多文档自动文摘中的特征组合优化①

Combination Optimization of Features in Multi – documents Automatic Summarization

刘茂福 李淑君 金可佳 张晓龙 （武汉科技大学 计算机科学与技术学院 湖北 武汉 430065）

摘 要：在分析当前多文档自动文摘方法中使用较多的特征基础上，提出了一种特征组合优化模型。该模型选用 tf * idf，句子位置及与标题句相似度来判断句子包含信息的重要程度，并加入了句子长度特征解决由 tf * idf 特征引起的长度偏长的句子占优势的问题，采用这 4 个特征来判断句子的重要性，并给每个特征指定权重来解决优化问题，实验结果表明特征组合优化模型在多文档自动文摘中的可行性。

关键词：tf * idf 句子位置 标题句相似度 句子长度 组合优化

1 引言

随着网络的迅速发展，人们接触到的数据急剧增多，当人们面对成千上万同一主题的网页，它们大多具有相同的信息，又包含少量不同的信息，如何快速准确地获取这些关键信息成为人们关注的问题。文本摘要可以帮助人们花更少的时间获得更多有用的信息。

文摘是准确全面地反映某一文本中心内容的简洁连贯的短文。自动文摘就是利用计算机自动地从原始文献中提取文摘。多文档文摘是将多文档集合中多次重复的信息以一次出现在文摘中，其他与主题相关的信息根据重要性及压缩比依次抽取的文本集合压缩技术[1]。目前多文档文摘的主要方法是将多文档集合作为一个整体研究，将文档集中的句子按其表达意思的相近程度组合聚类，然后从不同的类别中抽取文摘句。在国内外目前的自动文摘研究中，计算句子重要度用到的特征有：词频，位置信息，相似度等。这些特征从不同的方面体现了信息的重要性，如何更好的组合优化这些特征使文摘更准确更全面的反映原文信息是本文关注的问题。

统计特征是自动文摘中常用到的方法。统计方法的自动文摘系统是利用文章的形式特征来提取摘要，如词频、关键词、词的位置、词控制表和指示性的句子等[2]。但单纯以统计的方法来衡量句子中的词在文档的重要性，没有考虑其语义环境，同时忽略了文章的结构信息及包含信息的重要程度，此外，在采用词频特征时会使句子的重要性偏向于较长的句子，因此，本文采用四个特征：词的 tf * idf、句子的位置、句子与标题句的相似度以及句子长度特征来解决以上问题。本文以 DUC2001 语料为基础，以句子为基本处理单元，将句子的 tf * idf 与句子位置、标题句相似度以及句子长度特征相结合，并优化这四个特征的权重找出最佳的组合方式。

2 特征组合优化模型

模型采用自动摘录的多文档文摘方法。自动摘录（Automatic Extraction）将文本视为句子的线性序列，将句子视为词的线性序列[3]；按照句子的各个特征计算每个特征项的得分，并按一定方式组合优化各个特征项的权重得到句子的最终得分，按最终得分排序，抽取句子生成文摘。

模型主要分 4 个模块实现：预处理，独立特征计算，特征组合优化及文摘的生成。

2.1 预处理

预处理主要是将 DUC2001 语料中的每个文档划

① http://gate.ac.uk/

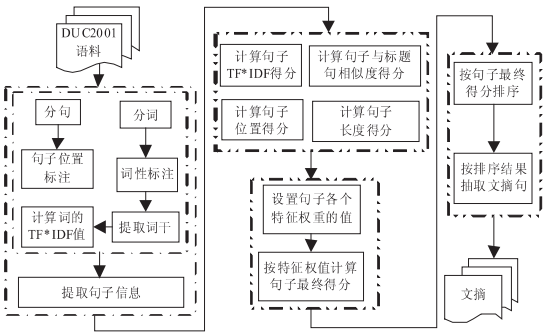

图1　模型过程图

分为句子单元。利用 GATE[①] 作为分词分句工具。该工具将每篇文章划分为一个个句子并指示出该句子在源文档中的位置，此外该工具还标识句子中每个单词的词性并提取该词的词干。根据该处理结果计算每个词 tf * idf 的值。将结果按一定的结构保存在文件中，最后按一定规则提取所需要的信息。

2.2 独立特征计算

根据预处理的结果，计算每个句子的各个特征项得分。

1）句子 tf * idf 得分

该特征表示若句子包含文档中重要的单词越多，该项得分越高。此项得分为句子中除去停顿词后所有单词的 tf * idf 值的总和。tf * idf 通用的计算方法是用词在文档中出现的频率与在大型语料中出现的文档频率的倒数之积作为短语统计信息含量的度量方法，其中文档频率是指语料库中包含该术语的文章数目。

