$开始迭代，迭代结果Wt^*一般说来总是要收敛到主特征向量，也就是特征值最大的那个特征向量，换句话说，如果只允许使用一个概念来概括所有语句的话，那么我们就要选择最重要的概念，也就是最大特征值对应的特征向量；如果排除掉主特征向量的作用，也就是说从和主特征向量正交的$n-1$维子空间中随机选取初始值，那么迭代结果一般说来会收敛到和次大特征值相应的特征向量，换言之，如果允许使用两个概念来概括所有语句的话，我们会优先选择最大和次大特征值所对应的特征向量。

如果我们以各个“隐含概念”ξ_i量为坐标轴，一个语句的坐标是其在概念方向上的投影，定义一个新的坐标系来表示所有语句，这个新的空间可以称为概念空间。

4. 隐含概念在文摘自动生成中的应用

使用上述的迭代方法，我们不仅能够合理地设置每条语句的权重，而且还能够挖掘出文本中隐藏的深层概念结构。每个隐藏概念也都有一个重要性度量，即与之相应的特征值λ_i来定量刻画。因此我们可以可以忽略一些不重要的概念。那些重要性特别低的概念不是文本意图的重点，蕴涵的信息量比较小，忽略掉并不会影响大局，因此可以作为噪声过滤掉；每条语句也有一个权重，反映该语句对相应概念的反映程度。因此我们直接摘取那些反映重要概念的重要语句，就能够有效地概括文本的主题。

我们对一些新闻题材的文本进行了测试，得到了较为满意的结果。在附录中列示了一篇新闻文本，使用上述方法求得重要的隐含概念以及各个句子的权重如下。我们按照权重的大小摘取一些句子组成摘要。

权重 句子序号	概念ξ_1	概念ξ_2	概念ξ_3	概念ξ_4
1	0.11	-0.04	-0.04	-0.03
2	0.19	0.004	-0.03	-0.02
3	0.08	-0.07	0.12	0.02
4	0.08	-0.05	-0.02	0.08
5	0.13	0.02	-0.05	-0.08
6	0.09	-0.04	-0.04	0.08
7	0.14	0.16	0.08	0.03
8	0.13	0.002	-0.03	0.02
9	0.12	-0.03	-0.04	-0.02
10	0.07	-0.09	0.10	-0.05

5. 结束语

在指示性文摘系统中，如何合理地衡量语句的权重是最核心的工作。我们采用迭代的方法求取语句的权重，证明了迭代是收敛的，而且采用这种方法求得的权重恰好能够反映出文本的深层概念和语义信息。实验结果表明直接摘取反映重要概念的重要语句，得到了较好的结果。

参考文献

[1] LSI meets TREC: A Status Report. In: the First Text Retrieval Conference(TREC1), D.Harman,ed., National Institute of Standards and Technology Special Publication 1993 S.T.Dumais

[2] Latent Semantic Indexing(LSI)and TREC-2. In: the Second Text Retrieval Conference(TREC2), D.Harman,ed., National Institute of Standards and Technology Special Publication,1994. S.T.Dumais

[3] 聚类分析 地质出版社 方开泰，潘恩沛，1982

[4] Cluster Analysis Heinemann Education Books Ltd. Brian Everitt, 1980

[5] 大规模中文语句的检索、分类与摘要研究 复旦大学学位论文 黄萱菁

[6] 网络的超搜索 科学美国人 No. 9 1999 Clever Team

[7] Authoritative sources in a hyperlinked environment, in: Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998 J. Kleinberg

[8] Hongyan Jing and Kathleen R. McKeown 2000. Cut and paste based text summarization. In Proceedings of NAACL 2000

附录：电讯稿原文

空军新招飞行学员心理素质明显提高

新华社北京5月19日电（通讯员李永芳）①据有关部门考查，空军自行招收的首批学员在飞行基础学校被淘汰和转入飞行学院后因技术原因停飞的比过去大大减少。②空军招收飞行学员工作办公室负责人在介绍这一情况时认为，飞行学员心理素质明显提高，显示了心理选拔在招飞工作中举足轻重的作用。

③我国空军目前应用的招飞心理素质内容，包括积极的飞行动机，良好的感知能力、记忆能力、思维判断能力、注意能力、空间定向能力、手足运动协调能力、抑制能力、应变能力和情绪稳定性10项个性心理特征。④它们互为一体，基本上能将飞行学员的心理状况全部调动起来，做到较全面、系统、综合的考查。

⑤国外招收飞行学员进行心理选拔，早在第一次世界大战期间就开始了。⑥据西欧一个国家统计，当时百分之九十的飞行事故是由于驾驶员不具备飞行的心理素质所造成的。

⑦中国空军1988年专门建立了招飞心理选拔机构。⑧目前已形成了一支包括有丰富飞行教学经验的飞行干部、航空心理学专业人员在内的招飞心理选拔骨干队伍。

⑨空军招收飞行学员工作办公室负责人告诉记者，据最近对入校新飞行学员的调查看，经过心理选拔的学员质量明显好于过去。学员整体素质好，⑩飞行动机端正，反应灵敏协调，记忆力、判断力、模仿力以及自控能力等比过去有很大提高。（完）

文摘：

⑤国外招收飞行学员进行心理选拔，早在第一次世界大战期间就开始了。⑦中国空军1988年专门建立了招飞心理选拔机构。⑧目前已形成了一支包括有丰富飞行教学经验的飞行干部、航空心理学专业人员在内的招飞心理选拔骨干队伍。②空空军招收飞行学员工作办公室负责人在介绍这一情况时认为，飞行学员心理素质明显提高，显示了心理选拔在招飞工作中举足轻重的作用。（完）