2）句子位置得分

本文把文档内容看成是句子的线性序列，每篇文档中第一句话最重要，其它句子按其位置重要性依次递减。位置越接近第一句此项得分越高。

3）句子与所属文档标题句的相似度得分

标题体现它所属文档中最重要内容，若每个句子与它的相似度越大则表明该句包含的重要信息越多，该项得分越高。此项得分采用向量内积的方法计算，并将结果标准化。

4）句子的长度等分

采用正态分布模型计算该特征项的得分。句子的长度越接近文档句子平均长度，此项得分越高。句子平均长度是同一主题下的所有文档的单词个数总和除以句子总数的值。

2.3 特征组合优化

这一部分给每个特征项设置一定的权重，每个特征项的最后得分为该特征项的得分乘以其权重，句子的最终得分为每个特征项的最后得分的总和。

2.4 文摘生成

根据句子最终得分排序，抽取满足用户要求的排在前面的句子作为文摘句并生成文摘。

3 特征选择

本文主要选取了四个特征：句子 tf * idf、句子位置、与标题句的相似度以及句子长度。

3.1 句子 tf * idf

文章的内容可视为由一些意义基元来表达的，词频统计的方法将词形看成是意义基元[4]。本文以句子为基本处理单元，将句子看成是词的集合，在文档中频率较高的词比较重要，而包含较多高频词的句子，其重要性也较大。

该特征得分是该句中每个非停顿词的 tf * idf 值的总和。

$$T_{i,k}=\sum_{w\in Sen_{i,k}} TF(w)*IDF(w) \quad (1)$$

$$TF(w)=\frac{N(w_{Dk})}{Sum(D_k)} \quad (2)$$

$$IDF(w)=\ln\frac{N(D_{set})}{N(D_w)} \quad (3)$$

公式(1)中 $Sen_{i,k}$ 表示第 k 个文档中第 i 个句子，w 表示该句中的非停顿词。$T_{i,k}$ 表示该句 tf * idf 特征项的得分。公式(2) TF(w) 表示 w 的 TF 的值，D_k 表示第 k 个文档，$N(w_{DK})$ 表示 w 在文档 k 中出现的次数，$Sum(D_k)$ 表示 D_k 中单词的总数；公式(3)中 IDF(w) 表示 w 的 IDF 的值，$N(D_{set})$ 表示同一个主题文档集中的文档的个数，$N(D_w)$ 表示在这个文档集中出现词 w 的文档个数。

公式(3)说明在同一个主题下的文档集中，同一个文档中同一个词的 TF 值一定相同，而不同文档中该词的 TF 值不一定相同，为了使所有的词 tf * idf 特征项的值具有可比性，本文采用均值的方法来统一计算的标准。

此外，从公式(2)和公式(3)及实验结果可以得出：一个文档中包含的单词的个数远远大于其中某个

单词在该篇文档中出现的次数,因此,每个词的 tf * idf 特征的值都非常小且每个句子的该特征项的值都不超过 1,因此,未对该特项进行标准化。同时由于选取的每个特征的得分都在 0 -1 之间,故各个特征之间具有可比性。

3.2 句子位置 Position

美国的 Baxendale 的调查结果显示:段落的论题是段落首句的概率为 85%,是段落末句的概率为 7% [5]。基于文档的这种结构,每篇文章的第一句话包含了最重要的内容。本文把文档看成句子的线形排列,第一个句子最重要,第二句次之,依次类推句子 i 在文档 k 中的位置得分[6,7]:

$$P_{i,k}=\frac{n-i+1}{n} \tag{4}$$

公式(4)中 $P_{i,k}$ 表示 $Sen_{i,k}$ 的位置特征项的得分;n 表示第 k 个文档的句子总数。

3.3 与主题句相似度 Similarity

基于文档结构,每篇文章的标题反映出该篇文章的主题思想,包含文档最重要的信息,而文档中的每个句子与主题有着某种联系,在一定程度上体现文档信息。本文把每篇文档中的标题句看成的最重要的句子,文档中的每个句子与标题句计算相似度,与标题句相似度较大者该特征项的得分较高。若源文档中没有标题则把文章第一句看作标题。

$$S_{i,k}=\frac{\left\langle \vec{Sen}_{i,k},\vec{D}_k\right\rangle}{\left\langle \vec{D}_k,\vec{D}_k\right\rangle} \tag{5}$$

$$\left\langle \vec{Sen}_{i,k},\vec{D}_k\right\rangle=\sum_{v\in Sen_{i,k}\cup D_k} v_{sj}*v_{dj} \tag{6}$$

公式(5)中 $S_{i,k}$ 表示 $Sen_{i,k}$ 的与主题句相似度特征项的得分,$\vec{D}_k$ 表示第 k 个文档的标题句向量。公式(6)中 $\left\langle \vec{Sen}_{i,k},\vec{D}_k\right\rangle$ 表示 $\vec{Sen}_{i,k}$ 与 $\vec{D}_k$ 的内积,v_{sj},v_{dj} 表示 $Sen_{i,k}$,$\vec{D}_k$ 中第 j 个分量,且该分量代表同一个单词。如果句子中包含该词则将该词对应分量值为 1,反之则为 0。

将标题句和每个句子向量化,向量化的句子不包含停顿词,计算内积时将相同单词对应的分量相乘,计算所有对应分量乘积的总和,再将该值标准化即除以标题句自身内积。

3.4 句子长度 Length

为避免由 tf * idf 特征所导致的得分偏向于长句子(单词较多的句子该项得分较高)问题,同时由于过长的句子包含一定的冗余信息,而过短的句子包含的信息较少,采用正态分布模型计算每个句子长度的得分,即每个句子的长度与文档集中句子平均长度相比,接近平均句子长度的句子得分较高,过长或过短的句子得分较低。这样在一定程度上避免了上述问题,使抽取的文摘句长度适中并尽量包含重要信息。

$$L_{i,k}=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7}$$

$$\sigma^2=\frac{\sum_{1}^{n}(x_i-u)^2}{n} \tag{8}$$

公式(7)中 $L_{i,k}$ 表示 $Sen_{i,k}$ 的句子长度特征项的得分;μ 为同一文档集中句子平均长度;x 为该句包含单词的个数。

公式(8)中 s 为每个句子与句子平均长度的均方差,x_i 为文档集中第 i 个句子的长度,n 为文档集中句子的总数。

通过给每个特征指定一定的权重计算句子的最终得分,并在实验中改变 4 个特征的权重,找出评测结果最好的特征组合及权重比值,在所有的实验结果中 4 个特征权重最好的比值为 1∶2∶1∶1。

$$Score(Sen_{i,k})=\omega_t T_{i,k}+\omega_p P_{i,k}+\omega_s S_{i,k}+\omega_l L_{i,k} \tag{9}$$

公式(9)中 $Score(Sen_{i,k})$ 表示句子 $Sen_{i,k}$ 的最终得分,ω_t,ω_p,ω_s,ω_l 分别代表上述 4 个特征的权重。

4 实验结果及评价

本文所有实验是基于 DUC2001 语料并用 Rouge 工具进行评测。DUC 是目前在多文档文摘领域最有影响的评测会议,由 NIST② 的系列会议之一 TIDES 赞助发起的文本理解会议。DUC2001③ 的语料是由 NIST 提供的 60 个相关文档集合,其中 30 个为训练集,30 个为测试集,每个集合已按照一定的标准分类。本文所有实验基于 30 个测试集进行的。ROUGE 方法是由 Lin 等于 2002 年提出并在 2004 年 DUC 上正式使用的。

这个评价方法是通过计算系统产生的文摘和由人工文摘间所重叠的单词数目来评价系统文摘[8]。

评测是通过自动生成的文摘与人工生成的文摘进行比较,主要参考三个评测量的得分:

- 准确率 P
- 召回率 R
- F – scores = 2P * R/(P + R)

召回率用来衡量系统生成文摘的信息覆盖率,而准确率衡量系统生成文摘的精度。F 测度来表示系统的总体性能。

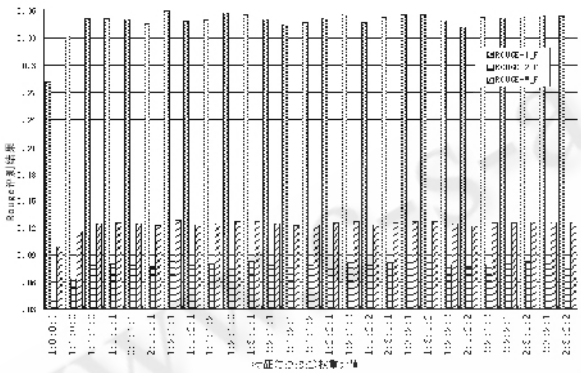

图 2　特征组合优化结果

tf * idf 特征是利用文章的形式特征来提取摘要,如关键词等,包含较多高频词的句子,其重要性也相对较大,因此将它与其它特征相结合,从不同的方面较全面的判断和比较各个特征间的优化关系。此外,段落的论题落在段落首句的几率远大于其它句子,故有必要提高段首句子位置的权重。在此基础上标题是整篇文章的核心内容,包含标题内容越多的句子在一定程度上越能反映出该句子的重要程度。然而,越长的句子虽包含的信息越多,但其中的冗余信息也相对较多,因此对句子的长度使用一个评测标尺,尽量使文摘句以最少的词包含最多的信息,从而可以更加全面的体现原文档集的核心思想。实验结果如图 2 所示:

实验主要分为 4 个部分:

1) 将 tf * idf 特征与其它特征组合,比较本文所选用的 4 个特征是否可以从不用的方面体现句子的重要性,并找出文摘评测结果最好的特征组合。所有实验结果表明将 4 个特征结合起来效果最好。

2) 提高一个特征的权重,比较结果是否有所提高。实验结果表明提高位置特征的效果最好,其中特征的权重比值为 1∶2∶1∶1。

3) 提高两个特征的权重,从不同的组合中找出最好的组合,以达到特征优化的目的,实验结果表明提高位置和长度特征或位置与主题句相似度特制的评测结果较好,但结果没有权重比值为 1∶2∶1∶1 的评测结果好。

4) 提高 3 个特征的权重。实验结果同样没有权重比值为 1∶2∶1∶1 的结果好。

表 1 所列出的特征组合中出现了的特征项的权重为 1,未出现的特征项的权重为 0。

表 1　特征组合

评测结果 / 特征组合	ROUGE – 1 F	ROUGE – 2 F	ROUGE – W F
T	0.28125	0.03926	0.09876
T + P	0.33210	0.06379	0.11577
T + P + S	0.35103	0.07819	0.12443
T + P + S + L	0.35135	0.07965	0.12484
P + S + L	0.35017	0.07796	0.12443

从表 1 可以看到:单独用文档中词的统计特征 tf * idf 来生成文摘,结果并不好;在生成文摘的过程中,把该特征与其它特征结合起来有效的提高了文摘的质量,将 tf * idf,位置及与标题句的相似度 3 个特征结合起来有效的提高了文摘的评测结果。在此基础上加入句子长度特征在一定程度上也提高了评测结果,但效果不太明显。若完全不考虑 tf * idf 使结果有所降低,将 4 个特征结合起来时评测结果最好。

表 2 中每个特征项前面的数值代表该特征项的权重。

表 2　特征优化

评测结果 / 特征组合	ROUGE – 1 F	ROUGE – 2 F	ROUGE – W F
1T + 2P + 1S	0.35763	0.08021	0.12669
2T + 2P + 1S + 1L	0.35044	0.07802	0.12439
1T + 2P + 1S + 1L	0.35909	0.08298	0.12763

从表 2 可以看到:优化 tf * idf,位置及与标题句的相似度 3 个特征结合的方式,有效的提高了文摘的评测结果,在优化后的组合中加入句子长度特征使评测结果有了明显的提高,可见句子长度特征在一定程度

上确实解决了由于 tf * idf 特征所引起的得分偏向于长句子的问题。

从所有实验中我们得出:tf * idf 特征反映了文档集中出现频率较高的重要词汇,单从统计特性上体现出句子的一定的重要性,却忽略了句子之间重要信息的包容性以及文档结构信息。标题句相似度特征有效的反映出句子中包含重要信息的程度;句子位置特征从文档的结构上充分体现重要信息的分布情况;句子长度则在一定程度上解决了句子过长或过短问题,弥补了 tf * idf 特征引起的得分偏向长句子的问题。

通过实验我们发现,优化后的特征组合大大提高了系统的性能,将 4 个特征结合并提高句子位置特征的权重有效的提高了文摘的质量。在所有实验中 4 个特征最佳的权重比值为 1.2:1:1。

5 结束语

本文采用了句子 tf * idf、位置、与标题句的相似度以及句子长度这四个特征作为最终得分的评判标准,通过组合方式找到了最佳的组合及它们的权重比值,有效的提高了文摘评测结果。虽然评测结果有了很大的提高,但在后处理部分以及句子相似度的计算中仍有不足,还需要进一步的研究和探讨。将外部资源引入句子相似度的计算以及如何提高后处理方法的效率,增加文摘的可读性是我们今后的研究方向。

参考文献

1 秦兵,刘挺,李生. 多文档自动文摘综述. 中文信息学报,2005,19(6):13-20.

2 吴岩,李秀坤. 自动文摘基集语句的提取与润色的数学模型. 计算机应用研究,2007,24(5):52-55.

3 刘挺,王开铸. 自动文摘的四种主要方法. 情报学报,1999,18(1):10-19.

4 万敏,罗振声,季妲,高小云. 基于概念统计的英文自动文摘研究. 计算机工程与应用,2002,38(24):7-9.

5 哈罗德-博科,查尔斯. L. 贝尼埃合著,赖茂生,王知津合译. 文摘的概念与方法. 书目文献出版社,1991.

6 Yuan Ding. A Survey on Multi-Document Summarization. Department of Computer and Information Science University of Pennsylvania, 2004.

7 Dragomir R. Radev, Hongyan Jing, Malgorzata Stys, Daniel Tam. Centroid-based summarization of multiple documents. Information Processing and Management. 2004, 40:919-938.

8 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. 中文信息学报,2005,19(2):93-99.